문학석사 학위논문

Developing Theory-Based Diagnostic Tests of English Grammar: Application of Processability Theory

처리가능성 이론에 기반을 둔 진단적 영문법 평가 시험 개발 연구

-

2014 년 2 월

서울대학교 대학원

영어영문학과 어학전공

Rosalie Hirch

Developing Theory-Based Diagnostic Tests of English Grammar: Application of Processability Theory

지도 교수 이용원

이 논문을 문학석사 학위논문으로 제출함
2014 년 2 월

서울대학교 대학원
영어영문학과 어학전공
Rosalie Hirch

Rosalie Hirch의 문학석사 학위논문을 인준함
2014 년 2월

위 원 장 박 용 배

부위원장 신 효 필

위 원 이 용 원

Abstract

An important consideration for developing diagnostic language tests is whether these should follow a particular course of study, or whether students and teachers are better served with theory-based tests. One particular grammar-acquisition theory that has garnered recent attention from language testers and SLA researchers is Manfred Pienemann's (1989) Processability Theory (PT). The Rapid Profile diagnostic test has already been developed based on this theory, and is currently in use; one substantial limitation of this test, however, is that it is a speaking test, and therefore difficult to administer in foreign countries. Furthermore, this type of testing requires language testers with either native or native-like proficiency, which may be restrictive in countries such as Korea, China, and Japan.

Researchers have developed other tests with different task types using PT. These tend to be productive tasks; however, two notable exceptions are those developed in Norris (2005) and Chapelle *et al*. (2010). Those studies employed PT to develop university-level placement tests appropriate for computer-based testing, with task types that attempt to imitate production (writing). These tasks can be fine-grained and allow for more control over the contexts that students are given, which well suits PT, and may be more accessible for foreign English language teachers in situations where there are few native English speakers. On the other hand, there are limitations; the test items may not test what they purport to, which leads to a false positive, indicating full acquisition on a grammar point not yet acquired. Another possibility is that the contexts are insufficient to show acquisition for another reason, such as topic unfamiliarity; in that case, the test would show a false negative, suggesting no acquisition when a grammar point has, indeed, been acquired.

The test developed for this study incorporates both types of tasks described above: writing and blended. The writing task is a story-telling task based on six pictures, designed to elicit the same types of grammar tested on the second half of the test. The grammar points are similar to those tested in Chapelle et al. (2010), with the exception that there are fewer, which accommodates context requirements while keeping the test at a reasonable length. The test was piloted twice to junior high school students from Korea, and adjustments were made before the final test was developed. This test was given to 200 Korean junior high school students. Students and teachers also received diagnostic feedback based on the test results. Qualitative and quantitative analyses were done on the results to analyze similarities and differences between the two task types and diagnostic information they offered.

The results suggested that the two task types performed similarly, in that they both showed implicational hierarchies comparable to those proposed in PT, but the blended-type tasks showed a tendency toward being less productive. Implications for diagnostic test designs are discussed.

Keywords: Diagnostic Language Assessment, Processability Theory, PT, Writing Assessment, Productive Skills, Productive Tasks, Placement Tests
Student Number: 2009-23815

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Among the various types of recognized language testing suggested by Alderson (2005)—proficiency, achievement, progress, placement, aptitude, and diagnostic—one that has seemed somewhat neglected until recent years is diagnostic (Donohue & Erling, 2012; Alderson, 2005). Far more attention has been paid to large-scale, high-stakes testing, especially proficiency, partially because of a better understanding of the reliability and validation processes for them, the need to consider not only test score reliability but social consequences (Bachman, 1990; Messick, 1995), and also because of the washback effect that such testing has on society as a whole, and the language classroom in particular (Bachman, 1990). Concerns about social consequences and washback effect have also led to a focus on more "communicative" forms of testing that resemble real-life use by combining skills in one task (Bachman, 1990; Oller, 1979). This type of testing has created its own problems, however; the interaction of skills is tremendously complex, placing far more demands on cognitive domains than do other subjects such as math (Buck & Tatsuoka, 1996). While tasks designed to emulate communicative ability do show us what test-takers can do—and by extrapolation, cannot do—the complexity of the interactions required to complete those tasks makes it very difficult to understand *why* test-takers cannot do, which puts an additional burden on language teachers

to help their students (Shohamy, 1992; Buck & Tatsuoka, 1996; Alderson, 2005; Jang, 2009b).

This is not to say that communicative tasks cannot be repurposed for diagnostic use, as there is a substantial amount of crossover among the different types of testing (Alderson, 2005; Jang, 2009a). Researchers have considered whether diagnostic language testing can ever be norm-referenced (Richards, 2008), and whether all criterion-referenced tests, regardless of their purpose, can be adapted for diagnostic purposes (Simpson & Arnold, 1983). Most attention, however, has been given to distinguishing among proficiency, achievement, placement, and diagnostic testing, particularly to those aspects that make each unique (Shohamy, 1992; Alderson, 2005). Proficiency testing is seen as broad tests that measure test-takers' overall or "real-life" language skills (Jang, 2008; Yin, 2011), while achievement testing measures a test-takers' knowledge or skills with reference to a particular course of study or treatment (Shohamy, 1982). These two types of testing, though often retrofitted for diagnostic purposes, are generally recognized as being more distinct in nature from diagnostic and also placement testing (Yin, 2011; Richards, 2008). One important element that distinguishes between proficiency and achievement, on one hand, and placement and diagnostic, on the other, is the direction in which the two types of testing face (Kunnan & Jang, 2009); the first two are, by their nature, backward-looking, in that they focus on what the learner knows, or "can do", meaning that they focus on what the student has already

accomplished, either in a specific class or overall (Yin, 2011). The other two are forward-looking, focusing primarily on what the learner *cannot* do, but needs to learn about in the future; they are intended to make predictions about what course of study would be most suitable for the student in the future, as opposed to evaluating the success of what the student has already done (Yin, 2011). For this reason, placement and diagnostic testing are seen as being essentially the same, as they inform an appropriate course of instruction; they attempt to determine what the learner does not know in an effort to facilitate learning (Richards, 2008).

These distinctions suggest that what really distinguishes among these types of testing is not so much the type of tasks or the design of test that are used, but the purpose for which these tests were originally designed. This focus on purpose is in keeping with the view of validity as being at least partially dependent on the way a test is used (Messick, 1995). An immediate observation of the difference in purpose between backward- and forward-looking tests is in the way they are *intended* to affect classroom instruction: while backward-looking tests may affect instruction, their influence is indirect, which is known as the washback effect; this is a consequence of the nature of backward-looking tests, which also tend to be high-stakes. Forward-looking tests, by contrast, are usually low-stakes. They are also specifically designed to *directly* affect classroom instruction, in an immediate way, and should therefore serve to counterbalance washback. Placement testing does so by putting students into a course of study that is at

an appropriate level to give them the maximum amount of new information (learning) while not exceeding what they are able to do. However, placement testing is aimed at identifying the weaknesses of a relatively large group of students (a class); diagnostic testing is directly aimed at individual students, at what each specifically does not know, rather than what the group in general does not know. Following this line through, achievement testing is somewhat similarly focused on the level of a class, while proficiency testing tends to have a much broader focus, on the widest range of subjects possible. These concepts of forward- and backward-looking testing are presented in Figure 1.1, which also shows the ways in which these tests overlap; each type can be used for a different purpose, though its effectiveness in that role is diminished.

The distinction between forward- and backward-looking, high- and low-stakes testing is an important one, since it affects test construction; although a task might be used for both proficiency and diagnostic testing, it

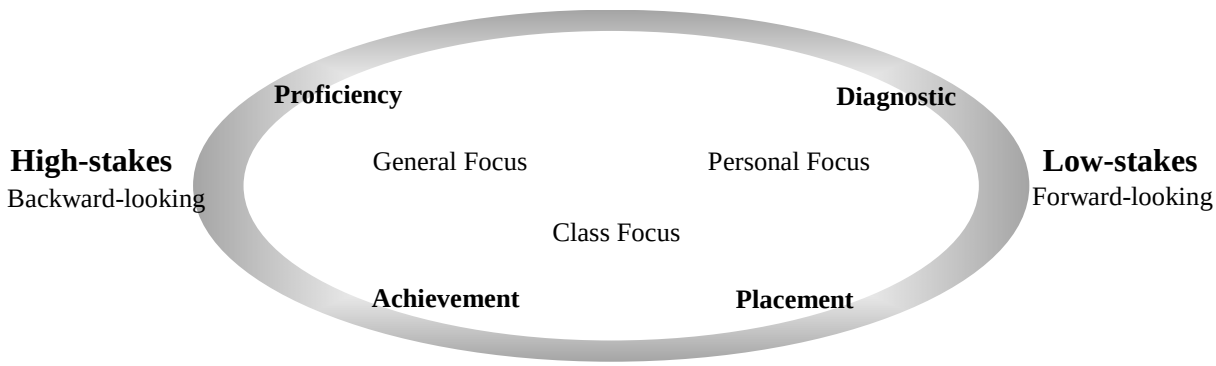


Figure 1.1 **Comparative Qualities of Different Test Types**

may not contain the same information value for both purposes (Shohamy, 1982; Alderson & Huhta, 2011; Donohue & Erling, 2012). In fact, researchers have found that diagnostic testing is likely to require a greater number of items than does proficiency testing in order to get sufficient information for diagnostic purposes (Jang, 2009b). The large number of test items required is due to the need for diagnostic tests to measure learner strengths and weaknesses, and the fact that this cannot be done at a general level, but must be specific (Alderson & Huhta, 2011). Even so, it should be noted that *all* language tests have at least some degree of diagnostic capability, depending on how the output is analyzed (Yin, 2011; Blatchford, 1971; Lee & Sawaki, 2009).

Overall, diagnostic language testing tends to be poorly defined and understood (Alderson & Huhta, 2011; Richards, 2008; Alderson, 2005). Among the difficulties is the fact that diagnostic language assessment has, in the absence of any definitive theories for language acquisition, been seen as the responsibility of the teacher in the classroom in relation to a particular course of instruction (Alderson & Huhta, 2011). Further complicating our understanding of diagnostic language assessment is that any diagnostic test—whether in the classroom, or in a medical setting, or for a mechanic inspecting a car—is usually defined in reference to a baseline “normal” that has proved difficult to define in second language studies (Alderson & Huhta, 2011). Nonetheless, there must be some way to create tests for diagnostic purposes that measure a student’s strengths and weaknesses, and to use this

information to guide classroom learning (Kunnan & Jang, 2009; Lee & Sawaki, 2009). In this sense of diagnostic testing measuring both strengths and weaknesses, there is both a backward-looking element to what the student can do (Jang, 2009b; Simpson & Arnold, 1983), and a forward-looking element to what the student cannot do, and should therefore work on, both in the short- and long-term (Simpson & Arnold, 1983; Shohamy, 1982). Overall, the purpose of diagnostic assessment—whether done by creating a new test or by repurposing an established test—must be to inform the stakeholders as to students' abilities or lack of abilities, which can effect change that will remedy the latter while enhancing (or at least, not being detrimental to) the former (Alderson & Huhta, 2011; Shohamy, 1992). In other words, diagnostic assessment consists of three fundamental components: diagnosis; feedback; remediation. All three must be considered in the course of making any diagnostic assessment.

The current study seeks to contribute to the dialogue on diagnostic language assessment generally by developing a diagnostic grammar test, utilizing a theory of grammar acquisition known as Processability Theory (Pieneman, 1998, 2005, 2011), rather than attaching the diagnostic assessment to a particular course of instruction. The primary purpose of this study is to attempt to validate the test that was developed; aspects of feedback and remediation, though considered, are components that require their own, additional studies. A secondary aspect of the study is to evaluate the theory for use on a diagnostic grammar test. The choice of grammar for

this study is, in some respects, a little unconventional, given the fact that so much focus in language acquisition and assessment has gone to communicative skills and away from traditional grammar-translation methods (Alderson, 2005; Ellis, N, 2008; Donohue & Erling, 2012). However, there are several strengths that a diagnostic test of grammar has over diagnostic tests of other language skills. Most of these reasons will be given in the section on diagnostic language tests below, but it is important to point out that grammatical ability is an element of language ability/competence that is relatively simple to identify and test, as compared to other, less understood and more complex skills such as reading and writing (Donohue & Erling, 2012; Alderson & Huhta, 2011), and furthermore, syntax plays an important role in both reading and writing, perhaps greater than that of vocabulary (Alderson & Huhta, 2011). Several studies have also suggested that grammatical ability develops alongside other communicative skills such as pragmatics (Håkansson & Norrby, 2005). Finally, automaticity plays an important role in communicability, and as an element of communication that seems to strongly rely on automaticity, grammar is a key element of communicability and general proficiency (Alderson & Huhta, 2011). These reasons, and others, will be explored further in the literature review section.

1.2 Research Questions

The current study sought to validate a diagnostic test of grammar that was developed by combining a theory of grammar acquisition known as Processability Theory (Pienemann, 1998, 2005, 2011) with characteristics of diagnostic language tests outlined in Alderson (2005). To further assist with developing appropriate items, this study employed tasks that were created for two previous studies: John Norris (2005) and Carol A. Chapelle, Yoo-Ree Chung, Volker Hegelheimer, Nick Pendar, and Jing Xu (2010; hereafter Chapelle *et al.*, 2010). The research questions focus first of all on validating the test for use in diagnosing grammar problems. As a corollary element, the questions also look at Processability Theory and how well it worked for developing a diagnostic test.

1. Can we achieve an acceptable level of score reliability for the grammatical diagnostic test used for this study?
2. Do the items for the grammatical diagnostic test work well at an item level in terms of item discrimination and difficulty? Were there any poorly performing items?
3. What are the relationships among the subtest, full test, and self-assessment?
4. What are the perceptions of the test from the viewpoint of the test-takers, raters, and teachers?
5. Are mastery and non-mastery patterns consistent with predictions based on the Processability Theory hierarchy?

1.3 Organization of the Thesis

This thesis is divided into six chapters. Chapter 2 is a review of the literature on Processability Theory and diagnostic language assessment, followed by a closer look at two particular studies that influenced the design of the grammar test for this study. Chapter 3 describing the methods used for the study, and Chapter 4 presents the results of the analysis, focusing on validating the diagnostic grammar test that was designed. Chapter 5 discusses the results of the analysis in order to evaluate the reliability and validity of the grammar test. The final chapter, Chapter 6, gives a few conclusions and makes some suggestions for future areas of research based on the results of this study.

Chapter 2

Literature Review

Given the lack of strong theoretical models for language acquisition, a substantial number of tests, regardless of their purpose, are built largely on teachers' and researchers' perceptions of what language proficiency is, and how it is best tested. The Literature Review presents the specific theory of grammar acquisition that the diagnostic test is built on and then describes the characteristics of diagnostic language tests. The Literature Review ends with the review of two previous studies that have item types which strongly influenced the test design.

2.1 Processability Theory

This section of the paper focuses on one theoretical approach to developing a diagnostic language test specifically on grammar, Processability Theory (hereafter, PT) (Pienemann, 1998, 2005, 2011). The theory has already been used to develop the Rapid Profile Test™ for diagnostics, but there remain some questions about its usefulness for diagnostic purposes, as well as some concerns about the theory itself. This discussion begins with the elements of Levelt's (1989) psycholinguistic model of speaking on which Processability Theory is based, followed by how PT uses Lexical Functional Grammar (LFG). The paper then gives an explanation of how these elements combine to create an implicational hierarchy, as well as the types of proofs the theory requires. It finishes with an explanation of how the theory views errors and learner variation.

2.1.1 Levelt's Speaking Model

Processability Theory (PT) began through investigations of developmental stages in German, and has since evolved into a universal theory of second language acquisition (Pienemann, 1998, 2005). It begins with the assumption that all learners develop internal grammars, or interlanguages, in a systematic way. Learners use this interlanguage to gradually automatize in the second language (L2) (Håkansson & Norrby, 2010). Numerous studies have been conducted in a variety of languages, including Arabic (Mansouri, 2005; Salameh, Håkansson, & Nettelbladt, 2004), Chinese (Zhang, 2005), Italian and Japanese (Di Biase & Kawaguchi, 2002), in order to demonstrate that PT may be a universal theory (Pienemann, 1998, 2005). What makes it universal is that, despite the different languages tested, the types of interlanguages produced by learners are consistent with the theory (Jansen, 2008; Di Biase & Kawaguchi, 2002). The development of PT's interlanguages has its origins in Levelt's (1989) model of language processing for speaking (Pienemann, 2011).

An essential element of Levelt's (1989) model of speaking is that the language processing required for speech production occurs incrementally (Dyson, 2008). The model postulates that a particular speech production is first conceived abstractly in the conceptualizer, where it is assigned meaning, and then passes to the formulator, where it is encoded (the full model from Levelt is presented in Figure 2.1 below). The first stage of the encoding process is grammatical encoding, which is where a speech utterance takes its

grammatical form. At this stage, words are placed into their lexical categories, and are then combined into an appropriate syntactic order. This grammatically encoded utterance then passes through the phonological encoder before becoming an overt speech act. Of interest to PT is what occurs in the grammatical encoder (Pienemann & Kessler, 2011). This encoding of grammar can be formed either through working memory or automaticity, which is to say, the speaker will either utilize cognitive resources (working memory), or will rely on implicit memory stores (automaticity) to encode grammar.

Central to the speaking model presented above is the way

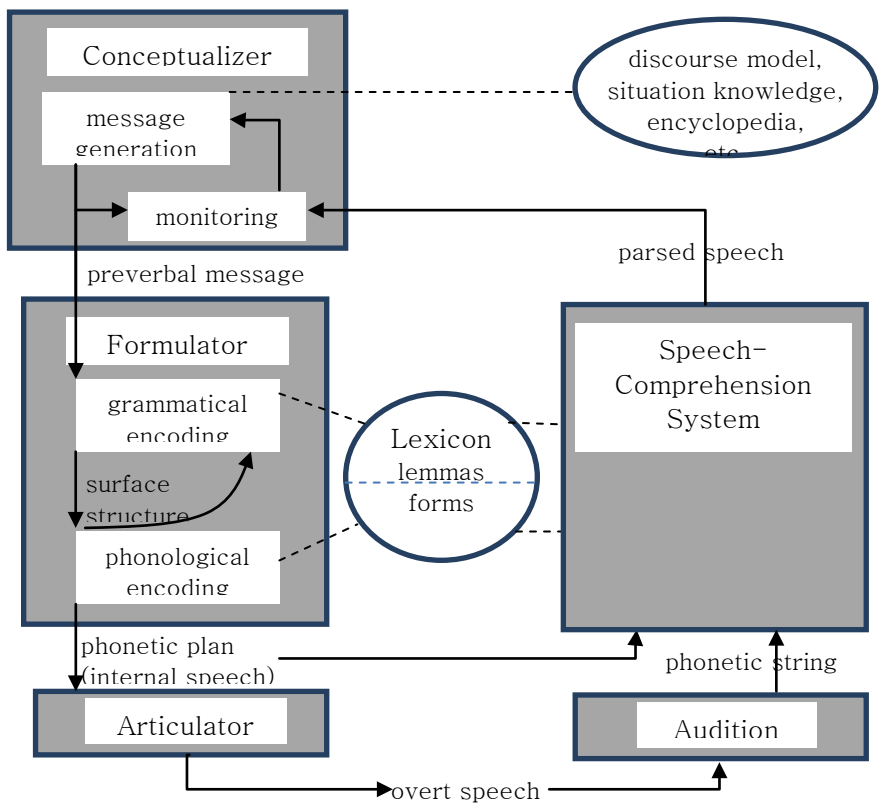


Figure 2.1 **Blueprint for the Speaker (Levelt, 1989)**

grammatical encoding interacts with working memory, which is where temporary attentive processes occur. Working memory is very limited in what it can process; only a few things can be processed at once, and therefore the working memory must be selective (Pienemann & Kessler, 2011). It is unable to give too much attention to grammar, as the majority of effort must go to propositional content. As a result, grammatical encoding must rely heavily on automatic processes, which are created incrementally from easier to more complex forms; thus, acquisition of grammar is automaticity (Håkansson & Norrby, 2007; Pienemann & Kessler, 2011). At the same time, it is not possible to learn complex grammar from the outset. In order for the L2 learner to acquire the grammar of the second language, the learner must break the grammar down into smaller, manageable pieces that will be acquired individually, thus explaining the series of grammars, or interlanguages, that a L2 learner will go through; these mark the gradual process by which the learner internalizes the L2 grammar (Pienemann & Kessler, 2011). The sequence of acquisition is determined by the nature of the processing procedure, according to how the learner breaks down the grammar (Ellis R, 2008; Pienemann & Kessler, 2011). A learner's reliance on simpler grammar may therefore be viewed as a type of solution, wherein the learner uses the minimum amount of grammatical knowledge necessary in order to communicate (Ellis N, 2008). The concept of incremental processing leads to the implicational hierarchy that is the core of PT. It also requires an interaction between explicit and implicit forms of knowledge

(Pienemann & Kessler, 2011).

Explicit knowledge is related to explicit learning; it is conscious knowledge that is only accessible through controlled processing, and is the type of knowledge that allows metalanguage (Ellis R, 2008). Explicit knowledge is what we tend to think of in relation to L2 acquisition, since much of L2 instruction (though not all) relies on metalanguage (Hulstijn, 2002). Implicit knowledge, by contrast, is automatic and therefore difficult to articulate consciously. It tends to be associated with L1 learning. Indeed, some theorists suggest it may only be acquired before a certain age, though this is an area of debate (Ellis R, 2008). Most language acquisition theorists agree that for L2 learning to occur, some aspect of language—particularly grammatical knowledge—must be transferred from explicit knowledge storage to implicit (Ellis R, 2008), or else that some other network of an implicit nature is created (Hulstijn, 2002). Certainly, implicit knowledge is required for the automaticity that is synonymous with language proficiency. Reliance on implicit knowledge is also an important aspect of PT, particularly in the types of tasks that can be used to support the theory (Baten, 2011). This will be elaborated further in Section 2.1.4 below.

2.1.2 Lexical Functional Grammar

Lexical Functional Grammar (LFG) is implemented in PT because it is compatible with Levelt's (1989) model of speaking, and because it is relatable to PT through feature unification (Håkansson, Pienemann, & Sayehli, 2002; Pienemann & Kessler, 2011). Furthermore, the use of LFG

makes it possible for PT to be extended to different languages, and for those languages to be compared (Håkansson & Norrby, 2005; Håkansson, Pienemann, & Sayehli, 2002). Finally, the psycholinguistic processes underlying LFG offer the framework for the stages and implicational hierarchy of PT (Di Biase & Kawaguchi, 2002; Pienemann, 1998). A brief explanation is given here of how LFG informs PT.

One of the key facets of LFG used in PT is the exchange of information between constituents required for L2 processing and acquisition (Pienemann & Kessler, 2011; Pienemann, 1998). This is based on feature unification, which is when functional (or feature) structures (f-structures) merge—for example, number agreement between a noun and a verb (Falk, 2001). Unification occurs at a variety of levels, all of which are integrated into the analysis of structure used by PT: lemma; phrase; clause; and sentence. Related to feature unification is Lexical Mapping Theory, which describes the link between argument structure (a-structure) and functional structure, or how semantic roles (argument) are expressed grammatically (function). One further set of structures that must be linked to function and argument is constituent structure (c-structure); this structure is the hierarchical component that represents the internal structure of a sentence, and is linked to f-structure through the Topic Hypothesis (Bresnan, 2001; Pienemann & Kessler, 2011; Dyson, 2008). The complex interaction of these three elements, represented in Table 2.1 below (in which morphology mapping and processing procedures represent feature unification) is what

allows a speaker to move from unmarked alignment to marked, and this in turn creates the context for the stages of PT (Pienemann & Kessler, 2011; Pienemann, 1998).

Table 2.1 Relationships among c-, a-, and f-structures (from Pienemann, 2011)

Morphology Mapping	Processing Procedures	Topic Hypothesis	Lexical Hypothesis
---	S'-procedure		Complex predicates
Inter-phrasal	S-procedure	Topicalisation of core arguments	Passive
Phrasal	Phrasal procedure	XP-adjunction	
Lexical Category	Procedure	Canonical order	Canonical order
None	Word lemma		

Before explaining the stages, it is important to understand the concept of unmarked and marked alignment in LFG and PT. Unmarked alignment is a direct relationship among the different structures, and therefore forms "the initial state of L2 development" (Pienemann & Kessler, 2011). An example would be a sentence such as "*John ate sandwich", a typical English SVO structure wherein John is simultaneously the subject, topic, and agent, fulfilling the basic roles for f-, a-, and c-structures respectively; this sentence would represent a basic structure that beginning English learners might be expected to make (note that this is not grammatically correct, but it *does* fit unmarked alignment for all three

structures). English sentences with unmarked alignment tend to have certain features in common: they are usually the most common; tend to be indicative and positive; have the broadest distribution, since they are appropriate in most or all contexts; do not use many pronouns; and tend to have more restrictions on subordinate clauses than on main clauses (Kroeger, 2004). Variations from these elements, through any of the structures, imply markedness, which requires an exchange of information across structures (Dyson, 2008). This exchange of information occurring on the constituent structure creates the stages for PT, since these represent the maximum amount of information that a L2 learner can process at a given point in the acquisition process (Jansen, 2008; Pienemann & Kessler, 2011; Pienemann, 1998). The framework is represented in Table 2.2.

Table 2.2 Developmental Stages (from Pienemann, 1998)

Developmental Stages
Stage 5—subordinate clause procedure
Stage 4—S-procedure
Stage 3—phrasal procedure
Stage 2—category procedure
Stage 1—lemma

The developmental routes along which a L2 learner acquires the second language proceed sequentially, following the speaking model set out

by Levelt (1989), along 5 stages delineated by PT (Jansen, 2008; Sakai, 2008; Pienemann, 1998). The first stage a learner must be able to process is the lemma stage, where a word is assigned meaning without any further context. Stage 2 is the category procedure; the lemma is assigned to a lexical category such as noun, verb, etc. Stage 3 is the phrasal procedure, where there is for the first time an exchange of information between the head and the parts of the phrase. The 4th stage is the S-procedure, having an interaction between elements at the level of the sentence, and the final stage, Stage 5, is the subordinate clause procedure, where there is an interaction among clauses (if it is applicable) (Pienemann & Kessler, 2011; Pienemann, 1998). This framework represents a strict and invariable order by which speakers process the language when they wish to produce a speech act (Dyson, 2008; Jansen, 2008). It is also important to recognize that these stages are not normative, or put another way, a learner does not necessarily have to produce the correct form in order to demonstrate knowledge of a stage; a learner who produces "*eated" as the past form of "eat" understands the grammatical concept of past tense, even if that learner does not know the correct lexical form that the past tense should take for that word (Håkansson & Norrby, 2008).[①] Furthermore, the stages are implicational; in order to acquire a higher stage, the learner must have acquired all of the stages below it (Jansen, 2008; Pienemann, 1998). It follows then that acquiring phrasal

[①] Though the theory does not make clear where correctly forming the past tense lies, we may guess it belongs to vocabulary. This remains unverified.

procedure necessitates first acquiring lemma and category procedures. These five developmental stages, and determining the markedness forms within each, comprise the basis for the implicational hierarchy of PT (Sakai, 2008; Pienemann, 1998).

2.1.3 Processability Theory's (Pienemann, 1998) Implicational Hierarchy for English

The developmental stages presented in Table 2.2 above present the gradual process by which a learner acquires a first language, but the process performs in the same way for a second language, with the difference that it now has a conscious element to it; beginning from Level 1, the learner first explicitly acquires the knowledge, but is later able to access the information through transfer to implicit knowledge resulting in automaticity (Pienemann, 1998; Norrby & Håkansson, 2007; Dyson, 2008). For English, the developmental stages convert into 6 levels for evaluating acquisition of both syntax and morphology, which are presented in Table 2.3. It should be noted that where syntax and morphology are at the same level, syntax usually creates the context for morphology, and is therefore assumed to be acquired before morphology (Pienemann, 1998; Jansen, 2008), though there have been some exceptions for this found in the data (Dyson, 2008). Nonetheless, these exceptions did not change the sequence of acquisition; merely the order in which morphology or syntax was acquired within one level.

Table 2.3 PT Levels of Development for English (from Pienemann, 1998)

Processing Procedures	**Information exchange + other principles**	**Morphology**	**Syntax**
6. subordinate clause-procedure	main & subordinate clause		cancel inversion *I asked when he could come home*
5. S-procedure	Topicalisation of core argument, information within S	inter-phrasal morph. (S) – SV-agreement (e.g. *Peter likes Mary*)	Do-2nd *Why does he like dogs?* Aux-2nd *When will she return?*
4. VP-procedure	information exchange within VP		Yes/no-inversion *Will she return?* copula inversion *Is he at home?*
3. phrasal procedure	information exchange within NP	phrasal morphemes NP agreement (e.g. *many dogs*)	Adv-fronting *(Then man sit on chair)* WH-fronting *(Why man sit on chair?)* Do-fronting *(Do man sit on chair?)*
2.category procedure	Unmarked Alignment—no information exchange	lexical morphemes Plural –s (*dogs*) *-ed* (PAST) *-ing* (PROG)	Canonical word order; SVO *(Man sit on chair)*
1. word/lemma access	word access, no information exchange	Words	---

The first level is lemma access, being words or word chunks before they have been assigned to categories. This is a difficult stage to pin down within PT. In fact, no studies conducted on PT have yet found a way to assess acquisition at this level; all have started with Level 2 or higher (Dyson, 2008). This causes complications for the theory which have neither been acknowledged nor redressed; any part of a theory that is not falsifiable cannot properly be called a theory (Popper, 1963). What makes this particularly concerning is that it is based on Levelt's speaking model, which is itself lexically driven (Gass & Selinker, 2008); this is a potentially major flaw. In order to maintain lemma as Level 1 in PT, a way must be found to test it. The second level is category procedure, which is the basic level at which studies begin, since this is the basis of unmarked alignment for morphology and syntax. At this level, there is no exchange of information among various elements. In morphology, this means that learners will be able to produce, for example, the plural form of a word (independent of contiguous words: i.e. learners will have knowledge of the form without necessarily having knowledge of appropriate context). Level 3 is where the first exchange of information takes place, within a noun phrase for morphology, whereas for syntax it is represented by adverb fronting or the formation of simple questions through putting "Do" or a Wh- word in front—note from the examples in Table 2.3 that the canonical order from Level 2 is maintained, with the exception of fronting; the addition of the adverb makes it marked. Level 4 is information exchange within a verb

phrase, which primarily involves inversion to form a yes/no question. Level 5 entails sentence level procedures, including subject/verb agreement for morphology, and the inclusion of “do” or auxiliaries in syntax, which builds on the fronting of Level 3. The final level is the subordinate clause-procedure, and for English this is cancel inversion in embedded questions. One important point is that these 6 levels represent the totality of what is analyzable; the theory can predict grammatical stages, but it cannot predict how development will occur within a stage (Baten, 2011). Are there stages, for example, by which one progresses in English from phrasal procedure to VP procedure? There are many other types of phrases, yet it is unclear how these fit into the developmental scheme, nor whether they all appear within the same stage. Thus, at the phrasal-procedure level (Level 3), it predicts that learners will be able to produce f-structure agreement within noun phrases, but not whether they will be able to produce other phrases such as adjective or preposition phrases. This is a potential shortcoming for diagnostic purposes, since the diagnostic information offered is very limited and, as will be seen in later sections, PT as it stands may ultimately prove more useful for determining placement levels rather than for diagnosing learner problems.

Another consideration is what, exactly, the levels are measuring; in other words, what is meant by “acquisition”? This is not an easy concept to define, and can even be specific according to individual teachers. PT measures acquisition in terms of emergence, which means the point at which

a learner spontaneously produces a certain form, or its first systematic use, but even this has no clear definition, and can be measured according to the number of times it is used accurately, or by the amount of time in between accurate usages (Gass & Selinker, 2008). PT theorists tend to define emergence in opposition to accuracy, since accuracy is viewed as a measure of consistency, whereas emergence is a measure of what the learner is trying to produce, or what base knowledge the learner has. In order to claim emergence for PT, most studies have adopted the measure of emergence as being 4 instances of correct usage within obligatory contexts (Pienemann & Kessler, 2011; Pienemann, 2005), though some measure it as a learner using a form correctly in 80% of obligatory contexts, depending on the type of task being used (Håkansson & Norrby, 2005). Several studies investigating PT have also found that the levels of the implicational hierarchy remain consistent, regardless of whether emergence or accuracy is used, suggesting a possible arena for future exploration of accuracy using PT (Baten, 2011; Håkansson & Norrby, 2010; Norrby & Håkansson, 2007). Of course, an essential requirement for determining emergence is the obligatory contexts and in particular, having an appropriate number of them. This requirement is crucial because the absence of a form in obligatory contexts shows non-acquisition, which, as the section on diagnostic tests will show, should be the primary object of study in diagnostic tests, and should be probabilistically determined in the psychometric modeling (Jansen, 2008; Sakai, 2008; Pienemann, 1998).

2.1.4 Tools for Measurement in Processability Theory

PT was originally intended for application only to speaking tasks, since it was felt that other types of tasks—including writing and multiple choice—allowed too much planning time, and therefore could potentially create a false positive② because test-takers tend to perform better on them (Håkansson & Norrby, 2007). The reliance on speaking tasks also fits somewhat more closely with current trends in testing toward communicative tasks (Ellis R, 2008). However, other researchers employed writing tasks using PT and found the output fit the PT hierarchy as well, particularly when the writing tasks were timed, which created some of the immediacy of speaking tasks while being productive (Håkansson & Norrby, 2010; Håkansson & Norrby, 2005; Chapelle *et al*., 2010). It should also be noted that writing tasks might be more accurate with low-level students, whose writing tends to be more like speaking anyway; even so, the results for all levels of learners were fairly accurate (Håkansson & Norrby, 2007). On the other hand, these same tasks have the inherent problem that was outlined in the introduction above; it is difficult to derive sufficient amounts of diagnostic information from them. This problem is compounded by the fact that PT only analyzes a relatively small number of grammar points within a potentially large collection of output, so that a great deal of effort goes into collecting a relatively small amount of information. This seems inefficient.

② False positives and false negatives will be discussed further in the section on diagnostic tests.

Another problem is that test-takers may be adept at avoiding certain difficult constructions, fearing that they could be penalized for inaccuracy more than for a lack of variety, even in informal testing situations (Ellis R, 2008). This does not mean the test-taker has not acquired the construction, even by emergence standards; the test-taker has merely circumvented obligatory contexts by employing avoidance strategies. This led to some researchers attempting multiple choice-style tasks, with mixed results; early studies suggested that multiple choice tasks are not compatible with PT, though the reason for this was unclear (Koizumi *et al*., 2012). Later studies, however, proved more successful when focusing on several important factors, especially keeping attention away from the grammatical structure that was being assessed through focusing test-taker attention more on content (which is also in keeping with the type of automaticity described in Levelt's model), and to making the test timed, much as was done in the writing tasks (Håkansson & Norrby, 2007). As will be seen in the Previous Studies section, changing the nature of the multiple choice tasks may also make them more productive for PT.

As was mentioned above, there is not yet any way to falsify Stage 1 of PT; however, the other stages are falsifiable, and therefore fit the classification for a scientific theory (Popper, 1963). Since PT employs an implicational hierarchy, there are specific quantitative methods that are required (Pienemann & Kessler, 2011; Pienemann, 1998). To begin with, group means may demonstrate overall patterns, but they can be misleading

at the individual level, since numerous individuals may display different patterns from the group as a whole and, depending on the number of variations, these may be lost within straight group means (Håkansson & Norrby, 2007; Ellis R, 2008). PT therefore calculates the scalability of the data by creating an implicational scale showing pluses for acquisition and minuses for nonacquisition, then counting exceptions. The calculation is the number of predicted cells (i.e. the total number of cells minus the exceptions) divided by the total number of cells to give the coefficient of scalability. In order to demonstrate a valid implicational analysis, this number must be over 90% (Pienemann & Kessler, 2011). Thus, calculating the accuracy of the implicational hierarchy requires developing a table similar to the one in Table 2.4. Note that, regardless of how extreme the exception may seem (Subject A, Level 4), this exception is still weighted as 1.

Table 2.4 Implication Scale and Calculation

Subject	Level 1	Level 2	Level 3	Level 4
A	+	-	-	+
B	+	+	-	-
C	+	+	+	-
D	+	+	+	+

Total # of cells: 16
Exceptions: 1 (Subject A, Level 4)
Coefficient of scalability:
15/16 = 0.94 or 94%

No research has been done into using other methods for calculating fit in PT. In fact, most work in validating this theory has been in confirming the developmental schedule across languages (beginning from Stage 2), and not on the many other predictions or claims. This is a shortcoming; properly

speaking, until all claims are subject to scientific testing, the theory remains weak. Tests must be devised and attempted for all the claims and predictions of PT. Subjecting a theory to falsifiability rarely destroys it, but rather strengthens it through verification and change (Kuhn, 1970). Therefore, PT—and all tests derived from it—will be better if subjected to falsifiability.

2.1.5 Learner Variation and Errors

In a sense, PT's approach to errors is similar to that of other forms of error analysis (e.g. Contrastive Analysis Hypothesis), with the exception that it has a specific system (Norrby & Håkansson, 2007). PT begins from interlanguage, which a learner designs to help solve the problem of making an utterance when the appropriate level has not yet been acquired, and when the resources in working memory are insufficient to give much attention to grammatical encoding (Håkansson & Norrby, 2007; Pienemann & Kessler 2011). According to PT, the learner has two ways of creating interlanguage: through deleting redundancy, or overuse of unmarked alignment (Pienemann & Kessler, 2011). Deleting redundancy is a common aspect of any language; English frequently makes use of pronouns, verb substitution, ellipsis, and other structures to avoid unnecessary repetition. Learners may likewise delete elements that might otherwise seem redundant, even though in English grammar they are not, such as eliminating one element in certain question inversions (see sentence 2 below). Or the learner may use the unmarked alignment in situations where it is inappropriate, such as maintaining SVO in question inversions (see sentence 3 below) (Pienemann

& Kessler, 2011; Pienemann, 1998). The possible outcomes are therefore:

Is she at home? (Target Sentence) (1)
She Ø at home? (Deleting Redundancy) (2)
She is at home? (Overuse of Unmarked Alignment—SVO) (3)

Although there may be some variation in what elements are deleted or overused, it is important to note that PT predicts: (a) all learners who have not acquired a stage will make one of these two general types of errors; (b) the learner will consistently make this type of error; and (c) depending on the type of error the learner tends to make, it could have detrimental effects on acquiring later stages, and may prevent further acquisition (Pienemann, 1998, 2005). As far as diagnostic testing is concerned, this method of analyzing errors is among the most productive aspects of the theory, since it suggests not only the error but, potentially, the learner's weakness, which is a major component of diagnostic testing. This is explored further below.

Another aspect of learner variation is L1 influence. Numerous theorists throughout SLA research have postulated some effect of L1 on L2 acquisition, though the degree of influence is not clear-cut (Alderson & Huhta, 2011). PT assumes that grammatical development and interlanguage are largely internal cognitive processes hardly affected by external factors, including L1. This is not to say that L1 has no effect whatsoever, but that the effect is dependent on those aspects of the grammar that are processable in the L2; otherwise, the interlanguage would become too complex and unmanageable (Håkansson, Pienemann, & Sayehli, 2002). The best option available to the learner, therefore, is to eliminate those L1 elements that

would make the interlanguage too complex, and focus instead on solving the language use problem through use of only L2 features (Pienemann, 1998; Pienemann & Kessler 2011). This hypothesis in PT is called the Developmentally Moderated Transfer Hypothesis, and it makes three predictions regarding L1 influence on L2 acquisition. First, even though a high-level structure in the L1 and L2 may be identical, it will not be acquired in the L2 until the appropriate level has been reached; thus, even if two languages formed verb phrases (Stage 4) in the same way, these would not be acquired in the L2 until all of the normal prerequisites for verb phrases have been acquired in that language—Stages 1-3. That said; the second prediction is that, when the learner has reached the appropriate level for that construction, the learner will easily acquire it because it is relatable to the L1. This means that, when the learner from above has acquired Stages 1-3 in the L2, verb phrases, which are similar in the two languages, will then be quickly acquired. Third, even in cases where the earliest structures between two languages are very different, the L1 structures will not be transferred to the L2 (Pienemann, 1998, 2011). In sum, the theory predicts limited interference from L1 in L2 acquisition; this is still a proposition that has not been proven.

2.2 Characteristics of Diagnostic Language Tests

This section was developed based on a list of 19 descriptors given in Alderson (2005) to build a general picture of what a diagnostic language test

should look like.[③] However, rather than look at the descriptors individually, they were grouped according to the general test characteristics that they address, and are discussed here under broader subject headings.

2.2.1 Conceptual Basis for Diagnostic Tests

In the early stages of diagnostic language testing research, most diagnostic tests were seen as being specific to some course of study, and many researchers still hold this view (Bachman, 1990; Kunnan & Jang, 2009). Certainly, when it comes to the usefulness of diagnostic information, teachers who need to utilize it in the classroom will likely have their own opinions about and approaches to language teaching, and may therefore find diagnostic information that is not directly related to their curricula to be not useful (Jang, 2009a). In this sense, the issue is less the particular course of instruction than it is the teacher's attitude toward language instruction and pedagogy generally (Jang, 2008). There is a more practical aspect to diagnostic tests being attached to a course of instruction though, since a diagnostic test for a course of study is more likely to be criterion-referenced than it is norm-referenced. After all, the purpose of a diagnostic test is not to rank the student in relation with others, but to understand the relationship between a student and a criterion, and what aspects of that criterion a student has not yet acquired (Blatchford, 1971). To say that a diagnostic test *must* be attached to a particular course or program, however, is too strong

[③] It should be noted that Alderson compiled this list based on information gathered from other research done on diagnostic language tests. The list has not been validated, but acts as a general guideline for researchers.

and limiting; any course of instruction is placed within the context of the education system as a whole, and thus a diagnostic test also relates to the general education system (Shohamy, 1982). Furthermore, a diagnostic test can be theory-based and even be norm-referenced—in fact, it may be more effective if it is (Alderson & Huhta, 2011; Alderson, 2011; Shohamy, 1982). Thus, tests based on a theoretical model could, with a deeper understanding of language ability, ultimately be more helpful for teachers than a test that is specific to a course of study (Kunnan & Jang, 2009). For these reasons, Alderson (2005) suggests either a theoretical or practical basis is appropriate for developing a diagnostic language test.

2.2.2 Focus on errors

The first two descriptors Alderson (2005) gives are based on test-taker strengths and weaknesses. Although most discussions of diagnostic language tests state that diagnostic tests should focus on strengths and weaknesses (Yin, 2011), the greatest information value is derived from the weaknesses, since it is the weaknesses that give direction to a future course of action (Sesli & Kara, 2012). There are three types of weaknesses that may be identified through diagnostic tests: student; instruction; and program. Student weaknesses can be general, referring to broad areas such as speaking or writing, or they can be specific, such as referring to an area of grammar, or even specific grammar points (Kunnan & Jang, 2009). Instruction weakness refers to a problem that the instructor may be having in communicating with a student or group of students, while program

weakness is a problem with the course of instruction as a whole (Shohamy, 1982). The difference between student and instructional/program problems is most readily observable in patterns of error across test-takers; for example, if one student in a class does not know the simple past form of “run”, that is likely a student problem. However, if an entire class does not know the simple past form of “run”, even when they know the past form of other irregular verbs, this indicates a problem with either the instruction or the program.④ Most diagnostic testing focuses on student diagnosis, as does this study, and finds instruction or program weaknesses incidentally to the main focus on students.

Other educational disciplines also use diagnostic tests to identify student weaknesses, so it is useful to examine at least a few of these to see what methods they use, in addition to reviewing those specific to second language acquisition (Alderson, 2005). For this study, 5 studies on educational diagnostic tests were reviewed in addition to several L2 language tests. Three of the tests in other disciplines were related to science (Arslan, Cigdemoglu & Moseley, 2012; Nehm, Beggrow & Opfer, 2012; Sesli & Kara, 2012); one was for L1 reading (Mokhtari, Niederhauser, Beschorner & Edwards, 2011); and one was for training teachers of foreign

④ In fact, a pre-pilot test for this study found precisely this problem—all of the students in one class, which was otherwise fairly high-level, did not know the proper past form of “run”. A quick look through several textbooks suggested that this had never been covered, even though “run” is a fairly common irregular verb. This problem therefore proved to be a program weakness, rather than student or instruction weakness.

languages (Richards, 2008). The most common characteristic among all of the studies reviewed was the focus on weaknesses, specifically, on errors and patterns in errors, and how to analyze these to reveal students' weaknesses. An important distinction must be made: a weakness is a problem with a student's knowledge or processing; an error is the observable outcome of a weakness. Therefore, we can speculate that it is not simply the error that is the object of investigation for diagnostic tests, but the *type of error*, since a specific type of error will reveal a specific weakness. These weaknesses can be identified as something the student has not yet learned (Simpson & Arnold, 1983), but they could also be misconceptions, meaning that the student has learned something incompletely or improperly (Nehm, Beggrow & Opfer, 2012). In the context of language acquisition, it may be possible to identify and differentiate between things that students have not learned and things that students have learned but do not know how to use properly, depending on how the test is constructed and the complexity of the skill required. The science tests in particular tended to utilize a two-, three-, or even four-tier system of testing in order to identify weaknesses through self-assessment (Arslan, Cigdemoglu & Moseley, 2012; Nehm, Beggrow & Opfer, 2012; Sesli & Kara, 2012). Tiered tests give insights into students' thinking in a similar way to how think-aloud protocols work. A two-tier item is a multiple choice task accompanied by a Likert-scale question asking the students about the confidence with which they answered. Three-tier items add a question

asking each student why they chose that answer—what the thinking was—and four-tier items include another Likert-scale question about the confidence of the students' thinking that led to the answer. These tiered task types give insights into how the students are thinking—where any problems may lie, and what thinking went into the problem. Though useful for examining students' thinking, it should be noted that these tiered methods are very involved when it comes to creating and interpreting the test scores; furthermore, these methods have primarily been tested on adult-level learners who may have a degree of self-consciousness not easily accessible for younger students such as were used in this study (Sesli & Kara, 2012; Nehm, Beggrow & Opfer, 2012). They have not yet been attempted in language tests. They are also not theory-based, but empirical methods; since this study focuses on creating a theory-based test, these methods were not used.

Several of the studies did mention the potential usefulness of theoretical models for acquisition, particularly those that involve hierarchies such as the Newman Error Hierarchy in math (White, 2005), hierarchies of language skills (Hulstijn, 2002), hierarchies within one language skill such as pragmatics or vocabulary (Håkansson & Norrby, 2005; Nation & Beglar, 2007), and implicational hierarchies for science (Simpson & Arnold, 1983). This is one area where PT strongly fits the template for a diagnostic test theory, since the implicational hierarchy it creates provides a suitable environment for diagnostic analyses. It also offers a method of analysis that

addresses not only the error but, to a certain extent, student weakness as well, making it potentially productive for diagnostic purposes. In the absence of a theory for acquisition or for analyzing errors, however, most studies relied on empirical methods to uncover patterns of thinking for why students make errors (Sesli & Kara, 2012; Nehm, Beggrow & Opfer, 2012; Simpson & Arnold, 1983). Qualitative methods of gathering information for diagnostic language tests are discussed further in the section on Quantitative and Qualitative Methods below.

Whether employing a theoretical model or qualitative methods, one of the most effective uses is in creating multiple choice-style tasks that, in some way, replicate common student errors in the distractors, thus revealing student weaknesses. Potentially the greatest benefit to developing such tasks is in validity. A common complaint about multiple choice tasks is that they lack surface validity because they do not have the same authenticity as other, integrated tasks (Bachman, 1990); however, when the distractors are based on authentic errors made by learners, it could add validity. But most importantly, these distractors have additional information value, because they tell the tester not only what the student does not know, but also point to reasons *why* the student likely does not know, i.e. the weakness (Richards, 2008; Yin, 2011; Simpson & Arnold, 1983).

Finding patterns of errors and understanding the thought processes that led to those errors gets to the heart of diagnosis, since they reveal the weaknesses in students' processing, and this is the information that is

actionable by all stakeholders, including students, teachers, administrators, and parents (Jang, 2008; Arslan, Cigdemoglu & Moseley, 2012). Focusing on errors in language assessment, however, requires a certain shift of focus, since the greater part of recent language assessment, being on proficiency, has focused on "Can-Do" statements (Bachman, 1990), but a focus on weaknesses is more likely to be described through "Can-Not-Do" statements (Alderson, 2011). However, "cannot do" seems too strong, and also too general. First of all, the errors that are addressed in diagnostic tests are a specific type of inability that is systematic, since they occur in recognizable patterns that have an identifiable source—as opposed to random errors, which are temporary, non-systematic, and may be caused by such factors as test environment or student emotional state (Bachman, 1990). We must therefore distinguish between systematic errors that identify student (or instructional, or program) weaknesses, and other errors that are external to students' language acquisition. Additionally, the weakness should never be seen as one that might never be overcome, as "cannot do" may imply (a substantial reason for focusing on "can do" in the first place), but as one which has not yet been overcome for some reason that is not fully understood or recognized. An analogy can be drawn to being in a restaurant waiting room—although you cannot eat while you are in the waiting room, it is not because you are physically incapable of eating, but because the conditions are not appropriate. Inability in such a context is a condition of the current state of affairs which may be remediable (once a table is

available). For these reasons, the current study proposes the term "*nondum* ability", or simply "*nondum*", from the Latin term *nondum*, meaning "not yet"—literally, "not yet ability". *Nondum* ability is defined for this study as an element which a second language learner has not yet acquired, whether because that element has not been taught to the learner or because of systematic misconceptions about that element, but which, *ceteris paribus*, that learner could still acquire. *Nondum* ability should be the object of any study on diagnostic tests. In PT, we can see learner error as important indicators of *nondum* ability.

2.2.3 Feedback

Alderson (2005) gives four descriptors relating to feedback, which is an important element of diagnostic assessment, since this is largely where the gap between tests and classroom instruction is bridged. Though not a major focus of this study, except in gauging teachers' and administrators' reactions to the grammar test developed, feedback still needs to be considered in any study on diagnostics, since properly validating a diagnostic language test requires developing meaningful feedback for all stakeholders (Yin, 2011). Feedback is the information given to the stakeholders that is intended to lead to change (Richards, 2008). Feedback must therefore not only have substantial information value, but should also be "translatable into instructional activities and actual strategies for teaching and learning. Thus, changes in the instructional system will take place in accordance with the feedback from tests" (Shohamy, 1982). These actions

and changes may include: aiding self-assessment and providing motivation for individuals; helping teachers individualize instruction for each student; giving teachers information for customizing courses; giving teachers feedback on the effectiveness of their teaching methods; and informing curriculum decisions (Yin, 2011; Bachman & Palmer, 2010; Alderson, 2005; Brown, 2004). Feedback should therefore ideally be accessible to a wide range of individuals with a diversity of interests and focuses (Richards, 2008).

The usefulness of feedback can be characterized in two ways: comprehensibility and applicability. Comprehensibility refers to how clear the meaning of the feedback is, and the types of information it includes (Kunnan & Jang, 2009; Donohue & Erling, 2012). A common complaint of much diagnostic feedback is that the information provided is couched in vague terms, such as “you have problems with vocabulary”, which superficially points to an error but does not give information that explains the weakness, and thus leads to no action (Jang, 2009a). Another problem is with evaluators that are poorly defined, such as “non-mastery” when the reason for the lack of mastery is not apparent (Donohue & Erling, 2012; Jang 2009b). Such feedback leads to uncertainty and frustration because a course of action based on it is not clear; often, a student knows when s/he has a problem, but does not know *why* (Jang, 2009b). To be actionable, the feedback must use specific and clearly defined terms that give underlying information about the cause of *nondum* ability. Applicability is the degree to

which feedback is actionable (Shohamy, 1992). Feedback must be applicable in that it leads to some change on the part of the stakeholders, but the degree to which each stakeholder may incorporate the feedback may vary (Richards, 2008). Students, who in some sense have the most control over what they will learn or what they want to focus on in their learning, will likely find feedback the most immediately applicable (Yin, 2011, Alderson, 2011). Teachers and administrators, on the other hand, are hampered by the fact that it may be difficult to incorporate feedback that has been developed for individual students into a curriculum that must be appropriate for the largest number of students possible; in order to be applicable for these stakeholders, the feedback may have to focus more on curriculum content (Yin, 2011). Either way, the process of validating a diagnostic language test would have to consider these aspects of feedback in order to justify the validity of the test.

2.2.4 Diagnostic test characteristics

As can be seen from the discussion on errors in Section 2.2.1 above, some form of multiple choice test tasks is desirable to give the maximum amount of information—both what and why—in as efficient a manner as possible. Multiple choice tasks are considered, in most respects, to be less authentic than other, communicative-style tasks (Purpura, 2004; Yin 2011); yet they are desirable here because of the information value that they have. Yin (2011) provides an apt analogy from medicine: a running test may more authentically mimic activities that one does in real life, but it cannot offer

the same diagnostic information that a comparatively inauthentic blood test can. Multiple choice tests also have other benefits, such as preventing biases through multiple choice tests' objectivity, as well as the fact that they will allow for quick scoring and feedback with the minimum amount of resources in human labor (though they require more effort to create than do some other task-types) (Richards, 2008; Norris & Ortega, 2003). Another aspect is that less authentic tasks often allow for more control over what is being tested; unlike in communicative tasks, test-takers cannot use avoidance strategies (Norris & Ortega, 2003). This was one of the major problems for PT discussed above; if a test-taker avoids a grammatical construction on a communicative task that would have been appropriate in the given situation, for example, it is impossible to tell from that whether the test-taker has acquired the form but is unsure or shy (an affective element unassociated with language ability), or does not know how to employ it properly, or has not been exposed to it yet. Thus, the diagnostic value of an avoided grammatical construction on a communicative task may be virtually zero. The fact that PT relies so heavily on productive tasks could therefore make it less valid for diagnostic purposes; a more ideal situation would be to combine the two forms—multiple choice and productive—in order to identify problems and the underlying errors in thinking. This was one of the reasons this study combined alternative task-types based on PT, which are described in the Previous Studies Section, with a writing task. Another was because multiple choice tasks can tend to lead to false positives when used

in isolation (Ellis R, 2008).

So far, these characteristics have referred to task types, but the test construction itself may be affected. An observation from attempts to retrofit proficiency exams is that the nature of these tests—being on a continuous scale and thus requiring items that vary from very easy to very difficult—makes them somewhat unsuited for diagnostic purposes (Jang, 2009b). In fact, the problem is the very easy and very difficult items; the majority of test-takers are expected to get the easy questions right and the difficult questions wrong, so these have little diagnostic value because, if a test-taker does get them wrong, it is difficult to determine why this may have occurred (Jang, 2009b). Thus, when it comes to designing diagnostic tests for specific classrooms, researchers can anticipate that either the majority (though not necessarily all) may be designed with a fairly narrow range of difficulties, relative to the test-takers, or that a specific group of test-takers will fall on a fairly narrow range of ability levels relative to the test. Outside of a classroom, or across a range of classrooms, a greater variety of difficulty levels may be observable on one diagnostic test. Therefore, a diagnostic test may have characteristics of either a criterion-referenced or norm-referenced test, depending on the scope of usage.

Two aspects not directly mentioned by Alderson but which are to a certain extent implied are the length of the test and the frequency with which diagnostic tests should be administered. It is suggested that diagnostic grammar tests should be long because of the range of contexts that must be

provided (Alderson, 2005; Alderson & Huhta, 2011), but this is also true of tests for other language skills (Richards, 2008; Blatchford, 1971). Regardless of the skill being tested, a variety of contexts must be provided, since a test-taker's ability may not be generalizable in all contexts (Nehm, Beggrow, Opfer & Ha, 2012). In the case of multiple choice tests, contexts are provided by the test, but in most communicative or integrated tasks, contexts must be provided by the test-taker, meaning that the tester must "wait" for a student to attempt a particular form; this may take considerably longer to gather (Norris, 2005). Thus, an examination of student writing for diagnostic purposes may take significant time and effort before an adequate number of contexts can be produced to recognize and diagnose a *nondum* ability; this is a potential weakness for PT. Gathering sufficient data is also a problem of reliability. If the test is to be sufficiently diagnostic, it must also provide a substantial amount of information; gathering that information, particularly if it is attempted in one test, would result in an instrument that is excessively cumbersome because it is too long (Blatchford, 1971). Therefore, rather than seeing a diagnostic test as one test, it may be more appropriate to view it as "a series of miniature tests on specific problems" (Shohamy, 1982). In this sense, a diagnostic test may be an accumulation of data gathered in a systematic fashion over a set period of time. This view of diagnostic tests necessitates regular administration, which is why they must be low-stakes. In this sense, writing tasks as required by PT may be more appropriate than multiple choice tasks, since writing tasks have less of a

surface similarity to “real” tests. On the other hand, simply the act of writing may cause anxiety; even more so in an age where the prevalence of integrated tests might make any productive task feel more like a high-stakes test situation. This requires further investigation.

2.2.5 Construct

Another important question for diagnostic language tests is the construct, or what the tests should be measuring (Alderson, 2005). Is it possible to make an adequate diagnostic grammar or vocabulary test? Are tests for language skills like speaking and reading easier to construct than those for language use, like grammar? And what is baseline or “normal”? The area of constructs—how narrowly they should be defined, whether a proficiency test is adequate for diagnostic purposes—is perhaps one of the most contentious areas (Alderson, 2011). The issue gets back to the question of the purpose for which a test has been developed; an integrated task focusing on proficiency could readily be employed for diagnostic purposes so long as the scoring rubric designed to evaluate it is sufficient for the purpose (Kunnan & Jang, 2009; Lee & Sawaki, 2009). In this sense, the fact that some tests such as speaking and writing may, on the surface, seem easier to construct, does not mean they will be easier to analyze since the same effort must go into *a posteriori* analysis of these tasks as goes into *a priori* analysis for multiple choice tasks (Sesli & Kara, 2012). In fact, when taking into account the requirements for constructing a diagnostic test, such as length, repetition, and the need for immediate feedback, speaking and

writing tasks that are designed to test proficiency may prove less efficient than other forms for diagnostic purposes. Thus, centering on low-level skills (grammar, vocabulary) could be a necessity, not only because higher-level skills (writing, speaking) are too complex and difficult to untangle but because low-level skills are more focused on *nondum* ability (Norris, 2005).

This is not to say that speaking and writing tasks, or even integrated tasks, cannot be used for diagnostic purposes; certainly, they can, depending on how they are analyzed. One possibility that emerges from the literature on diagnostic language assessment is splitting up the elements within a skill into different dimensions (Richards, 2008; Buck & Tatsuoka, 1996). This follows the example of diagnostic tests in other educational disciplines, which frequently rely on hierarchies to analyze errors (Simpson & Arnold, 1983). An example of this is the Newman Error Hierarchy for math, which delineates a process by which students solve a problem. Analyzing errors according to this hierarchy helps teachers identify precisely where an error occurs, and thus to direct assistance to a specific area of the problem, rather than on the problem as a whole (White, 2005). A hierarchy is also the cornerstone of the Vocabulary Size Test, which divides words into levels according to frequencies in corpus data (Nation & Beglar, 2007). In these hierarchical models, the issue is less one of analyzing low-level or high-level skills in order to determine acquisition, but more about finding the developmental sequence, if one exists, in which these skills are acquired. This is also important in understanding what is “normal” for an L2 learner;

empirical methods require creating an “everyman native speaker” to which learners are expected to fit, but a hierarchy creates a theoretical model for fluency. The latter is easier to identify and aspire to.

2.2.6 Quantitative and Qualitative Methods

One area that is not covered by Alderson (2005) is the quantitative and qualitative methods for measuring the reliability and validity of diagnostic language tests. Other educational diagnostic tests utilized a wide variety of qualitative methods to gather and analyze data for making tests, including interviews; written responses to questions; drawings; worksheets; word-association tests; concept maps; focus groups; written feedback from instructors and tutors; tiered tests; and classroom observations (Sesli & Kara, 2012; Mokhtari, Niederhauser, Beschorner & Edwards, 2011; Donohue & Erling, 2012; Richards, 2008). Other methods commonly used in developing language tests to help understand test-taker thought processes include think-aloud protocols, eye-tracking devices, audio and video recordings, timing devices, and others. These are all internal to test-takers, but other educational tests emphasize gathering information external to a learner that still affects the person, including the home and school environment; study practices; and even cultural identity groups and general social practices (Donohue & Erling, 2012). This information is largely gained through interviews, and such extensive information gathering turns each student into a case study of sorts (Mokhtari, Niederhauser, Beschorner & Edwards, 2011). This type of in-depth analysis is more practical for a teacher and

likely not feasible for developing a test, but it could be useful to keep these data-collection points in mind when designing a test and particularly when designing a study.

A large part of what distinguishes a diagnostic test from other forms of tests, as has been discussed at some length already, is not so much the way that the test has been formed—the task types used or the length of the test—but on the way that the test has been analyzed (Richards, 2008; Blatchford , 1971). Thus, any test, whether proficiency, achievement, or placement, has at least some diagnostic value, depending on how the researcher approaches the data (Lee & Sawaki, 2009; Blatchford, 1971; Jang, 2009b). A proficiency test can therefore give diagnostic information if it is analyzed in a fine-grained way; breaking down the items into their constituent elements, perhaps through Item Response Theory. Yet, the focus on determining weaknesses through errors brings up another important distinction; proficiency and achievement tests tend to focus primarily on ability, whereas diagnostic tests (and to an arguable degree, placement tests) tend more in the opposite direction, toward *nondum* ability. The assumption of ability is described in terms of the relationship it has with item responses, which relationship Lord (1952) defines: “the probability that an examinee will answer an item correctly is a normal-ogive function of his ability.” But what, then, is an error? And in particular, how are we to treat errors when each error has a meaning of its own, pointing to a difference source for *nondum* ability? The difference is in the object of study, and the question

relates to the quantitative methods used to analyze errors, as opposed to keys.

Another issue in diagnostic tests is that of false positives and false negatives, the former being when a skill appears to be acquired but actually is *nondum*, and the latter being the appearance of *nondum* ability when a skill has actually been acquired. These have differential consequences that will mostly be felt on individual learners. A learner who receives unnecessary remedial instruction may become bored and lose interest; on the other hand, a learner who does not receive needed remedial instruction may become frustrated and lose the motivation to study. These problems must be minimized in order to effectively enhance education, and therefore may require a complex model of analysis. For all these reasons, some researchers feel that classical test theory may not be adequate to the task of diagnostics (Blatchford, 1971; Shohamy, 1982), though there is some evidence that CTT can be effective for analysis (Richards, 2008). Other methods proposed for analyzing diagnostic tests include multidimensional IRT models, Bayesian procedures, regression-based approaches to subscale scores, and structural equation modeling (Blatchford, 1971; Stone, Ye, Zhu & Lane, 2010; Rupp, Templin & Henson, 2010). This question is well beyond the scope of this study (or the abilities of this researcher) to address in a direct way, but because they are not very well-understood, they should at least be introduced as part of the discussion.

2.3 Designing PT Task Types

This section focuses on two particular studies that utilized tests designed by using PT: “Using Developmental Sequences to Estimate Ability with English Grammar: Preliminary Design and Investigation of a Web-based Test” by John Norris (2005); and “Towards a Computer-delivered Test of Productive Grammatical Ability” by Chapelle *et al.* (2010). Both studies sought to develop tests for use in computer-based testing using developmental sequences to determine levels (Norris, 2005). Both studies were conducted on university-level students from a wide variety of backgrounds. It is important to note that these tests are *not* diagnostic tests, but are rather intended to be placement tests; this points to the potentially problematic aspect of PT as being less suitable for diagnostic purposes, and more suitable for placement purposes. Nevertheless, the tasks used in these studies offer an interesting option with potential to be developed for diagnostic tests, not just placement tests. Using these tests as a basis for designing a diagnostic test also gives some insight into the differences in constructing tests for these specific purposes and, more importantly, suggests a way PT may be adapted for use outside productive tasks. The placement tests had some features that were distinctly different from those of diagnostic tests as described in Alderson (2005), including the fact that there was no focus on learner errors or variation, as well as some differences in length and repetition of tasks. However, there are two major aspects that formed the basis for the current study. First, the types of tasks that were

created for the two placement tests imitated productive tasks while still maintaining qualities of multiple choice tasks, thus conforming to requisites for PT and diagnostic tests in an interesting and potentially useful way. Second, some of the grammar tested varied slightly from that of PT while still being based on principles of analysis set out in PT, particularly in the Norris (2005) study.

One of the starting points for both studies was the necessity to gain both a large volume and a wide variety of information from test-takers; the more valid and reliable information decision-makers have, the better informed they are for the decision-making process (Chapelle *et al*., 2010). In the case of Norris (2005), which was developing a test for online placement testing, a productive sample was unlikely to elicit the variety needed, and the test was developed as a possible substitute for a writing test. It therefore featured quite a few contexts, and a wide variety of them. In this respect at least, it more closely resembled the description of diagnostic tests given above. The test developed for Chapelle *et al*. (2010), by contrast, was being given as supplemental information accompanying a productive writing test for placement purposes; it used tasks that were similar to those of Norris (2005), but gave fewer contexts. Some tasks from Norris (2005) were sentence completions, either with pictures or embedded within a passage. Most tasks were designed as word-order tasks, in combination with sentence completion tasks. An example of one such task is given in Figure 2.2 below. The task resembles a productive task in that it requires the test-taker to

complete the sentence by putting the words in order, but represents multiple choice in that the test-taker must choose which elements belong in the answer, and which are distractors. It also attempts to imitate overuse of redundant elements proposed in PT as a specific interlanguage, since a learner may want to interject the referent subject "who" in front of "will". Chapelle *et al*. (2010) built on this design by Norris, adding several elements. To begin with, the tasks had more contextual content, which was a strategy that helped to distract from the grammatical element being tested (Norris & Ortega, 2003). Furthermore, some had a missing element that had to be added in or a word that needed to have a changed form, which imitates the redundancy error type of PT in which a learner tends to drop words or elements

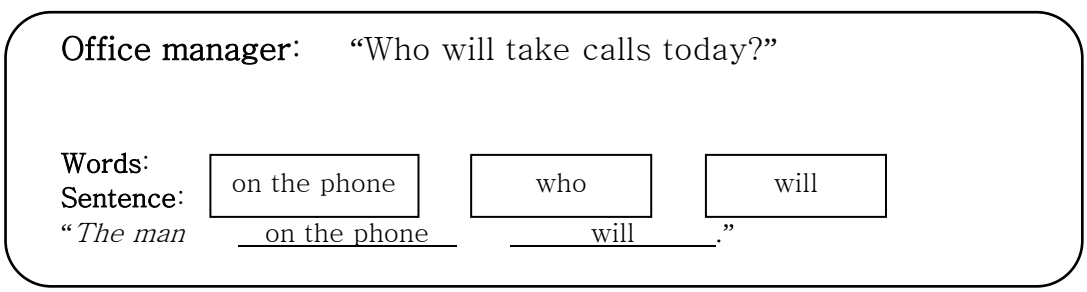


Figure 2.2 **Task Sample from Norris (2005)**

that seem to be unnecessary for one reason or another. Overall, the goal of such item-types is to first of all maximize the amount of information given through preventing any avoidance strategies; if a test-taker cannot produce the target response in the obligatory contexts, it is a clear indication by the

standards of PT that the structure has not been acquired (*nondum* ability). Second, this format would likely prevent test-takers from improving performance by learning from the tasks (Norris & Ortega, 2003). For these reasons, these task types were adapted for use in this study.

Another interesting element of both studies was the methods used for choosing the grammar tested. Although PT is intended to test certain very specific grammar points outlined in Table 2.3 above, Norris (2005) assumed that the top level, which tests relative clauses, might be further explored by utilizing different types from the embedded question grammar, which then "might tap even higher degrees of processing, and therefore help to distinguish among more advanced examinees" (Norris, 2005). Adding more levels would also be useful, since it may be difficult to elicit certain grammar structures; for instance, even in a speaking task with a native English interlocutor, it may be difficult to create sufficient obligatory contexts to demonstrate acquisition or non-acquisition of embedded questions (Hulstijn, 2002). Therefore, trying other relative clauses may be fruitful from an information-gathering point of view. Chapelle *et al*. (2010) went a little further, through using analysis of pre-pilot tests and researcher judgments to determine grammar points that may be productive on a test based on PT. Both of these concepts for expanding on the grammar tested by PT were explored in the current study.

Chapter 3

Method

This section looks at the methods for designing and piloting a diagnostic grammar test. It begins by describing the participants for the pre-pilot and main studies, followed by a detailed description of how the grammar test was designed, the design of a writing test for comparison purposes, data collection procedures, scoring, and finally, analysis of the data.

3.1 Participants

There were 219 students in the main study, plus 25 in the first pre-pilot and 42 in the second pre-pilot, all ranging in age from 10-16 and grades 3-9. This section describes the 219 students from the main study. Students were drawn from a variety of after-school educational institutes, and there was no common curriculum or program among them. All students are native Korean speakers, and come from mid-sized Korean cities outside of Seoul. None of the participants has lived or studied overseas for longer than a month. The number of years the students estimated they had studied English in Korea ranged from 1 year to 8; all started English studies in elementary school, usually grade 1. A brief questionnaire asked students what they felt they did best and worst. When it came to various language skills, students indicated that they tend to feel most comfortable overall with reading and listening, and least comfortable with grammar; attitudes toward speaking, writing, and vocabulary were in the middle. Students were frank

with their feelings about studying English: some said they like it because it is fun and interesting, some said they dislike it because it is hard, but the majority expressed mixed or even bland feelings; they accept it as something they have to study. These are pretty natural feelings, all-in-all. Finally, students were asked which of the tests for this study—grammar or writing—they found easier. Surprisingly, given what their teachers said about them preferring traditional grammar tasks (see Section 4.4 below), the majority said they found the writing test to be easier. When asked why, the majority wrote either “it was easier” or “I like writing”. Descriptive statistics for the participants of the main study are in Tables 3.1 and 3.2, divided according to gender and then grade. “Grammar Test” refers to the multiple choice test that was created for this study, while “Writing Test” refers to the short writing task that the students completed. Both tests are described further below.

Table 3.1 Descriptive Statistics for Students based on Gender

	N	Ave. Age	Grade Range	Grammar Test			Writing Test		
				Mean	StDev	Range	Mean	StDev	Range
Male	101	13.1	4-9	0.48	.20	0.04-0.90	3.4	1.8	0-7
Female	118	13.0	3-8	0.50	0.19	0.02-0.87	3.6	1.8	0-8
Total	219	13.0	3-9	0.49	0.19	0.02-0.90	3.5	1.8	0-8

Table 3.2 Descriptive Statistics for Students based on Grade

	N	% Girls	% Boys	Grammar Test			Writing Test		
				Mean	StDev	Range	Mean	StDev	Range
Gr. 3&4	8	50.0	50.0	0.56	0.15	0.35-0.79	3.9	1.2	2.5-5.5
Gr. 5	64	53.1	46.9	0.45	0.18	0.10-0.85	3.3	1.8	0-7.5
Gr. 6	89	59.6	40.4	0.50	0.20	0.13-0.87	3.3	1.8	0-8
Gr. 7	39	51.3	48.7	0.47	0.19	0.02-0.79	3.8	1.6	0-7
Gr. 8&9	19	36.8	63.2	0.58	0.22	0.04-0.90	4.2	2.4	0-7
Total	219	53.9	46.1	0.49	0.19	0.02-0.90	3.5	1.8	0-8

3.2 Instruments

This section gives the process used to combine the theoretical backing from diagnostic testing and PT described above, using the task-types developed in Norris (2005) and Chapelle *et al*. (2010). It begins with grammar selection, and then tasks and tests.

3.2.1 Choosing the grammar

The complete grammar points usually tested in PT are in Table 2.3 above. This study attempted to expand on that list of grammar points. The first stage in selecting grammar was to see which predicted grammar could be derived from a writing task. When initially creating a writing task for the study design, it proved exceedingly difficult to elicit any question forms predicted in PT in a standard, story-telling writing task. This was the case even when native speakers completed the task, and even when the picture

task seemed to specifically require questions. The writing task was effective in eliciting from native English and high-level ESL speakers all other forms of grammar pursued in this study, so the choice was either to create a second writing task that would be effective at encouraging writing questions—thus making an already difficult test and study design more difficult—or drop the aspect of questions from the test design. The latter was chosen. As a result, the final grammar test focuses primarily on the morphology from PT, as shown in Table 3.3.

Table 3.3 Grammar tested from PT

	Morphology	**Example**	**Syntax**	**Example**
Stage 6 Sent'				
Stage 5 Sent.	3sg-s	He eats		
Stage 4 V-Phrase				
Stage 3 Phrase	Pl-agree	Two cats		
Stage 2 Categ	Past –ed	Played		
Stage 1 Lemma			BWO	SVO

Note: BWO=*Basic Word Order*; SVO=*Subject, Verb, Object, or the basic word order for English.* 3sg-s=*Third person singular agreement using "s"*

Additional grammar points were added based on several aspects of PT and on grammar points that seem difficult for Korean ESL learners. Two

of these are determiners and non-count nouns, which should fit into the theory at the lower phrase (stage 3) and category (stage 2) stages, respectively. Another is prepositions and subordinating conjunctions, specifically “during/while”, “before”, and “after”. As prepositions, these tend to be misused (especially during), and as conjunctions, there is a strong tendency to be strongly influenced by L1 word order, which may generally result in Korean ESL learners’ overreliance on the word order:

Because [cause], [effect] (4)
While [event 1], [event 2] (5)

And an underrepresentation of:

[effect] because [cause] (6)
[event 1] while [event 2] (7)

One further problem related to that of subordinate clauses is verb tense, specifically combining past and past continuous in a sentence such as:

Susan decorated a cake while John was playing tennis. (8)

All of these grammar points were tested in the current study, and are presented in Table 3.4 alongside the grammar points from PT, in the predicted pattern that would most likely suit PT. Also, it proved impossible to test anything at the level of verb phrase, so Stage 4 remains blank.

Table 3.4 Grammar Tested for This Study

	Morph	Ex.	Syntax	Ex
Stage 6 Sent'			Sub. Conj. (while/before/after)	He worked while she was talking.
Stage 5 Sent.	3sg-s	He eats		
Stage 4 Vb Phr				
Stage 3 Phrase	Pres. Cont. Pl-agree Det-agree	He is walking. Two cats A cat	Prep. Phr. (during/ before/after)	…after school.
Stage 2 Categ	Past –ed Count/non	She played Furniture		
Stage 1 Lemma			BWO	We went home.

3.2.2 Grammar Task and Test Design

PT was originally designed based on Levelt's (1989) speaking model, and therefore focused in its early development on assessing grammar in speaking (Pienemann, 1998). However, subsequent studies have been conducted on writing that suggested the hierarchy remains stable within writing contexts as well, and is therefore applicable to both productive skills (Håkansson & Norrby, 2007; Pienemann, 2011). Thus, the test designed for this study relied on two other studies that applied PT to writing tests, Norris (2005) and Chapelle *et al*. (2010). In the first part of test design, a test was

developed using tasks that were almost identical to those of Norris (2005) and Chapelle *et al*. (2010), and this was pre-piloted to a group of 25 junior high school students fitting the profile of the participants described above. Although the tasks were similar to those of Norris (2005) and Chapelle *et al*. (2010), they differed at the test level in that there were a greater number of tasks in the overall test for the current study to provide the requisite number of contexts for each grammar point required by PT and diagnostic assessment. Also, since the test was being given to middle and elementary school students, the tasks were shorter, resembling Norris's (2005) tasks more so than Chapelle *et al*.'s (2010). Also, in order to minimize semantic problems, the tasks focused entirely on concrete nouns and action verbs, as these are usually the first taught to beginning learners. Even so, this first pre-pilot proved too difficult for students; in particular, the tasks were too unfamiliar, and the test "leapt" a bit in difficulty from section to section, thus proving too challenging for most students. Many were unable to complete the test. Based on the feedback from the first pre-pilot, a second test was created and a second pre-pilot given. The second test added some item types in order to create a more gradual progression through the grammar points and task types. The results of the second pre-pilot test were far more satisfactory. The only major change made for the final version of the test was to increase the number of tasks per section, in order to give 5 to 6 contexts for each grammar point~~s~~ instead of 4. It should be noted that the test was given to native English speakers at all stages to ensure that the

desired target grammar was the most appropriate solution to each task.

Another element that was explored was how to focus test-takers' attention away from the grammaticality of the task and onto content, which would better support a claim that the test accesses implicit knowledge (Ellis R., 2008). This was done in a variety of ways, including using pictures. An example of one such method is given in Figure 3.1, developed for the grammar test in this study. The goal of this task was to focus the test-takers' attention on content—in this case, the number of objects in the picture—rather than on the fact that the task is testing knowledge of the plural form of the noun.

1. The woman walked with ___two dogs___ yesterday. (dog)

Figure 3.1 **Task Sample for the Current Study**

Finally, the test design employed a time constraint, which has been suggested to increase the likelihood that a task will access implicit knowledge (Richards, 2008; Baten, 2011). The two pre-pilot tests were used to set the time limit for the final version of the test. Since the test was timed, the instructions were given in both English and Korean, so that the time constraint would be testing grammatical ability, not the ability to read instructions. It should also be noted that this is one area where using computers, as suggested in Alderson (2005), might have been most

beneficial in this study; since the entire test was timed, some students may have taken longer on earlier sections that tested low-level grammar, and may thus have drawn more on explicit knowledge in these sections, rather than implicit. The final version of the grammar test, along with the instructions in both languages, is given in Appendix A.

3.2.3 Writing Task Design

A writing test was designed originally for comparison on PT, but later proved necessary for the purposes of a general proficiency score. The writing task was designed in part through work on an earlier study on complexity, accuracy, and fluency (CAF), using studies by Tavakoli and Foster (2011) and Ong and Zhang (2010). The tasks in those studies were designed in such a way as to give the maximum opportunity primarily for complexity and fluency, though not necessarily accuracy (Ong & Zhang, 2010). This was done through picture tasks that gave test-takers opportunities for increased complexity through +/- multiplicity and +/- fluidity, where multiplicity refers to how many activities are going on in one picture frame (foreground and background activity), and fluidity refers to the continuity among pictures. Adding or subtracting these elements can increase complexity within sentences and within the composition as a whole (Tavakoli & Foster, 2011). For this study—in which complexity in particular was sought—the writing task was designed in such a way that background action frequently occurred at the same time as the foreground action was occurring (+multiplicity), which was intended to encourage production of

subordinate clauses using “while”, “before”, and “after”. Furthermore, there was not always clear continuity/relationships among the pictures (-fluidity). These are devices intended to elicit output of greater complexity, if the writers are able to produce it, but to still offer the opportunity for lower-level learners to create output (Tavakoli & Foster, 2011). The writing test was first piloted on English L1 adults and high-level ESL learners, and produced more than enough contexts for all of the grammar points given in Table 3.3. However, since the writing test was meant primarily to test grammar and not vocabulary, the test-takers were given a list of vocabulary words to help them. The final version of the complete writing test is given in Appendix B.

3.2.4 Feedback

Although feedback was not a focus of the study, some feedback was offered to students in exchange for participation in the study. Two different forms of feedback were given to the schools, one in paragraph form and the other more like a report card (see Appendix F for examples). Administrators and teachers reviewed the feedback and were interviewed to ascertain which type they preferred, and what information they took from the feedback.

3.3 Data Collection Procedures

In keeping with Alderson’s (2005) characteristics, the tests were administered in an environment that would be comfortable and familiar for students, which was at their after-school institutes. All students were assured that this would not affect their grades, and instructions were given in Korean

and English in order to further reduce stress. Students were also allowed to ask questions during the test if they were confused about the format of a question, and teachers were instructed to give explanations without giving away any answers. Teachers reported that few students asked questions. The test would ideally be given in a computer-based format, as per Alderson's 19th characteristic; this would have allowed for more control over timing, as well as quicker scoring. However, as not all schools had equal access to computers, the test was a paper-and-pencil format. The first part of the test was the writing task, and the second part was the grammar test. This was to ensure that there was no undue influence on the writing task from the grammar test; students may have been tempted to use some of the forms from the grammar tasks in their writing. Although this effect was not observed on the pre-pilots, it seemed a reasonable precaution. Each test had a cover page with the student's name, and the last page was a brief questionnaire (which was in Korean). Students were instructed to only write their names on the cover page, and not on any other pages, in order to maintain privacy.

3.4 Scoring/Rating the Grammar and Writing Tests

Scoring was done in two parts. First, the tests were scored solely for the grammar points that were the focus of this study. To this end, a scoring page was created in order to express the different aspects of PT, and especially to give the maximum amount of information possible about errors. There were 6 raters for the first rating, all native English speakers with some

experience in teaching ESL; some are still teachers, but most currently work in academic publishing as writers and editors. Four of the six have extensive experience in language assessment, both in creating general proficiency tests, and scoring speaking and writing tests. Four were females, two males, ranging in age from 31-41. Raters were given a training session during which they were informed about the theory and provided sample tests to score together with the researcher. Since both the grammar and writing tests had absolute scores for the grammar, meaning that there were binary responses (present or not present), each test was rated only once, and then checked by the researcher to ensure the raters understood the method. Raters had the option on the grammar test to mark a problematic response as "incorrect", "missing", or as "non-target" for when an item was grammatically correct, but did not match the target answer in some way. These were counted as "incorrect" for the total score, but freed the raters from the psychological strain of marking something as incorrect when it had no grammatical errors but was merely not the desired response. On the writing test, raters counted the number of times a test-taker attempted to use a grammar point, and whether it was correct or incorrect. Then, for both the grammar and writing tests, the raters had to give a brief description of the nature of the error: a missing word or words; words that were added unnecessarily; or other problems such as tense or word order. All of these elements are reflected in the scoring sheet, in Appendix C.

One point that should be noted about the scoring was that one

section, Section 6, was primarily intended to test SVO word order. According to PT, this should be the starting point of acquisition for all ESL learners, and is therefore the least marked form. This proved to be the case, as all students got the items in this section correct for SVO word order. There was a small number of students (16) who made mistakes, but these were not with word order; thus, it was safe to say that all students had, indeed, acquired SVO. This section did not figure in the results section of the study, since it is impossible to correlate when every student gets the answer correct.

The second part of scoring was done solely to the writing test, and was a holistic score on an 11-point scale (ranging from 0 for insufficient writing sample (one sentence or less), to 10 for superior performance) which was developed by the researcher based on the ACTFL guidelines (2012). This scoring was done to give an overall proficiency score for each student. Each writing test was rated for overall proficiency by two raters, and any discrepancies over 2 points were rated again by a third rater. The scoring rubric is given in Appendix C.

3.5 Data Analyses

The data were first entered into MS Excel, which was used to calculate means and other descriptive statistics, and then transferred to SPSS 21 for more complex analyses. Investigating the reliability of the grammar test was done through Cronbach's Alpha, and inter-item correlation and covariance matrices. The writing test scores were checked for inter-rater

reliability using various indices such as agreement rules, Cohen's Kappa, and inter-rater correlations. For the performance at the level of items, item difficulty and discrimination were calculated. To determine the relationships among the sub-test, test, and proficiency (writing test) scores, correlations were calculated. All of these were done on SPSS. For calculating PT, the standard is to create an implicational scale and calculate the coefficient of scalability, which was done for both the grammar and writing tests, the latter for comparative purposes. These analyses were done on Excel, as were all graphs and tables presented in this paper.

The standard method for representing an implicational hierarchy is to list the subjects in the first column, in order of most levels acquired to least. Since representing 219 subjects is a bit cumbersome for this paper, a different method was used, as illustrated in Table 3.4. The first column shows the highest level that a subject has acquired, regardless of whether the subject acquired the levels in order, and what percentage of students acquired that level. Gaps in a level show that there were subjects who did not acquire the lower levels before acquiring the higher, and therefore did not fit the implicational hierarchy. In the hypothetical chart offered below, it was possible to acquire 3 grammar points. The first group of students acquired the 3rd grammar point; however, not all acquired all 3 levels. The first row within that group shows the number who acquired all 3 levels; the number is given in the second last column, followed by the percentage of those who acquired level 3 that the number represents. The next few rows

show the exceptions, gaps being levels that the subjects did not acquire, followed by the number and percentage of the group. The next group acquired level 2 as its highest level and includes the non-fit subjects; the final two groups are those that acquired level1 and no levels. In the results section, one chart is given for the grammar test, another for the writing test.

The columns within the middle levels section show the fit and non-fit patterns of acquisition for the grammar points. The order of the columns is determined by the number of students which acquired that particular grammar point. So, in the chart below, Column A was the grammar point acquired by the greatest number of students, while Column C was the grammar point acquired by the fewest students. These levels equate to stages in the implicational hierarchy, and should ideally occur in the predicted order of the implicational hierarchy. In the results section, a key is given for each chart to show the order of acquisition of the grammar points in order to make comparison of the test results easier.

Calculating the coefficient of scalability for this table requires counting the number of exceptions. In this case there are 15 (Each gap square times the number of students for that square). There are 40 subjects (Total for N=40), and 3 levels, so that means there is a total of 120 squares possible. Subtract the exceptions from the total (120-15=105), then divide that number by the total to get the coefficient of scalability: 105/120=87.5%. Therefore, this particular table would *not* meet the criteria for fitting the implicational hierarchy, as set out in PT (>90%).

Table 3.5 Example of Implicational Hierarchy Design for this Study

	A	B	C		
	■	■	■	5	35.7%
	■		■	4	28.6%
Level 3			■	3	21.4%
35%		■	■	2	14.3%
Level 2	■	■		9	69.2%
32.5%		■		4	17.4%
Level 1	■				
27.5%	■			11	100.0%
0 Levels					
5%				2	100.0%

Chapter 4

Results

4.1 Descriptive Statistics for the Grammar and Writing Tests

The descriptive statistics for the grammar test are given in Table 4.1, for two different forms: Version 1 is the test as it was originally designed; Version 2 is the test minus the first subsection (items 1-10), which tested non-count nouns and determiners with count nouns. This latter was done *post-hoc*, after viewing the results of the initial analysis. There were several items that had exceptionally poor correlations, the most notable being in Subsection 1, the items of which tended to have extremely low and even negative correlations with items from other subsections and even from items within the same subsection. This was the main factor in the decision to complete calculations for two different forms of the test: one complete, and one without Section 1.

Table 4.1 Descriptive Statistics for Two Versions of the Grammar Test and Writing Test

	N	Items	Mean	SD	Median	Mode	Range
Version 1	219	52	25.6	10.1	25	15	1-47
Version 2	219	42	20.3	9.0	20	19	0-40
Writing	219	1	3.0	1.8	4.0	4.5	0-8

Inter-item correlations are given in Appendix D. These were, for the most part, moderate to low; the highest correlations tended to be within one

section, as would be expected. It is also somewhat expected that items across sections may not correlate as well, since they are testing different grammar points.

Descriptive statistics for the holistic scores on the writing test are also given in Table 4.1. In addition to the holistic scoring, complexity scores were calculated for each composition based on t-unit analysis (Hunt, 1965). These statistics are given in Table 4.2. The analysis shows that the compositions were generally very short, with the average around 68 words and 11 t-units (an independent clause and its dependent clauses) per composition, and few dependent clauses per t-unit (about 11% of t-units had a dependent clause). Of these dependent clauses, about 19% were the target subordinate clauses (using while, before, and after). The rest used when (43.1%), and because (37.6%). There were no other variations present in the data. Another point that should be noted is the placement of the subordinate clauses: 77% of the subordinating conjunctions appeared in the beginning of the sentences, and only 33% in the middle. There was some variation among conjunctions, with “because” appearing almost equally in both positions (53.5% in the beginning of the sentence; 46.5% between two clauses); however, over half of those in the beginning of a sentence were incorrectly used (52.6%). The most common error was for “because” to clearly appear at the beginning of a sentence, but in a way that should have conjoined two clauses: “Today I late school. Because my pets chase to me. And I fall down.” The rest of the conjunctions overwhelmingly appeared in the

beginning of the sentence: 91.5%, with only 8.5% appearing in between two clauses.

Table 4.2 Textual Characteristics of Essays Written by Participants

N	Ave. Word Count	Range Word Count	Ave t-unit Count	Words per t-unit	Words per Clause	Clauses per t-unit	Target Clauses
219	**67.83**	**0-242**	**10.78**	**6.30**	**5.69**	**0.11**	**0.19**

4.2 Reliability Statistics for the Grammar and Writing Tests

Cronbach's Alpha was calculated for each subsection of the test, for the test as a whole, and for the test without the first section, which was done because the first section (determiners and non-count nouns) performed so poorly overall. The overall reliability estimate is quite good, though there is room for improvement, particularly in the later sections of the test.

Table 4.3 Score Reliability Coefficients for each subsection, the whole test, and the test without section 1

Section*	Det	NC	PN	Past	PrC	SVsg	SVpl	Prep	Sub Clause			SCT	Test	PTest
									A	B	C			
Number of items	5	5	5	5	5	6	4	5	4	4	4	12	52	42
Alpha score	0.18	0.7	0.88	0.85	0.93	0.92	0.73	0.76	0.73	0.74	0.61	0.83	0.92	0.93

**Note*: Det=*Determiners*; NC=*Non-count nouns*; PN=*plural nouns*; Past=*Past tense*; PrC=*Present continuous*; SVsg=*Singular subject/verb agreement*; SVpl=*plural subject verb agreement*; Prep=*prepositions*; SubClause=*Subordinate clauses, Sections A, B, and C*; PTest=*test without Section 1*

The writing tests were scored holistically to give a common, external measure of proficiency. Five raters were used from the original scoring. Inter-rater correlations were quite high, though Kappa was a little low. This likely occurred because only about half of the scores were of perfect agreement, while an almost equal number were of adjacent agreement, differing by 1. Two scores differed by 2 points, even after being re-rated by a third person. Spearman's rho and Cronbach's alpha values were both quite high. Overall, the holistic scoring of the writing test was adequately strong.

Table 4.4 Inter-rater Reliability Statistics

	N	Correlation	Kappa	Perfect Agreement	Adjacent Scores	Perfect+ Adjacent	Rho	Alpha
Writing Test	219	0.92	0.41	0.49	0.49	0.99	0.91	0.96

4.3 Performance of Items

Item difficulty and discrimination indices were calculated. Table 4.5 is presented on the following page. Item difficulty was calculated by subtracting the mean for the task from 1, and item discrimination through corrected item-total correlation. Figure 4.1 shows the data in graph form. Among the most discriminating items were those testing plural nouns, past tense, present continuous, and singular subject verb agreement, as well as some of the items testing subordinate clauses. The items with the poorest discrimination all came from the first subsection testing non-count nouns and determiners with count nouns: in fact, as the chart shows, one item (2b)

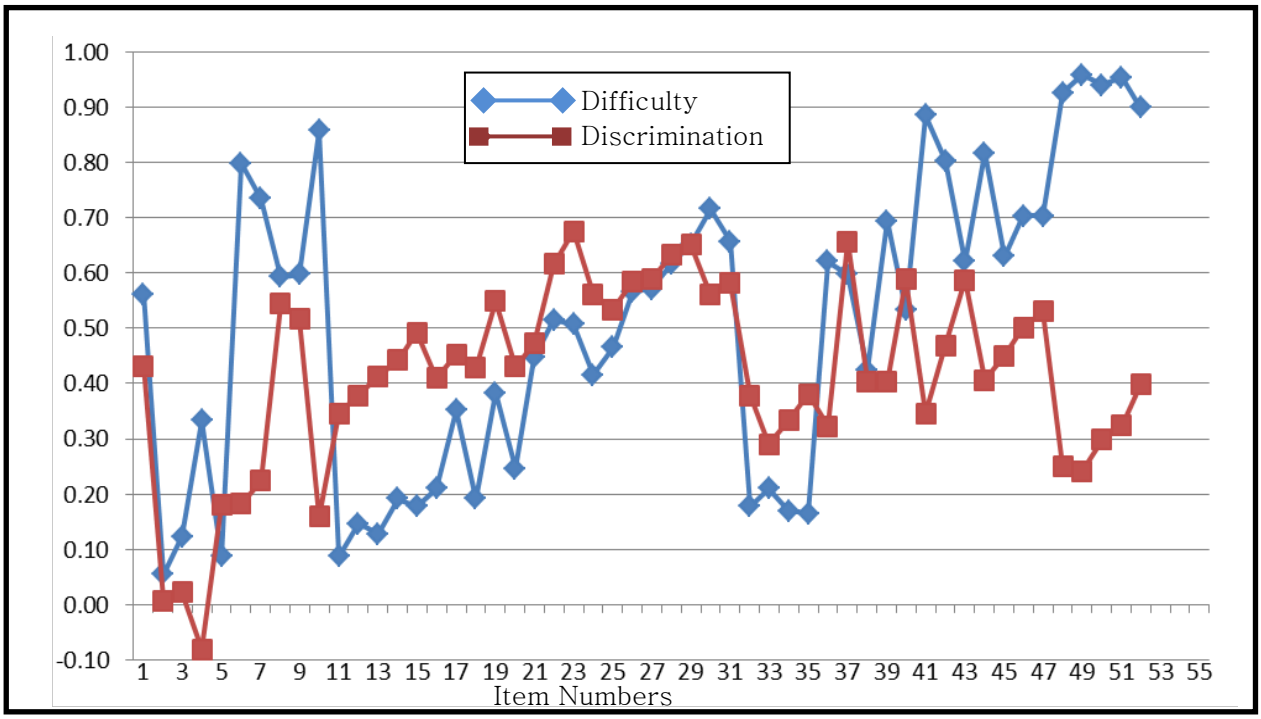


Figure 4.1 **Item Difficulty and Discrimination (Corrected Item-Total Correlation)**

had negative discrimination. This is in keeping with other findings for this subsection found in the analyses done above. The last few sections, which all tested subordinate clauses, had lower discrimination rates; however, it should be noted that a number of students did not finish these sections. One final note is that the two easiest—and least discriminating—subsections in both forms are creating the plural form of nouns and plural subject/verb agreement. This observation is in keeping with Processability Theory, which predicts that these two forms will be acquired early by all ESL learners.

Table 4.5
Item Difficulty and Discrimination Matrix

	Item Difficulty	Discrimination	Alpha if Item Deleted
Det1	0.56	.432	.925
Det2	0.05	.007	.927
Det3	0.12	.022	.928
Det4	0.33	-.082	.930
Det5	0.09	.181	.927
NC1	0.80	.182	.927
NC2	0.74	.224	.927
NC3	0.59	.544	.924
NC4	0.60	.516	.924
NC5	0.86	.159	.927

Section 1 Only

	Item Difficulty	Discrimination	Alpha if Item Deleted
PlurN1	0.09	.352	.928
PlurN2	0.15	.374	.928
PlurN3	0.13	.418	.928
PlurN4	0.19	.459	.928
PlurN5	0.18	.491	.927
Past1	0.21	.435	.928
Past2	0.35	.481	.927
Past3	0.19	.457	.928
Past4	0.38	.569	.926
Past5	0.25	.445	.928
PrC1	0.45	.471	.928
PrC2	0.52	.625	.926
PrC3	0.51	.673	.925
PrC4	0.42	.570	.926
PrC5	0.47	.549	.927
SVSg1	0.57	.583	.926
SVSg2	0.57	.590	.926
SVSg3	0.62	.628	.926
SVSg4	0.65	.643	.926
SVSg5	0.72	.553	.927
SVSg6	0.66	.572	.926
SVPl1	0.18	.363	.928
SVPl2	0.21	.285	.929
SVPl3	0.17	.339	.929
SVPl4	0.16	.392	.928
Prep1	0.62	.315	.929
Prep2	0.60	.649	.926
Prep3	0.42	.393	.928
Prep4	0.69	.407	.928
Prep5	0.53	.588	.926
SCA1	0.89	.345	.928
SCA2	0.80	.479	.927
SCA3	0.62	.578	.926
SCA4	0.82	.402	.928
SCB1	0.63	.448	.928
SCB2	0.70	.504	.927
SCB3	0.70	.512	.927
SCB4	0.93	.245	.929
SCC1	0.96	.242	.929
SCC2	0.94	.302	.929
SCC3	0.95	.318	.929
SCC4	0.90	.402	.928

The remaining sections, after Section 1 was removed from the test.

4.4 Comparison of Subsections, Total Score, and External Measure

The score derived from the writing test was used as a proficiency score, and this was compared with the subsections of the test as well as with the test as a whole. The correlation matrix is given in Table 4.6. Disattenuated correlations are given in Table 4.7; the correlations went up slightly when corrected for attenuation, which is expected. Since the test was a grammar test, it would not likely correlate very highly with a writing test; thus, a score of 0.61 (0.67 after correcting for attenuation) is acceptable (Hatch & Lazaraton, 1991). This correlation was also higher than the correlations for any of the subsections with the writing score. The subsections correlated in a fairly moderate band with both the overall grammar (0.51-0.75) and writing scores (0.31-0.50), with the exception of determiners, which correlated comparatively poorly with both the grammar (0.32) and the writing scores (0.23). Though the non-count nouns correlated moderately well with both (0.57 and 0.34 respectively), the performance of determiners is yet another indication that there was a problem with the first section of the test. Apart from that, one other section that did not perform well was the third part of the subordinate clause subsection, which had a lower correlation with the grammar and writing tests than did the other two subordinate clause sections (0.51 with grammar and 0.31 with writing, versus the other two subordinate clause sections, which were 0.67 and 0.64 with grammar and 0.42 and 0.46 with the writing score). Overall, the total (aggregate) score for the subordinate clauses tended to correlate most highly

Table 4.6 Correlations among Subsection, Test, and Proficiency Scores

	PlN	Past	PrC	SVSg	SVPl	Prep	SubCl				Test Total	Writing Score
							A	B	C	Tot		
PlN	1											
Past	.37**	1										
PrC	.29**	.34**	1									
SVsg	.28**	.42**	.43**	1								
SVpl	.38**	.36**	.27**	.25**	1							
Prep	.28**	.33**	.46**	.45**	.26**	1						
SCA	.21**	.28**	.40**	.40**	.26**	.53**	1					
SCB	.23**	.34**	.27**	.38**	.23**	.53**	.56**	1				
SCC	.15*	.18**	.26**	.39**	.11	.39**	.50**	.42**	1			
SCT	.25**	.34**	.39**	.48**	.26**	.60**	.87**	.86**	.69**	1		
Test	.55**	.65**	.70**	.75**	.51**	.73**	.67**	.64**	.51**	.76**	1	
Writing	.36**	.43**	.44**	.37**	.33**	.47**	.42**	.46**	.31**	.50**	.61**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

of all the subsections with both the grammar and proficiency scores. One final point is that the correlations were recalculated without the first section of the grammar test, and all correlations of the subsections with the grammar test went up; however, the correlation between the total grammar test score and the proficiency score went down slightly, from 0.624 to 0.617. This is a puzzling and unexpected result, though likely not significant.

Table 4.7 Disattenuated Correlation Scores

	PlN	Past	PrC	SVSg	SVPl	Prep	SubCl				Test Total	Writing Score
							A	B	C	Tot		
PlN	1											
Past	0.43	1										
PrC	0.32	0.38	1									
SVsg	0.32	0.47	0.46	1								
SVpl	0.48	0.46	0.33	0.30	1							
Prep	0.35	0.51	0.54	0.54	0.35	1						
SCA	0.27	0.36	0.48	0.49	0.35	0.71	1					
SCB	0.28	0.42	0.33	0.46	0.31	0.70	0.76	1				
SCC	0.20	0.25	0.35	0.52	0.17	0.57	0.75	0.63	1			
SCT	0.29	0.41	0.44	0.54	0.34	0.76	1.00*	1.00*	0.97	1		
Test	0.61	0.74	0.76	0.81	0.63	0.88	0.82	0.78	0.68	0.87	1	
Writing	0.40	0.48	0.47	0.40	0.41	0.56	0.51	0.56	0.41	0.57	0.67	1

**Note: These scores were over 1.00. Disattenuated scores exceeding 1 indicate that "measurement errors are not randomly distributed" (Schumacker & Muchinsky, 1996).*

4.5 Responses to Questionnaires and Interviews

Among the greatest limitations for this study was the lack of interaction the researcher was allowed with teachers and students; due to rigorous academic schedules for both at after-school institutes, institute administrators highly restricted the researcher in the number of questions that could be asked of students in questionnaires, and would only allow interviews with teachers, not questionnaires. In this regard, the students

generally thought the grammar test was more difficult than the writing test (65.3% thought the writing test was easier). The teachers generally felt that the grammar test would be fairly easy for the students, and were somewhat surprised by the rather low scores many received. All rightly predicted that the writing scores would tend to be fairly low; as one teacher noted, Korean students usually do not like writing, and avoid it where possible. The institute administrators gave more comments; overall, they thought the tests were an appropriate level of difficulty, though they were also surprised that the students did not do as well, overall, as expected. The raters generally felt that most of the grammar test was an appropriate difficulty, though they felt the last sections, subordinate clauses, may have been a bit difficult.

4.6 Assessing the Implicational Hierarchies

Processability Theory requires building an implicational scale and then calculating the coefficient of scalability. For this study, the scale was calculated in three different ways: according to the grammar for PT; using the proposed hierarchy from Table 3.3 above; and using the proposed hierarchy without determiners and non-count nouns. The stages across the top are from Table 3.3. Since determiners and non-count nouns proved unreliable, and furthermore the coefficient of scalability was low, that table is not discussed here, but is presented in Appendix E. Tables 4.8 and 4.9 show the results using the grammar that is strictly from PT. The coefficient for the grammar test on this scale was 96.8%, and for the writing test it was 100%, both of which meet the requirements for PT (Pienemann, 2011). It is

also important to note that the order of the hierarchies came out differently from that predicted in PT. The writing test came out closest to that predicted in the theory; plural nouns were not well represented in the writing samples, so it is impossible to say if this was generally unacquired or caused by some form of avoidance. The grammar test had the first two stages reversed.

Table 4.8 PT Implicational Hierarchy—Grammar Test

	A	B	C		
Level 3 29.2%	■	■	■	51	79.7%
	■		■	9	14.1%
		■	■	4	6.3%
Level 2 36.5%	■	■		72	90%
		■		8	10%
Level 1 24.2%	■			53	100%
Level 0 10.0%				22	100%

A—Stage 3: Plural Nouns
B—Stage 2: Past Tense
C—Stage 5: 3rd Person Singular Subject/Verb Agreement

Table 4.9 PT Implicational Hierarchy—Writing Test

	A	B	C		
Level 3 0%	■	■	■	0	0.0%
Level 2 1.4%	■	■		3	100%
Level 1 55.3%	■			121	100%
0 Levels 43.4%				95	100%

A—Stage 2: Past Tense
B—Stage 5: 3rd Person Singular Subject/Verb Agreement
C—Stage 3: Plural Nouns

The second scale discussed here is similar to the full one proposed in Table 3.3 above, except without determiners and non-count nouns. The scale is in Tables 4.10 and 4.11. This time, the coefficient for grammar was 90.9%, and 99.8% for writing, both of which again meet the threshold of

90%. One other important element to point out is that all of the coefficients of scalability for the writing test are unnaturally high, since the writing samples were, for the most part, too small to be very effective. In fact, none of the test-takers could claim to have acquired subordinate clauses or prepositions in writing samples, though several test-takers did provide at least a few contexts. It seems likely that, given two to three more writing tasks, there would be a higher number of acquisitions for both subordinate clauses and prepositional phrases. Another important point is that, in this case, prepositions proved harder than predicted, for both test types.

All of these hierarchies were also calculated using frequency to approximate an accuracy measure. This was done in order to compare the emergence criteria with a measure of accuracy in order to see how they compare, and also to give another way to explore the validity of the exam. An implicational hierarchy was developed for all 3 different hierarchies in this study, but only the scale for the extended hierarchy without Determiners and Non-Count Nouns is presented here. The other two can be found in Appendix E. The coefficient of scalability for the grammar test was 93.1%, and for the writing test, it was 90.0%, both of which are within the acceptable range for an implicational hierarchy. It is worth noting that, when calculated according to frequency, the grammar test had a higher coefficient of scalability than the writing test did for all scales (see Appendix E). Furthermore, the scale itself came out fairly similar, with a couple of exceptions. The potential implications of these observations are discussed in

the next section, below.

Table 4.10 Implicational Hierarchy for Grammar without Section 1

	A	B	C	D	E	F		
6 25.1%	■	■	■	■	■	■	22	40.0%
	■	■	■			■	7	12.7%
	■	■				■	7	12.7%
	■	■	■		■	■	6	10.9%
	■		■		■	■	4	7.3%
	■	■	■	■		■	4	7.3%
	■	■			■	■	2	3.6%
	■				■	■	1	1.8%
	■		■			■	1	1.8%
	■					■	1	1.8%
5 11.4%	■	■	■	■	■	■	8	32.0%
	■	■	■		■	■	8	32.0%
	■	■			■	■	3	12.0%
		■			■	■	1	4.0%
			■	■	■	■	1	4.0%
		■	■		■	■	1	4.0%
		■	■	■	■	■	1	4.0%
	■		■		■	■	1	4.0%
	■	■		■	■	■	1	4.0%
4 11.4%	■	■	■	■	■		17	68.0%
	■	■	■		■		4	16.0%
			■	■	■		1	4.0%
	■			■	■		1	4.0%
		■	■	■	■		1	4.0%
	■		■	■	■		1	4.0%
3 23.7%	■	■	■	■			40	76.9%
		■	■	■			4	7.7%
	■			■			5	9.6%
				■			3	5.8%
2 16.4%%	■	■	■				29	80.6%
		■	■				7	19.4%
1 5.5%	■						12	100%
0 6.4%							14	100%

A—Stage 3: Plural Nouns
B—Stage 2: Past Tense
C—Stage 3: Present Continuous
D—Stage 5: 3rd Person Singular Subject/Verb Agreement
E—Stage 3: Prepositions
F—Stage 6: Subordinate Clauses

Table 4.11 Implicational Hierarchy for Writing without Section 1

	A	B	C	D	E	F		
6	■	■	■	■	■	■	0	0.0%
5	■	■	■	■	■		0	0.0%
4	■	■	■	■			0	0.0%
3	■	■	■				0	0.0%
	■		■				3	100.0%
2	■	■					2	0.0%
1	■						121	83.3%
0							94	16.7%

A—Stage 2: Past Tense
B—Stage 3: Present Continuous
C—Stage 5: 3rd Person Singular Subject/Verb Agreement
D—Stage 3: Plural Nouns
E—Stage 3: Prepositions
F—Stage 6: Subordinate Clauses

Table 4.12 Implicational Accuracy Hierarchy for Grammar Test without Section 1

Levels	A	B	C	D	E	F	#	%
Level 6 2.7%	■	■	■	■	□	■	**1**	50.0%
	■	■	■	■	■	■	1	50.0%
Level 5 23.7%	■	■	■	■	■	■	**19**	36.5%
	■	■	■	□	■	■	12	23.1%
	■	■	□	■	■	■	6	11.5%
	■	□	□	□	■	■	4	7.7%
	■	□	■	□	■	■	4	7.7%
	■	■	□	□	■	■	5	9.6%
	■	□	■	■	■	■	2	3.8%
Level 4 15.1%	■	■	■	■			**20**	60.6%
	■	■	□	■			3	9.1%
	■	□	□	■			4	12.1%
	■	□	■	■			2	6.1%
	□	□	■	■			1	3.0%
	□	■	□	■			2	6.1%
	□	□	□	■			1	3.0%
Level 3 19.2%	■	■	■				**31**	73.8%
	■	□	■				8	19.0%
	□	■	■				2	4.8%
	□	□	■				1	2.4%
Level 2 16.4%	■	■					**30**	83.3%
	□	■					6	16.7%
Level 1 15.5%	■						**34**	100.0%
0 Levels 9.1%							**20**	100.0%

A—Stage 3: Plural Nouns
B—Stage 2: Past Tense
C—Stage 3: Present Continuous
D—Stage 5: 3rd Person Singular Subject/Verb Agreement
E—Stage 3: Prepositions
F—Stage 6: Subordinate Clauses

Table 4.13 Implicational Accuracy Hierarchy for Writing Test without Section 1

Levels	A	B	C	D	E	F	#	%
Level 6 3.7%	■	■	■	■	■	■	0	0.0%
						■	3	37.5%
	■	■	■			■	2	25.0%
		■				■	1	12.5%
	■					■	1	12.5%
	■	■				■	1	12.5%
Level 5 6.0%	■	■	■	■	■		0	0.0%
	■				■		6	46.2%
					■		3	23.1%
				■	■		1	7.7%
		■		■	■		1	7.7%
		■			■		1	7.7%
	■	■	■		■		1	7.7%
Level 4 6.0%	■	■	■	■			1	7.7%
				■			6	46.2%
	■			■			2	15.4%
			■	■			1	7.7%
	■	■		■			1	7.7%
		■		■			1	7.7%
		■	■	■			1	7.7%
Level 3 8.8%	■	■	■				2	10.5%
	■		■				7	36.8%
			■				5	26.3%
		■	■				5	26.3%
Level 2 12.4%	■	■					16	59.3%
		■					11	40.7%
Level 1 26.3%	■						57	100.0%
0 Levels 36.9%							80	100.0%

A—Stage 3: Plural Nouns
B—Stage 2: Past Tense
C—Stage 3: Present Continuous
D—Stage 3: Prepositions
E—Stage 5: 3rd Person Singular Subject/Verb Agreement
F—Stage 6: Subordinate Clauses

Chapter 5

Discussion

The discussion section addresses each of the research questions to elaborate on the results from above, and to look at possible implications for PT, this particular diagnostic test design, and diagnostic tests generally.

5.1 Can we achieve an acceptable level of reliability for the grammatical diagnostic test used for this study?

The answer to this question appears to be “Yes”, though with a caveat. The reliability score for the whole test was fairly high, 0.926, and improved slightly when the first section was removed, to 0.929. More telling are the mean, median and mode scores, all of which are almost identical when the first section is deleted. Inter-item correlations and covariances also tended to improve somewhat when the first section is deleted. All of these results suggest that there may have been a particular problem with the first section, which included determiners and non-count nouns, and that this section may not be reliable. This section was therefore analyzed in greater depth.

Within the context of this study, there were three possible reasons why this section was problematic. This first has to do with PT and the nature of the implicational hierarchy. It is possible that the test results are an indication that determiners and/or non-count nouns may not be compatible with the theory, which would confirm at least that much of the implicational hierarchy as set out in PT. This may also give some information on the

nature of non-count nouns, which are notoriously difficult for Koreans to acquire; perhaps, instead of being an element of grammar, it should instead be viewed as a semantic issue, and should be taught as such. This would also explain why the non-count nouns in particular came out so high on the expanded implicational scale, at the very top; they may not properly be considered grammar problems at all.

A second possibility is one that several raters and teachers referred to, and that has to do with the fact that the first section was the only one that required leaving a gap in some items. Raters and teachers felt that some students may have instinctively felt the need to fill in those gaps, even when they knew they were not supposed to. Suggestions for improving this included offering students a different word to put there, such as “some”, or giving them the option of writing Ø, as is done in some classrooms for tasks of this type. Another problem some raters experienced is that, after a while, they felt that “the” *could* have been used in some of those non-count gaps, even though the contexts were clearly inappropriate for them. Anyone who has worked on a large-scale language test will sympathize; after spending a great deal of time looking at items from every possible angle to ensure there is only one answer, test-makers tend to imagine any possible scenario, even ones that do not exist within the given context. However, it is worth noting that, for that reason, greater context may have been helpful in order to obviate too much imagination.

A third possibility is in the form of the tasks themselves. Upon re-

evaluating them, it is clear that these are the most overtly explicit tasks on the test. Originally, this task-type was used in order to provide a simple beginning for the test-takers, to ease them into more difficult and less familiar tasks. The use of pictures was initially explored, but this was only a surface distraction, as they were not needed to answer the items, so it was rejected. As a result of the lack of distractions, the test-takers were possibly more focused on the fact that these were grammar tasks, and thus were not accessing implicit knowledge stores. This observation—that students may have viewed these tasks more as grammar than as communicative tasks—is bolstered by the comments made by raters and teachers above that they felt making the students leave some gaps open for non-count nouns was difficult for the students, as they intuitively want to fill a gap. The suggestion of using Ø in particular would make this a more explicit exercise, rather than less. It is unclear which of these problems is most prevalent—possibly a combination of them—but at any rate, the first section was clearly a problem. This was why two test forms were evaluated throughout the results section.

This last point about whether Section 1 taps implicit or explicit knowledge is particularly important as it suggests that, if the first section performed poorly because it was too explicit, perhaps this indicates that the rest of the test was more successful at utilizing implicit knowledge. This suggests that those strategies employed in Norris (2005) and Chapelle *et al*. (2010) could be at least partially effective for imitating productive skills

while still maintaining many of the qualities of multiple choice tasks, especially ease and speediness of correcting and giving feedback. Regardless of whether the test itself fits the PT hierarchy, this task-type would likely be appropriate for use on a diagnostic test. Further research needs to be done in analyzing the error types and designing tasks—this is discussed in the future research section—but at least this format appears to be a helpful complement to writing tasks for diagnosing grammar problems in a semi-productive way that could help student writing.

Another element that must be mentioned is the disparity in the coefficient of scalability between the grammar and writing tests. In all cases, the implicational hierarchy for the writing test had a higher coefficient than did that for the grammar test, which indicates that the writing test may have been more reliable, but there are several other possibilities as well. One is that the writing sample was too small, and a larger writing sample would likely have resulted in a score closer to that of the grammar test. Another possibility is that the grammar and writing tests each assessed a slightly different aspect of language acquisition. The grammar test may have been a more accurate measure of recognition and recall, while the writing test could be directed more towards production. This latter was the basic assumption utilized in preparing feedback for this study; acquisition in the grammar test but not in the writing test was interpreted as meaning that the student had acquired recognition of a grammar point, so the recommendation was to encourage the student to produce the grammar point more, in this case, by

writing. However, this is not at all proven, and requires further research to support such a claim.

One other point that should be noted is that the final section, the subordinate clauses, was also fairly difficult, and perhaps not as discriminating. However, this is in part because a number of students did not finish this section, though the reasons why they did not complete it are unclear. The teachers felt the students had sufficient time, so it seems unlikely that this was the problem. It may be that test-takers were avoiding these items for other reasons, either because they felt they could not solve them, or because they simply had not learned these forms yet, and did not want to attempt them. In particular, the last four items in subordinate clauses (the entire last section) were frequently left blank. These items also required the most cognitive effort, since test-takers had to rearrange the words, add a word, and also change the form of one word. It may be that this was too much effort, particularly for the age group, and students simply gave up. In this case, it could be that these different elements should *not* be combined in one task, at least for this age group.

5.2 Do the items for the grammatical diagnostic test work well at an item level in terms of item discrimination and difficulty? Were there any poorly performing items?

Since the problems with the first section have already been addressed, this section will focus on the remaining items of the test, particularly as the chart in Figure 4.1 shows the difficulty and discrimination of them. Most of

the items performed in a relatively predictable way, with a few exceptions, two of which—the first preposition item and the first subordinate clause item, both being quite high in difficulty, and low in discrimination—are explored here to see what kind of information they offer for making improvements to the test.

The first preposition item, which is number 26 in Figure 4.1 above, was difficult, but had low discrimination. The target sentence for this item (with the part the students had to solve underlined) was "We have it before lunch on Mondays. / We have it on Mondays before lunch"—there were actually two possible target answers given the instructions and the words the students had. Two more answers were possible, classified as "non-target" on the scoring sheet because they were grammatically correct but did not follow the instructions precisely: these were "We have it before lunch", and "We have it on Mondays". This task-type required test-takers to put the words in order and delete one extra word, in this case "the". A review of the error types test-takers made on this item revealed 4 basic types: overuse of "the"; word order, namely putting the preposition after the noun; failing to use "before"; and other, random errors (such as deleting "lunch"). Of these, by far the majority was overuse of "the": this error comprised 53.3% of the total errors, as opposed to 26.3% for putting the preposition in postposition; 11.7% failing to use the preposition; and 8.8% other (it must be noted that some test-takers made more than one mistake, for example, using "the" but not using "before"). This is an interesting finding as regards the testing of

determiners and non-count nouns; this item apparently tested two different grammar points, and the determiner appears to have been significantly more difficult than the preposition. This raises the question of how to test determiners at all for diagnostic purposes, and what exactly is the nature of the difficulty with determiners. The other interesting finding is that the next most common error may have been one of L1 influence, namely placing the preposition in postposition (inverting the noun and preposition). This may be an area for falsifying the Developmentally Moderated Transfer Hypothesis, which states that there will not be an observable L1 influence. However, other languages would need to be similarly tested in order to adequately challenge the DMTH—this observation merely gives insight into how it might be tested.

The first subordinate clause item was also sample sentence (8) used above, “Today, Susan decorated a cake while John was playing tennis. Several raters suggested that the test-takers made errors on this item because they did not read the instructions properly; they had to add one word, in this case “while”. The high percentage of students who did not use any subordinating conjunction may, indeed, indicate that many test-takers did not read or follow instructions. However, most errors matched observations made by several raters that the overwhelming majority of students tended to either overuse “and” or underuse cohesive devices in their writing; this is perhaps an issue of complexity, but it is also related to emergence and implicit knowledge in that the grammar test results *did* seem to reflect the

actual results from the writing test. This is an important finding, and indicates that this test task has promise, and therefore more should be done to evaluate whether there was a problem with the instructions or not. It should also be noted that 18.7% of the students who dropped the conjunction completely also made word order errors, typically inverting the verb to make “Today, Susan decorated a cake was John playing tennis”. This is a substantial result, but is confusing. Whether this happened because they do not understand subordinate clauses, or perhaps they were thinking of interrogative forms (as per PT), is unclear. More research needs to be done to understand why test-takers would want to invert the subject and verb in that context. Neither L1 nor PT seems to explain it, but the number is too high to be purely random.

Those were two tasks that performed poorly; what about a task that performed well? One item that showed promise was the second item in the first subordinate clause section, the target answer of which was “Scott was talking on the phone before he went to school this morning.” It was one of the most discriminating items on the test, and also had a moderately high difficulty level. The task measured two different target grammars: the use of the conjunction “before” in the middle of the sentence, as opposed to at the beginning of the sentence; and present continuous tense, through omission of “was” in the words given to the students. There were 146 errors in total (keeping in mind that numerous test-takers made 2 errors on one task); of these, 71 (48.6%) involved misplacement of the conjunction, and 75 (51.4%)

involved not using “was”. This result suggests several things about this type of task. This item worked because it combined two elements that clearly interact within the sentence; the conjunction and the appropriate tense. This is in opposition to the preposition item, which tested two grammar points that may *not* have a direct effect on one another, even though they do frequently appear together: prepositions and determiners. Furthermore, the information that can be gleaned from the errors on this task is very specific. Misplacement of the conjunction “before” suggests some interference in ignorance of the visual cues (which not only had time signatures, but were also placed visually to reinforce the before/after relationship). An error with the tense with correct placement of the conjunction could indicate the test-taker is unaware of the aspectual relationship between the two clauses, and that the present continuous is appropriate in this situation. But an error with both grammar points could indicate that the test-taker simply has not learned the form yet. Each error type—and it is significant that there were only these three general types (A, B, A+B)—points to a specific underlying cause of *nondum* ability that identifies a student weakness, and also suggests a course for remedy. This is a task type that has the potential to perform well for diagnostic purposes.

Overall, this qualitative analysis of weak and strong items suggests that the items on this grammar test do have considerable potential, both for use on a diagnostic test and as an alternative or supplementary method for testing implicit knowledge. More careful examination of test-taker error-

types should be done in order to strengthen the items that have already been developed, and enhance the types of diagnostic information that can be gleaned from them.

5.3 What are the relationships among the subtest, full test, and self-assessment?

As was mentioned above, the writing test was very short and perhaps not the best measure of proficiency; a longer writing sample would have been better (though it is unclear how much more would have been sufficient). Nevertheless, even a correlation of 0.62 between the overall grammar test and writing test suggests that there is a relationship between the two. This could support the assumption from Levelt's Speaking Model that fluency requires automaticity of grammar, since some correlation was observed in this study between grammar and the results of a timed writing test (Levelt, 1989).

A look at the subsections reveals that the subsection which correlated highest with both the grammar and writing test was the aggregate score for the subordinate clauses, which seemed to be the strongest predictor among the subsections. The result is interesting, since none of the individual subordinate clause sections correlated very highly with either test, suggesting that the various forms of testing subordinate clauses may have worked well together; the fact that slightly different task-types performed well in tandem is an important idea to explore for the other grammar points. Generally speaking though, the high overall correlation of the subordinate

clauses makes sense, since subordination is an issue of complexity and accuracy, which, along with fluency, form the major components of proficiency (Skehan, 2009). Prepositions also correlated fairly highly, just under subordinate clauses, which could provide some justification for the assumption that acquiring prepositions is necessary in order to acquire subordinating conjunctions, though this is a preliminary judgment. If accurate, this also supports the acquisition order set out in PT based on LFG. Outside of the subordinate clauses, the *lowest* correlation scores came from two of the elements that are predicted to be in the lower levels of PT, plural subject/verb agreement and plural nouns. Past tense, which should also be acquired early, also had a low correlation score with the grammar and writing tests (0.65 and 0.43 respectively). Singular subject/verb agreement correlated fairly well with the grammar test, but quite poorly with the writing test, which may be related to the fact that many test-takers had a hard time with maintaining aspect in their writing; while the instructions explicitly told test-takers to write in the past tense, and also provided two sentences that began the story from a 3rd person perspective, many test-takers switched partway through their writing to 1st person present tense. Interestingly, most of them *began* writing past tense 3rd person, and therefore created enough contexts to achieve acquisition of past, but then changed after a few sentences, or went back and forth between the two. For these reasons, perhaps the results are showing a tendency on the part of the raters to penalize test-takers for switching to the present tense, but failing to

use it properly. But this is a weak speculation; certainly, this requires closer analysis of error patterns in both the grammar test and the writing samples.

Another preliminary observation is that the PT hierarchy may not be a strong indicator of success in proficiency, since these levels did not tend to correlate well with the writing test. It is important to note that most levels of morphology were not tested in the study, since it proved rather difficult in a writing test. Yet, this as well needs to be considered; while the PT hierarchy may be effective for speaking assessment, which is transparently interactive, it may not be appropriate for most forms of writing. This type of assessment may be most appropriate for very specific forms of writing such as texting or online chatting, but might not be very useful for academic or business writing. The expanded hierarchy, by contrast, had elements that correlated much better with the implicational hierarchy, but it is impossible to tell at this point in time if it would be as useful for speaking assessments. It may be that writing has much different hierarchies than speaking does. The two hierarchies should be further explored in this way, considering correlation to complexity measures for both speaking and writing, which will assist greatly with offering feedback that is useful for test-takers.

A general observation is that it is unclear, at this time, whether the grammar test is actually measuring emergence/acquisition, or accuracy, or some combination of the two. Given the hierarchy that came out of these correlations, particularly at the lower levels, certainly the case can be made for it being a measure of acquisition as stipulated by PT. On the other hand,

the higher levels in the hierarchy may be more accurately measuring complexity, accuracy, and fluency (CAF), even if they do fit into a hierarchical pattern. Still, the pattern is unclear. The key may be in the error analysis, which was too broad an area of research for this study.

5.4 What are the perceptions of the test from the viewpoint of the test-takers, raters, and teachers?

The majority of academic pursuits that occur at institutes such as the ones used for this study deal largely in test-prep, particularly for national-level tests and school entrance exams. Teachers and institute administrators were very interested in the grammar test from the beginning,[⑤] since education is a major element of Korean society, institute schedules are hectic, and there is little time for teachers to get to know their students' ability levels. A test such as the one developed for this study might assist in quick yet personal evaluations of students, which was a need expressed by all of the administrators and instructors who participated in this study. The fact that it is backed by theory and therefore has a universal aspect to it was also appealing, as it might offer stakeholders a detailed comparison with other learners. An unexpected discovery through this research was that most of the after-school institutes, especially those outside of Seoul, do not track student progress in any significant way, such as through regular evaluation—the measure of success is whether students do well on their

[⑤] Originally, the study was intended to have fewer test-takers, but several institute administrators became overly enthusiastic and added subjects on their own initiative.

tests *outside* the institute, not whether they are proficient. There is little in the way of speaking or writing practice, which is part of the reason why national-level Korean tests are attempting to introduce more productive tasks, to hopefully stimulate positive washback effects. Institute administrators were therefore happy to participate in the study in order to get some feedback on student levels. To give them feedback, a simple form was devised which interpreted the results rather conservatively, in lieu of the question of whether the grammar test accurately represents implicit knowledge. On the form, students who successfully completed 4 tasks in a skill section on the grammar test were said to have partially acquired that grammar skill; full acquisition required 4 contexts on the writing task as well. There was also a section for making recommendations for students, which mostly consisted of recommending that students who had not acquired a grammar point practice it more in worksheets (explicit knowledge), while students who had partially acquired an ability were advised to practice with writing it more in stories and expositions (implicit knowledge). Institute administrators and teachers were interested in the test because they felt it was helpful to have some measure, other than national and entrance exams, by which to evaluate student progress; the more so since no grade was involved in the test, which was also appreciated. When asked if they thought using these types of tests would be useful, they felt that they would, depending on how detailed and applicable the feedback was.

The raters who rated the test just for the grammar points (not the

holistic score) spent the most time with the test, and therefore had more feedback to offer. One predictable and perhaps somewhat obvious comment they all had, when comparing the grammar and writing tests, was that it was much easier to score errors on the grammar test than it was on the writing test. Most mentioned trying to interpret what the students were thinking in the writing test in order to understand the exact nature of the error, whereas the thinking behind errors in the grammar test was clearer and easier to define. This is an important observation when it comes to finding not only student errors, but also student weaknesses, for diagnosing *nondum* ability. Raters did feel that students tended to make consistent errors across the two tests; in other words, if a test-taker made errors with verb tense on the grammar test, they were likely to do so on the writing test as well. Furthermore, if a test-taker was likely to omit words on an item, they were likely to make the same type of error on other items, and on the writing test as well. Raters also felt that there were some commonalities among the test-takers, meaning that many test-takers made similar mistakes on the same items. More specifically, some raters felt that test-takers displayed at least a moderate influence from L1, particularly in the subordinate clause sections of the test, though there may have been other external influences as well. This suggests that analysis of student errors might have to be analyzed in two different ways. First, analysis of error patterns within a student, to see if that student makes certain types of errors. Second, analysis of errors across students, to see if there are some common errors among them. The first type

of error is internal and predicted by PT; the second is external, and predicted by other models of diagnostic testing. Most raters seemed to feel that the test was an adequate difficulty level for the students, although they felt the subordinate clause section may have been a bit difficult.

This leads to more critical comments which focused on the subordinate clause section, where several raters felt that students had not really read the instructions, or else did not understand them despite the fact they were written in both Korean and English (see the tests in Appendices A and B). This was already discussed somewhat above, in the analysis of the first item in the subordinate clauses section. One point that can be made is that raters sometimes sympathized strongly with the test-takers, to the point that two raters actually expressed feeling like they thought a test-taker *could* have solved the problem, but somehow missed something in the instructions or the examples, and that they felt badly for the students. The reaction is also particularly interesting, given that the subordinate clause section did correlate most highly with both the total grammar score and the holistic writing score, which suggests it functioned fairly appropriately; nevertheless, more time spent in explaining the instructions to test-takers and ensuring that they understand the tasks completely may produce different results in future research, and is therefore worth investigating. One would presume, if such a test were to be integrated into a classroom situation, that test-takers would gradually become more familiar with the style of the tasks. It is important to ensure that the usefulness of these tasks is not diminished by

familiarity with them.

5.5 Are mastery and non-mastery patterns consistent with predictions based on the Processability Theory hierarchy?

There were 3 possible PT hierarchies proposed in this study. The first was the implicational hierarchy as delineated in Pienemann (1998); in this study, past tense, which should have been the first acquired, followed by plural subject/verb agreement and plural nouns at the next level (in no specified order), and finally singular subject/verb agreement. This was not how the results stood for either test, though again, the writing sample was quite small, so it is difficult to make too many conclusions based on this. Still, in the writing sample, past tense was acquired first, though not by all students. Second was 3rd person subject/verb agreement, which may seem contrary since test-takers were instructed to write in the past tense; writing-paragraph tasks should be carefully designed to see student acquisition of present and past tense. As noted above, however, many test-takers switched from past to present tense, and therefore raters were told to count those instances of present tense as well. Numerous contexts were therefore possible for 3rd person singular subject/verb agreement, and here test-takers performed much as would be expected, in that only 3 test-takers matched the emergence criteria for it; all others, if they used it, failed to achieve emergence. No test-takers matched the criteria for plural noun acquisition, despite the fact that there were opportunities for it. However, most test-takers focused on the action and the major participants in the pictures, rather

than the objects around them, so there were few contexts created for plural nouns.

The grammar test had at the lowest level plural nouns, followed by past tense, 3rd person singular subject/verb agreement. According to PT, past tense should have been acquired first; that the hierarchy did not follow in that sequence could mean several things. It could mean that those tasks were not, in fact, accessing implicit knowledge, but were still too explicit. On the other hand, it could indicate a potential inconsistency in the PT hierarchy, especially since the actual numbers were fairly high and, in some cases, very close in numbers. A third option is that there was some other interference, perhaps from the task type, which led to more false positives in plural nouns or false negatives in past tense. In the absence of greater evidence, particularly as the writing sample was small and insufficient to challenge PT, it is much safer to say that the first or third suggestion is more likely, but this exercise does at least suggest a potential method for falsifying the theory.

The full implicational hierarchy that was proposed above is discussed only briefly here to mention the problem of determiners and non-count nouns. Determiners came in at a surprisingly low level for both tests, particularly for the writing test, where it was at a very similar level to the past tense. However, this may have been because of a “guessing” strategy that many test-takers seemed to take whereby they added determiners to every noun, and in some instances, even to adjectives and verbs for good

measure. This was frequently the problem with the non-count nouns as well. One observation is that the test-takers were almost shooting blindly, hoping that if they kept using determiners, at least some would be correct. In this sense, can we say that a test-taker has “acquired” the use of determiners? By PT criteria, which was itself drawn from other sources (Pienemann, 1998), it would count, since there was often at least 4 correct contexts. From this point of view, emergence as correct usage in 4 contexts may not be a sufficiently complete or complex measure of acquisition for some grammatical forms.

Finally, there is the modified form of the full hierarchy proposed, without the determiners and non-count nouns. The writing test was not bad at predicting the hierarchy at the lower levels, but was rather erratic at the higher levels. This may have been simply a consequence of the small writing sample, but it may also point to an interaction between the subject matter and test-taker fluency. As was noted above, most test-takers tended to focus on the participants in the story, rather than on events or observations that could be made around the main action. Is this a consequence of cognitive load, of an insufficient amount of attention being given to content in the working memory? Or are there other factors going on that are too complex to extricate from the written text? This highlights one of the problems of using productive and integrated tasks for diagnostic purposes; it is difficult to separate the various elements involved in order to decide which is causing the student weakness, what is the exact nature of the

nondum ability. A great deal of data must be gathered to properly assess test-takers, particularly those at lower levels or who are in the early stages of acquisition, which may require substantial time and effort. As a result, there are likely to be numerous false positives when relying solely on productive tasks for diagnostic purposes.

By contrast, the grammar test performed fairly well with respect to the proposed hierarchy, with an acceptable coefficient of scalability at 90.9%. There were two notable exceptions in the scale: the past tense, and prepositional phrases, both of which came out at levels that were higher than predicted. In the case of the section on past tense, this could have been a task effect from being combined with the section on present continuous, which may have confused test-takers; in fact, a fair number thought that they were supposed to choose between either past continuous or present continuous, not simple past or present continuous.

As to the preposition section: there are a couple of possibilities. On the one hand, it may have had the problem of testing too many elements at once, as was mentioned above. On the other hand, this may be indicative of an issue mentioned in the Literature Review section, namely, how development works *within* a section. Although preposition phrases are at the phrase level, and within the tree structure of LFG, appear at a lower level than verb phrases or sentences, they may in fact be as difficult as, or even more difficult than, those elements. In this sense, if the PT hierarchy *were* to be expanded, a more complex system may be required, such as overlapping

difficulties or pathways, both of which are discussed further below. Overall, more attention needs to be given to error patterns and difficulty levels in these task-types, in order to devise better tasks for them, as well as to ensuring that test-takers understand the instructions. This is an area for future research.

Although the grammar test seemed to perform well and has some promise with further development, it should be pointed out that it is not yet a convincing replacement for an actual productive task. While it certainly seems effective, and there is evidence that it does access implicit knowledge in at least some of the tasks, it gave too many false positives to be confirmed as actually reproducing the qualities of productive tasks. Comparison of the emergence and frequency hierarchies supports further that there was an underlying difference between the two tests. The order of the grammar acquired on the grammar test was the same when calculating according to emergence or frequency, while the order changed somewhat on the grammar test. Perhaps more significant is the fact that the coefficient of scalability went up by over 2 points for the grammar test when going from emergence to accuracy, but went down by almost 10 points for the writing test. This may have been because the writing sample was too small; on the other hand, it might indicate that the writing test was more effective at accessing implicit knowledge through production, as evidenced in the emergence hierarchy, while the grammar test targeted more explicit knowledge through recall and recognition. In this sense, perhaps the best

approach would be to combine the two test types in order to create a much broader picture of each student's abilities and *nondum* abilities. A much larger writing sample must be collected for comparison before that can be confirmed. At this time, it can only be said that they likely work well together in identifying errors and weaknesses and diagnosing *nondum* ability. At the same time, it may also be appropriate to explore other confirmatory measures than the coefficient of scalability used in current PT studies, as well as exploring other models than an implicational hierarchy for development of grammatical ability.

Chapter 6

Conclusions and Future Research

6.1 Evaluating the Instrument Developed for this Study

Validating a test is an iterative process. This study was the first pilot of this instrument, and there is still a lot that can be done with the data before conducting a second pilot. To begin with, this study was unable to examine test-taker errors. The next stage is to mine the data, both in the grammar and writing tests, to find any significant patterns, either within a test-taker (internal) or across test-takers (external), and to attempt to quantify that data in a meaningful way. As with other educational diagnostic tests, this information can then be used to improve the test so as to better understand why test-takers make the mistakes they do. Multi-stage testing may also be helpful, as well as think-aloud or recall protocols. All of this information will help to improve the diagnostic information that can be gleaned from this test. Future research should therefore focus more on understanding what typical errors test-takers make, and why they make them. In this sense, PT offers valuable insight into how to evaluate different learner errors for analysis.

Along these lines, more research is needed to better understand how different grammar points do or do not interact. The problems with determiners and non-count nouns were unexpected, and suggest that perhaps they need to be treated differently in testing and teaching. On the other hand, combining conjunction and tense worked well and provided immense

diagnostic information, perhaps more than would have been provided by testing either grammar point on its own. This seems to be the way in which a hierarchy appropriate for diagnostic grammar testing could be built. But this could be applied beyond grammar; reading, for example, requires several layers of comprehension. Finding ways to test isolated interactions will improve the diagnostic value of an item while still making a form that is efficient for use in the classroom.

Additionally, it is essential to understand how such a test can be integrated into a classroom situation; after all, the purpose of this assessment is to assist students in overcoming hurdles in language development. This requires providing feedback that is clear and meaningful for all of the stakeholders. This is also part of the validation process, since it relates to the consequential aspects of the test (Messick, 1995). This is a key element, since it is the point at which the diagnostic test intersects directly with the classroom. Feedback must give an appropriate amount of actionable information in a format that is easy to understand (Jang, 2009a). An observation from this study was that stakeholders preferred a report card-style format to a paragraph format (see Appendix F for examples of each).Stakeholders also found the information interesting and potentially useful, though they were unsure of how to use it. Therefore, future research should also focus on integrating diagnostic testing into the classroom. Such research would require longitudinal studies in order to evaluate the effect that such testing has, not only on the development of student abilities, but

also on psychological attitudes such as anxiety and motivation.

Creating and then validating diagnostic feedback is a substantial and difficult aspect of validating a diagnostic test. In the context of this study, it is not enough to validate the instrument and the hierarchy that was used to develop it. Researchers must understand the dynamics of the classroom, current theories on language education and development, and how theories about language assessment interact with these elements. What useful information can be generated from a hierarchy? And there is also the issue of explaining the difference between emergence and acquisition, and why acquiring something in the sense of emergence may not reflect a high level of accuracy, nor even the ability to obtain high scores on proficiency exams. In this sense, unless explained in a way that will be clear and useful, information from implicational hierarchies may prove overly frustrating for those attempting to achieve some educational goals. Ultimately, we do not yet understand the way emergence, accuracy, and proficiency test scores interact. Perhaps this should be the first line of inquiry, before developing more tests and attempting to validate them.

One other, pedagogical element of this study that should be of interest, at least within the Korean context, is that of the performance of students. There was little difference between the output of grades 5/6 and 7/8. It is difficult to draw too many conclusions from this, as the sample size for grades 7/8 was rather small. Furthermore, it is unclear as to what should be considered normative: at what age ought students to have acquired

subordinate clauses, for example? Future research may be interested in using the theory to develop tests for evaluating the effectiveness of Korean ESL programs, and whether these grammar points are being taught in a way and at an age that is appropriate. This is a different type of diagnostic test, but certainly one that is addressed elsewhere in the literature (Kunnan & Jang, 2009; Shohamy, 1982).

There were several limitations in this study that should affect future research. The first two have already been mentioned above—the lack of feedback from students and teachers, as well as the small size of the writing sample. Both should be overcome in future studies. In particular, a much larger writing sample, perhaps as many as 4-5 similar tasks, would be required to confirm that the grammar test is accessing implicit knowledge, as well as to adequately support adding further levels to PT. One further limitation is that this study was conducted only on Korean students, and is therefore only generalizable to that group. This is appropriate for controlling social factors; on the other hand, it greatly reduces the generalizability of the findings.

6.2 Evaluating Processability Theory

Evaluating a theory is never easy—it takes time and patience, and by the criteria of falsifiability, it is a never-ending process. Furthermore, the data that has been gathered is only generalizable to Korean elementary and middle school students, although the methods that have been used for this study may be transferrable to other, international contexts. Still, it is

worthwhile in that, the more evidence gathered, the more researchers and practitioners can understand the process of acquiring a language. In this case, there were no strong conclusions, but there are several tentative ones that can offer direction for future research.

To begin, the PT hierarchy, as it stands now, seems plausible, but limited. It is difficult to add to the hierarchy, but the hierarchy in its current form has only limited application. It cannot offer the fine-grained analysis and feedback required for diagnostic feedback. In this sense, perhaps the PT hierarchy is not suitable for use in diagnostic situations, which require testing a wide range of grammar points at a much greater variety of levels than it currently proposes. Instead, it may be more appropriate for placement tests, since the developmental stages are fairly broad. This criticism is bolstered by the problems that were encountered by raters scoring the writing tests for grammar. Their frustrations help to explain why productive tasks are sometimes problematic to analyze for diagnostic information; it is difficult to separate the elements of writing in such a way as to get at the underlying student weakness that will explain *nondum* ability. Though there may be some potential in some of the levels attempted in this study, and in the methods used to create them, they have not yet been confirmed, and there is certainly a problem of establishing difficulty levels. Future research may try to replicate the results of this study, in Korea or elsewhere, with a larger body of writing for comparison purposes. It would also be helpful to attempt more robust measurement methods, in order to see if this makes a

difference in the hierarchy, and whether other levels could be added as a result.

One question that arises out of the research done in this study is the nature of potential levels that could be developed for PT, and what these mean for teaching and testing grammar. One example is the fact that determiners seemed so different throughout the test. This could be partially explained by the first section being aimed more toward explicit knowledge than later sections were, but that alone does not explain the issue: Why then did it seem to make the preposition item so difficult and ineffective? This seems to indicate that there may be some grammar points that are more related to accuracy and others that are more appropriate for measuring fluency and complexity. A comparison of the two types of errors on the preposition item highlights this matter: a native English speaker can easily understand "*We have math class before the lunch on Mondays", but not "*We have math class lunch on Mondays". The first sentence is an auxiliary problem, the second fundamental, since the error in the first does not impede comprehension, while the second does. It may be that PT primarily evaluates some aspects of comprehension, and is most appropriate for that purpose. If so, then further levels could reflect this concept; at least, it is worth exploring.

The Developmentally Modified Transfer hypothesis is also a bit troubling in that it does not seem to adequately explain L1 influence; at least, not enough aspects of it. The DMT hypothesis has not been adequately

challenged, but it should be, because it offers potential for understanding an area of language acquisition that is of great concern. There are some aspects of the current study that could be used to explore the DMTH—for example, prepositions vs. postpositions, or placement of subordinate clauses. For these reasons, the DMT hypothesis seems a fruitful area for future research, exploring ways that an L1 may or may not influence L2 acquisition. It may also assist in giving further insight into which grammar points could increase PT's implicational hierarchy.

Most importantly though, for this researcher at least, is that future research should explore more theoretical approaches to diagnostic testing. Although somewhat critical of PT, the results of this study show that some aspects of the theory may be further investigated to find useful ways of developing diagnostic tests. It is important to continue testing the theory in order to understand language acquisition in ways that are meaningful for testing and instruction.

References

ACTFL Proficiency Guidelines. (2012). American Council on the Teaching of Foreign Languages. Accessed online at: http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Alderson, J. Charles (2000). *Assessing Reading*. Cambridge UP: Cambridge.

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

Alderson, J. Charles, & Huhta, Ari. (2011) Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? *EUROSLA Yearbook*. 11: 30-52.

Arslan, H.O., Cigdemoglu, C., Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International Journal of Science Education*. 34(11): 1667-1686.

Bachman, Lyle F. (1990). *Fundamental Considerations in Language Testing*. Oxford UP: New York.

Bachman, Lyle, & Palmer, Adrian. (2010). *Language Assessment in Practice*. Oxford UP: New York.

Baten, Kristoff. (2011) Processability Theory and German case acquisition. Language Learning. 61(2): 455–505.

Blatchford, Charles H. (1971). A Theoretical Contribution to ESL Diagnostic Test Construction. TESOL Quarterly. 5(3): 209-215.

Bresnan, Joan. (2001). *Lexical-Functional Syntax*. Blackwell: Malden.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. Language Testing. 15(2): 119–157.

Chapelle, Carol A., Chung, Yoo-Ree, Hegelheimer, Volker, Pendar, Nick, and Xu, Jing (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*. 27(4): 443–469.

Chapelle, Carol A., Enright, Mary K., & Jamieson, Joan M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*™. Routledge: New York.

Di Biase, Bruno, & Kawaguchi, Satomi. (2002). Exploring the typological plausibility of Processability Theory: language development in Italian second language and Japanese second language. *Second Language Research.* 18(3): 274–302.

Donohue, James P., & Erling, Elizabeth J. (2012). Investigating the relationship between the use of English for academic purposes and academic attainment. *Journal of English for Academic Purposes*. 11: 210-219.

Dyson, Bronwen. (2009). Processability Theory and the role of morphology in English as a second language development: a longitudinal study. *Second Language Research*. 25(3): 355-376.

Ellis, Nick. (2008). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The*

Modern Language Journal. 92(2): 232-249.

Ellis, Rod. (2008). Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics*. 18(1): 4-22.

Falk, Yehuda N. (2001). *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. CSLI: Stanford.

Gass, Susan M. & Selinker, Larry. (2008) *Second Language Acquisition: An Introductory Course*. Routledge: New York.

Håkansson, Gisela, & Norrby, Catrin. (2005). Grammar and pragmatics: Swedish as a foreign language. *EUROSLA Yearbook*. 5: 137-161.

Håkansson, Gisela, & Norrby, Catrin. (2007) Processability Theory applied to written and spoken L2 Swedish. *Second Language Acquisition Research: Theory-construction and Testing*. Ed. Mansouri, Fethi. Cambridge Scholars: Newcastle. 95-118.

Håkansson, Gisela, & Norrby, Catrin. (2010) Environmental influence on language acquisition: Comparing second and foreign language acquisition of Swedish. *Language Learning*. 60(3): 628-650.

Håkansson, Gisela, Pienemann, Manfred, & Sayehli, Susan. (2002). Transfer and typological proximity in the context of second language processing. *Second Language Research*. 18(3): 250-273.

Hatch, Evelyn, & Lazaraton, Anne. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. Heinle & Heinle: Boston.

Hulstijn, Jan. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research*. 18(3): 193-223.

Hunt, Kellogg W. (1965). *Grammatical Structures Written at Three Grade Levels*. National Council of Teachers of English: Champaign, Ill.

Jang, Eunice E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.). *Towards adaptive CALL: Natural language processing for diagnostic language assessment*. Ames, IA: Iowa State University. 117-131

Jang, Eunice E. (2009a). Demystifying a Q-Matrix for Making Diagnostic Inferences About L2 Reading Skills. *Language Assessment Quarterly*. 6(3): 210-238

Jang, Eunice E. (2009b). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*. 26(1): pp. 31-73.

Jansen, Louise. (2008) Acquisition of German word order in tutored learners: A cross-sectional study in a wider theoretical context. *Language Learning: A Journal of Research in Language Studies*. 58(1): pp. 185-231.

Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika*. 39(1). 31-36.

Koizumi, Rie, Sakai , Hideki, Ido, Takahiro, Ota, Hiroshi, Hayama, Megumi,

Sato, Masatoshi, & Nemoto, Akiko. (2011). Development and Validation of a Diagnostic Grammar Test for Japanese Learners of English. *Language Assessment Quarterly*. 8(1): 53-72

Kroeger, Paul R. (2004) *Analyzing Syntax: A Lexical-functional Approach*. Cambridge UP: New York.

Kuhn, Thomas S. (1970). Logic of Discovery or Psychology of Research?. in *Philosophy of Science: The Central Issues*. Ed. Martin Curd & J.A. Cover. W.W. Norton: New York. 11-20.

Kunnan, Antony J., & Jang, Eunice E. (2009) Diagnostic Feedback in Language Assessment. in *The Handbook of Language Teaching*. Ed. Michael H. Long and Catherine J. Doughty. Blackwell: New York. 610-625.

Lee, Y.-W. & Sawaki, Y. (2009). Cognitive Diagnosis and Q-Matrices in Language Assessment. *Language Assessment Quarterly*. 6(3): 169-171.

Levelt, Willem J.M. (1989) *Speaking: From Intention to Articulation*. MIT: Cambridge.

Lord, Fredric. (1952). *A Theory of Test Scores*. Psychometric Society: ETS.

Messick, S. (1995). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. 50(9): 741-749.

Mokhtari, Kouider, Niederhauser, Dale S., Beschorner, Elizabeth A., Edwards, Patricia A. (2011). FAD: Filtering, Analyzing, and

Diagnosing Reading Difficulties. *The Reading Teacher*. 64(8). 631-635.

Nation, I. S. Paul & Beglar, D. (2007) A vocabulary size test. *The Language Teacher* 31(7): 9-13.

Nehm, Ross H., Beggrow, Elizabeth P., Opfer, John E., & Ha, Minsu. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher*. 74(2): 92-98.

Norris, John M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. Long (Eds.), The handbook of second language acquisition. Cambridge: Blackwell. 717-761.

Norris, John. M. (2005). Using developmental sequences to estimate ability with English grammar: Preliminary design and investigation of a web-based test. *Second Language Studies*. 24(1): 24-128.

Ong, Justina, Zhang, Lawrence J. (2010). Effects of Task Complexity on the Fluency and Lexical Complexity in EFL Students' Argumentative Writing. *Journal of Second Language Writing*. 19(4): 218-233.

Pallotti, Gabrielle. (2009) CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics*. 30(4): 590-601.

Pienemann, Manfred, & Keßler, Jörg-U. (2011) *Studying Processability Theory*. John Benjamins: Philadelphia.

Pienemann, Manfred. (1998) *Language Processing and Second Language Development: Processability Theory*. John Benjamins: Philadelphia.

Pienemann, Manfred. (2002). Issues in second language acquisition and language processing. *Second Language Research*. 18(3): 189–192

Pienemann, Manfred. (2005). *Cross-Linguistics Aspects of Processability Theory*. John Benjamins: Philadelphia.

Popper, Karl. (1963) Science: Conjectures and Refutations. in *Philosophy of Science: The Central Issues*. Ed. Martin Curd & J.A. Cover. W.W. Norton: New York. 3-10.

Purpura J (2004). *Assessing Grammar*. Cambridge UP: Cambridge.

Rapid Profile Test. Accessed online at: http://kw.uni-paderborn.de/institute-einrichtungen/institut-fuer-anglistik-und-amerikanistik/personal/pienemann/rapid-profile/

Richards, Brian J. (2008) Formative assessment in teacher education: The development of a diagnostic language test for trainee teachers of German. *British Journal of Education Studies*. 56(2): 184-204

Rupp, André A., Templin, Jonathon, & Henson, Robert A. (2010) *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press: New York.

Sakai, Hideki. (2008) An analysis of Japanese university students' oral performance in English using Processability Theory. *System*. 36(4): 534-549.

Salameh, Eva-Kristina, Håkansson, Gisela, & Nettelbladt, Ulrika. (2004). Developmental perspectives on bilingual Swedish-Arabic children with and without language impairment: a longitudinal study.

International Journal of Language & Communication Disorders. 39(1): 65–91.

Schumacker, RE, & Muchinsky, PM. (1996). Disattenuating correlation coefficients. From http://www.rasch.org/rmt/rmt101g.htm

Sesli, Ertugrul, & Kara, Yilmaz. (2012). Development and application of a two-tier multiple-choice diagnostic test for high school students' understanding of cell division and reproduction. *Journal of Biological Education*. 46(4): 214-225

Shohamy, Elana. (1982). Affective Considerations in Language Testing. *The Modern Language Journal*. 66(1): 13-17

Shohamy, Elana. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal*. 76(4): 513-521

Simpson, Mary, & Arnold, Brian. (1983). Diagnostic tests and criterion-referenced assessments: Their contribution to the resolution of pupil learning difficulties. 20(1): 36-42.

Skehan, Peter. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*. 30/4, 510-532.

Stone, Clement A., Ye, Feifei, Zhu, Xiaowen, & Lane, Suzanne. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. Applied Measurement in Education. 23(1): 63-86.

Tavakoli, Parvanine, Foster, Pauline. (2011). Task Design and Second Language Performance: The Effect of Narrative Type on Learner Output. *Language Learning*. 61(s1): 37-72.

White, Allen L. (2005). Active mathematics in classrooms: Finding out why children make mistakes–And then doing something to help them. *Square One*. 15(4): 15-19. Accessed online at: http://www.decd.sa.gov.au/northernadelaide/files/links/newman2.pdf

Yin, Muchun (2010). Informing pedagogy with diagnostic language test results: A language assessment action research study. *International Journal of Educational and Psychological Assessment*. 6(2): 3-20.

Appendix A—Final Grammar Test (With Target Answers)

Grammar Test (문법 테스트)

You have 30 minutes to complete the test. 30 분 안에 모든 테스트를 마쳐야 합니다.

Part 1

Fill in the blanks with the correct determiner, either "a" or "the" or nothing. Some blanks may NOT need a determiner.

빈칸에 알맞은 한정사를 기입해주세요. "a" 또는 "the" 중에 기입해주시길 바라며 한정사를 **기입하지 않아도** 되는 빈칸도 있습니다.

1. Who is that beautiful woman?
   She is __the__ best actress in __a (the)__ popular new movie.

2. What does that store sell?
   They sell ____ furniture and ____ clothing.

3. What is inside that bag?
   It is __a (the)__ gift for __the/an__ English teacher.

4. What is in those bottles?
   This bottle has ____ water, and this bottle has ____ juice.

5. What should we bring to the party tomorrow?
   I will bring __a (the)__ big cake and you can bring ____ fruit.

Part 2

Fill in the blanks with the correct words to describe each picture. You will need 1 number (one, two, three, four) and 1 noun. The noun is given to you, but you might have to change the form of the noun.

그림을 묘사할수 있는 알맞은 단어를 사용하여 빈칸을 채워주세요. 갯수를 표현하는 숫자 (one, two, three, 그리고 four 와 같은) 와 명사 하나를 사용하셔야 합니다. 명사는 이미 주어졌으나, 명사의 형태를 적절하게 변경해야 할 수도 있습니다.

Example: The man bought __one jacket__ at the store. (jacket)

1. The woman walked with __two dogs__ yesterday. (dog)

2. Josh gave his teacher __four flowers__ for her birthday. (flower)

3. I read __three books__ for my English class. (book)

4. I have __four pictures__ of my family at home. (picture)

5. I saw __two cars__ in an accident last week. (car)

Part 3

Fill in each blank with the correct verb to describe the picture. You will need 1 pronoun (he, she, they) and 1 verb. Write the verb in the present tense form. Use the pictures to get the correct words.

빈칸에 알맞은 단어를 기입해 주세요. 대명사 (he, she, they) 그리고 동사 하나가 필요합니다. 동사는 **현재 시제입니다**. 그림들을 활용하여 알맞은 단어를 넣어주시길 바랍니다.

Example: __They play__ soccer after school every day. (play)

1. __They sing__ songs at school every week.

2. __She (He) writes__ letters to her friends every day.

3. __They (She) make(s)/bake(s)/etc...__ cookies for fun on the weekend.

4. __He sleeps/rests/etc...__ in his little bed every night.

5. __She talks/etc...__ her mother on Sundays.

Part 4

Fill in the blanks with the correct tense form. Some sentences are in the past tense, and some sentences are in the present continuous tense. Read each sentence carefully, and write the correct form in the blank. The verb is given to you, but you have to change the form of the verb.

빈칸을 정확한 동사의 올바른 시제를 사용하여 넣어주세요. 몇몇 문장은 **과거시제를** 사용하여야 하며 몇몇 문장은 **과거진행형 시제를** 사용하여야 합니다. 각 문장을 유심히 읽고, 빈칸을 올바르게 채워주세요. 문장의 옆에 동사가 주어졌으나 동사의 형태를 변형하여야 합니다.

Examples: Mary __walked__ to her friend's house last night. (walk)
Two people __are walking__ to their friend's house now. (walk)

1. Last night, Jenny __chatted__ on the Internet. (chat)

2. Right now, many people __are riding__ this bus. (ride)

3. That man __is watching__ a TV show now. (watch)

4. Several people __ate__ a large dinner last weekend. (eat)

5. Yesterday afternoon, Cindy ___packed___ her bags for a trip. (pack)

6. This woman ___is playing___ basketball now. (play)

7. Last Friday, John and Jack ___ran___ in a race. (run)

8. Only one person ___picked___ up medicine yesterday. (pick)

9. Jenny and Judy ___are carrying___ some packages now. (carry)

10. Right now, these men ___are moving___ a big chair. (move)

Part 5

Fill in each blank with the correct verb to describe the picture. You should write each verb in the present tense form. Use the picture to get the correct word.

그림을 잘 묘사할수 있는 알맞은 동사를 사용하여 빈칸을 채워주세요. 동사의 **현재형 시제를** 사용하여 주세요. 그림을 잘 활용하여 알맞은 단어를 사용하여야 합니다.

Example: The children ___jump___ every day.

1. The girl ___plays___ her guitar every day.

2. The man ___opens/etc...___ his umbrella when it rains.

3. The kids ___wash/clean/etc...___ cars every weekend.

4. The boy and girl ___drink___ tea every afternoon.

5. Santa ___gives/etc...___ gifts to children every Christmas.

Part 6

Put the words in the correct order. You must use ALL of the words. Do not change the form of any words.

단어들을 알맞은 순서로 배열해주세요. **모든** 단어들을 사용하셔야 합니다. 단어의 형태를 변형 하지는 마세요.

saw | a bird | Jenny

Example: Jenny saw a bird in the park today.

Jane | studying | Math | was

1. Jane was studying Math very hard all night.

washing | their car | Alex and Peter | were

2. Alex and Peter were washing their car last weekend

pictures | Betty and Sue | were | taking

3. Betty and Sue were taking pictures in the park yesterday.

was | Kevin | soda | drinking

4. Kevin was drinking soda during lunch.

Part 7

Put the words in the correct order. You must use more than one word, but you MIGHT NOT use ALL of the words. Do not change the form of any words.

단어들을 올바른 순서로 배치하세요. 하나의 단어 이상을 사용해야 하지만 **모든 단어들을 사용하지는 않아도 됩니다**. 단어들의 형태는 바꾸지 마세요.

Example:

this | school | to | after | afternoon

What do you want to do today?

Let's go to a movie after school this afternoon.

on Mondays | lunch | before | the

1. When does your class have Math class?

We have it before lunch on Mondays / on Mondays before lunch.

the | later | storm | during

2. Did you go outside when it was raining?

No, I stayed home during the storm.

school | after | and | every day

3. What musical instrument do you practice?

I practice piano after school every day / every day after school.

before | 9 o'clock | when | last night

4. When did you finish your homework?

I finished it before 9 o'clock last night / last night before 9 o'clock.

while | winter | during | the

5. Did you do anything fun in Canada?

I went skiing during the winter.

Part 8

Put the words in the correct order. You must USE ALL of the words, and you might have to add more words. Do not change the form of any words.

단어들을 올바른 순서로 배치하세요. 주어진 모든 단어들을 사용해야 하며, 다른 단어들을 추가하여야 할 수도 있습니다. 단어들의 형태는 바꾸지 마세요.

Example:


8:00 am | 9:00 am

breakfast | before | eats

Every day, Mary exercises before she eats breakfast.


10:00 am | 10:00 am

playing | tennis | John | was

1. Today, Susan decorated a cake while John was playing tennis.


8:00 am | 9:00 am

on the phone | Scott | before | talking

2. Scott was talking on the phone before he went to school this morning.


3:00 pm | 1:00 pm

homework | he | his | finished

3. This afternoon, Mark took a nap after he finished his homework.

to music | this | listening | while | girl

4. This girl was listening to music while she folded her clothes last night.

Part 9

Put the words in the correct order. You must use more than one word, but you might not use all of the words. You MIGHT have to CHANGE the form of some words.

단어들을 올바른 순서로 배치하세요. 하나의 단어 이상이 필요하지만 모든 단어들을 사용하지 않아도 됩니다. 몇 몇 단어들의 형태를 바꿀 수도 있습니다.

Example:

history | teach | he | every | when

The teacher reads the textbook when he teaches history.

after | clean | my | I | room | the

5. What will you do this afternoon?

I want to read a book after I clean my room.

at the theater | they | while | a movie | watching | and

6. The couple ate popcorn while they watched a movie at the theater / at the theater while they watched a movie.

shopped | while | then | are | you

7. Can I get something for you at the store?

Please get me some batteries while you are shopping.

new dress | she | putting on | her | also | after

8. Greg took a picture of Tina after she put on her new dress.

Part 10

Put the words in the correct order. You must USE ALL of the words, and you MIGHT HAVE TO ADD more words. Also, you MIGHT have to CHANGE the form of some words.

올바른 순서로 단어들을 배치하세요. **모든 단어들을** 다 사용하여야만 하며, **몇 몇 단어들을 추가할** 필요도 있습니다. 몇 몇 단어들의 형태는 바뀔수 도 있습니다.

Example:

finishing | my homework | after

Can you help me move this box?

Yes, I can help you after I finish my homework.

bought | can | one | I

9. Are you going to get a new car soon?

No, I have to save more money before I can buy one.

12:00 pm

11:00 am

them | up the mountain | after | yesterday

10. Mark and Ellen had a picnic (yesterday) after they climbed (etc...) up the mountain (yesterday).

study | was | she | in school

11. How did Mary pay for her university degree?

She taught after-school classes while she was studying in school.

2:00 pm

4:00 pm

tree | plant | Mark | before | in the yard

12. Last Saturday, Jane painted the outside of the house before Mark planted a tree in the yard.

This is the end of the test. 여기가 테스트의 마지막 입니다.

Please answer the questions on the next page. 다음 페이지의 질문들에 대한 답변을 적어주세요.

Appendix B—Final Writing Test

Writing Test—작문테스트

Look at the pictures on page 2. Then write a story based on the pictures. Your story should be at least 150 words long. Write the story in the past tense. You may use the back of this page if you need more paper.
2 페이지 있는 그림들을 바탕으로 스토리를 작성하세요. 공간이 필요할 경우 뒷 면을 활용하시면 됩니다. 스토리는 적어도 150 단어 이상으로 작성해주셔야 합니다. 스토리를 **과거시제를** 활용하여 작성해 주세요. 공간이 더 필요할 경우 뒷 페이지를 사용하여도 좋습니다.

Some helpful words are given to you on page 3. 도움이 되는 단어들을 3 페이지에 제공하였습니다.

The first two sentences are given to you below. You have 15 minutes to complete the whole test.
첫 문장은 아래와 같이 제공됩니다. 15 분 안에 모든 테스트를 마쳐야 합니다.

Brad had a bad day yesterday. He was hurrying because he had to get to school at 9 am.

Backpack (배낭)

Bus stop (버스정류장)

Crossing guard (건널목 안전요원)

Chase (쫓아가다)

School nurse (학교간호사)

Bandages (붐 대)

Hurt (다치다)

Yell at (혼내다)

Appendix C
PT Scoring Sheet & Proficiency Scoring Rubric

Scoring Sheet ID #:_____

Grammar Test

Sect.	Pg	Item #	Target	No Ans.	Incor.	Non-target	Underuse	Overuse	Other
Determiner	1	1a							
		1b							
		3a							
		3b							
		5a							
No Determ	1	2a							
		2b							
		4a							
		4b							
		5b							
Plural Noun	2	1							
		2							
		3							
		4							
		5							
Past tense	4-5	1							
		4							
		5							
		7							
		8							
Pres. Cont.	4-5	2							
		3							
		6							
		9							
		10							
S/V Agree—Sing	3	2							
		4							
	4-5	5							
		3							
		6							
S/V Agree—Pl	6	1							
		2							
		5							
	3	1							
		3							
	4-5	2							
		9							
		10							
	6	3							
		4							

Scoring Sheet ID #:_____

Sect.	Pg	Item #	Target	No Ans.	Incor.	Non-target	Underuse	Overuse	Other
Preposit	8	1							
		2							
		3							
		4							
		5							
Subordinate Clause	9-14	1							
		2							
		3							
		4							
		5							
		6							
		7							
		8							
		9							
		10							
		11							
		12							

Writing Test	Contxts	Correct (Target)	Missing	Incor.	Non-target	Underuse	Overuse	Other
Determiner								
No Determiner								
Plural/Non-C Noun								
Past								
Pr./Past Cont.								
S/V Agr. (Sg)								
S/V Agr. (Pl)								
Prep. (bef/aft/dur)								
Sub. Clauses (bef/aft/wh)								

Dep. Clauses (Total)		3	2	1	# T-Units:	#Words:

Additional Comments:

Writing Proficiency Scoring Rubric

The following rubric is a 10-point scale rubric based on the American Council for Teachers of Foreign Languages (ACTFL) guidelines for writing proficiency. For the original version, which was based more on academic/business writing, see http://actflproficiencyguidelines2012.org/writing#Intermediate

One important point to keep in mind when scoring: students were given the following words: backpack; bus stop; crossing guard; chase; school nurse; bandages; hurt; yell at. The student must show variety beyond these words in order to be considered as having varied vocabulary. For example, a student who uses both “backpack” and “bag” alternately likely has greater vocabulary depth than a student who only uses “backpack”.

Writing » Superior (10)

Writers at the Superior level develop topics in a way that moves beyond the concrete to the abstract. They demonstrate the ability to explain complex matters, and to present and support opinions by developing cogent arguments and hypotheses. Their treatment of the topic is enhanced by the effective use of structure, lexicon, and writing protocols. They organize and prioritize ideas to convey to the reader what is significant. The relationship among ideas is consistently clear, due to organizational and developmental principles (e.g., cause and effect, comparison, chronology). These writers are capable of extended treatment of a topic which typically requires at least a series of paragraphs.

Writers at the Superior level demonstrate a high degree of control of grammar and syntax, of both general and specialized vocabulary, of spelling or symbol production, of cohesive devices, and of punctuation. Their vocabulary is precise and varied. Writers at this level direct their writing to their audiences; their writing fluency eases the reader’s task.

At the Superior level, writers demonstrate no pattern of error; however, occasional errors may occur, particularly in low-frequency structures. When present, these errors do not interfere with comprehension, and they rarely distract the native reader.

Writing » Advanced (7-9)

Writers at the Advanced level can narrate and describe in the major time frames of past, present, and future, using paraphrasing and elaboration to provide clarity. Advanced-level writers produce connected discourse of paragraph length and structure. At this level, writers show good control of the most frequently used structures and generic vocabulary, allowing them to be understood by those unaccustomed to the writing of non-natives.

Advanced High (9)

Writers at the Advanced High sublevel are able to write with significant precision and detail. Their writing tends to emphasize the concrete aspects of topics. Advanced High writers can narrate and describe in the major time frames, with solid control of aspect. In addition, they are able to demonstrate the ability to sometimes write at a Superior level, but are not able to maintain this level consistently. They have good control of a range of grammatical structures and a fairly wide general vocabulary. They often show remarkable ease of expression, but patterns of error appear. The linguistic limitations of Advanced High writing may occasionally distract the native reader from the message.

Advanced Mid (8)

Writers at the Advanced Mid sublevel demonstrate the ability to narrate

and describe with detail in all major time frames with good control of aspect. Their writing exhibits a variety of cohesive devices in texts up to several paragraphs in length. There is good control of the most frequently used target-language syntactic structures and a range of general vocabulary. Most often, thoughts are expressed clearly and supported by some elaboration. This writing incorporates organizational features both of the target language and the writer's first language and may at times resemble oral discourse. Writing at the Advanced Mid sublevel is understood readily by natives not used to the writing of non-natives.

Advanced Low (7)

Writers at the Advanced Low sublevel demonstrate the ability to narrate and describe in major time frames with some control of aspect. They are able to combine and link sentences into texts of paragraph length and structure. Their writing, while adequate to satisfy the criteria of the Advanced level, may not be substantive. Writers at the Advanced Low sublevel demonstrate the ability to incorporate a limited number of cohesive devices, and may resort to some redundancy and awkward repetition. They rely on patterns of oral discourse and the writing style of their first language. These writers demonstrate minimal control of common structures and vocabulary associated with the Advanced level. Their writing is understood by natives not accustomed to the writing of non-natives, although some additional effort may be required in the reading of the text.

Writing » Intermediate (4-6)

Writers at the Intermediate level can create with the language and communicate simple facts and ideas in a series of loosely connected sentences. They write primarily in present time. At this level, writers use basic vocabulary and structures to express meaning that is comprehensible to those accustomed to the writing of non-natives.

Intermediate High (6)

Writers at the Intermediate High sublevel can write compositions and simple summaries related to school experiences. They can narrate and describe in different time frames when writing about everyday events and situations. These narrations and descriptions are often, but not always, of paragraph length, and they typically contain some evidence of breakdown in one or more features of the Advanced level. For example, these writers may be inconsistent in the use of appropriate major time markers, resulting in a loss of clarity. The vocabulary, grammar and style of Intermediate High writers essentially correspond to those of the spoken language. Intermediate High writing, even with numerous and perhaps significant errors, is generally comprehensible to natives not used to the writing of non-natives, but there are likely to be gaps in comprehension.

Intermediate Mid (5)

Writers at the Intermediate Mid sublevel can write short, simple communications and compositions in loosely connected texts about daily routines, common events, and other personal topics. Their writing is framed in present time but may contain references to other time frames. The writing style closely resembles oral discourse. Writers at the Intermediate Mid sublevel show evidence of control of basic sentence structure and verb forms. This writing is best defined as a collection of discrete sentences and/or questions loosely strung together. There is little evidence of deliberate organization. Intermediate Mid writers can be understood readily by natives used to the writing of non-natives.

Intermediate Low (4)

Writers at the Intermediate Low sublevel can create statements and formulate questions based on familiar material. Most sentences

are recombinations of learned vocabulary and structures. These are short and simple conversational-style sentences with basic word order. They are written almost exclusively in present time. Writing tends to consist of a few simple sentences, often with repetitive structure. Topics are tied to highly predictable content areas and personal information. Vocabulary is adequate to express elementary concepts. There may be basic errors in grammar, word choice, punctuation, spelling, and in the formation and use of non-alphabetic symbols. Their writing is understood by natives used to the writing of non-natives, although additional effort may be required.

Writing » Novice (1-3)

Writers at the Novice level can provide limited formulaic information on simple forms and documents. These writers can reproduce practiced material to convey the simplest concepts. In addition, they can transcribe familiar words or phrases, copy letters of the alphabet or syllables of a syllabary, or reproduce basic characters with some accuracy.

Novice High (3)

Writers at the Novice High sublevel are able to express themselves within the context in which the language was learned, relying mainly on practiced material. Their writing is focused on common elements of daily life. Novice High writers are able to recombine learned vocabulary and structures to create simple sentences on very familiar topics, but are not able to sustain sentence-level writing all the time. Due to inadequate vocabulary and/or grammar, writing at this level may only partially communicate the intentions of the writer. Novice High writing is often comprehensible to natives used to the writing of non-natives, but gaps in comprehension may occur.

Novice Mid (2)

Writers at the Novice Mid sublevel can reproduce from memory a modest number of words and phrases in context. They can supply limited information on simple forms. Novice Mid writers exhibit a high degree of accuracy when writing on well-practiced, familiar topics using limited formulaic language. With less familiar topics, there is a marked decrease in accuracy. Errors in spelling or in the representation of symbols may be frequent. There is little evidence of functional writing skills. At this level, the writing may be difficult to understand even by those accustomed to non-native writers.

Novice Low (1)

Writers at the Novice Low sublevel are able to copy or transcribe familiar words or phrases, form letters in an alphabetic system, and copy and produce isolated, basic strokes in languages that use syllabaries or characters. Given adequate time and familiar cues, they can reproduce from memory a very limited number of isolated words or familiar phrases, but errors are to be expected.

Appendix D
Inter-Item Correlation Matrix

Inter-item Correlation Matrix

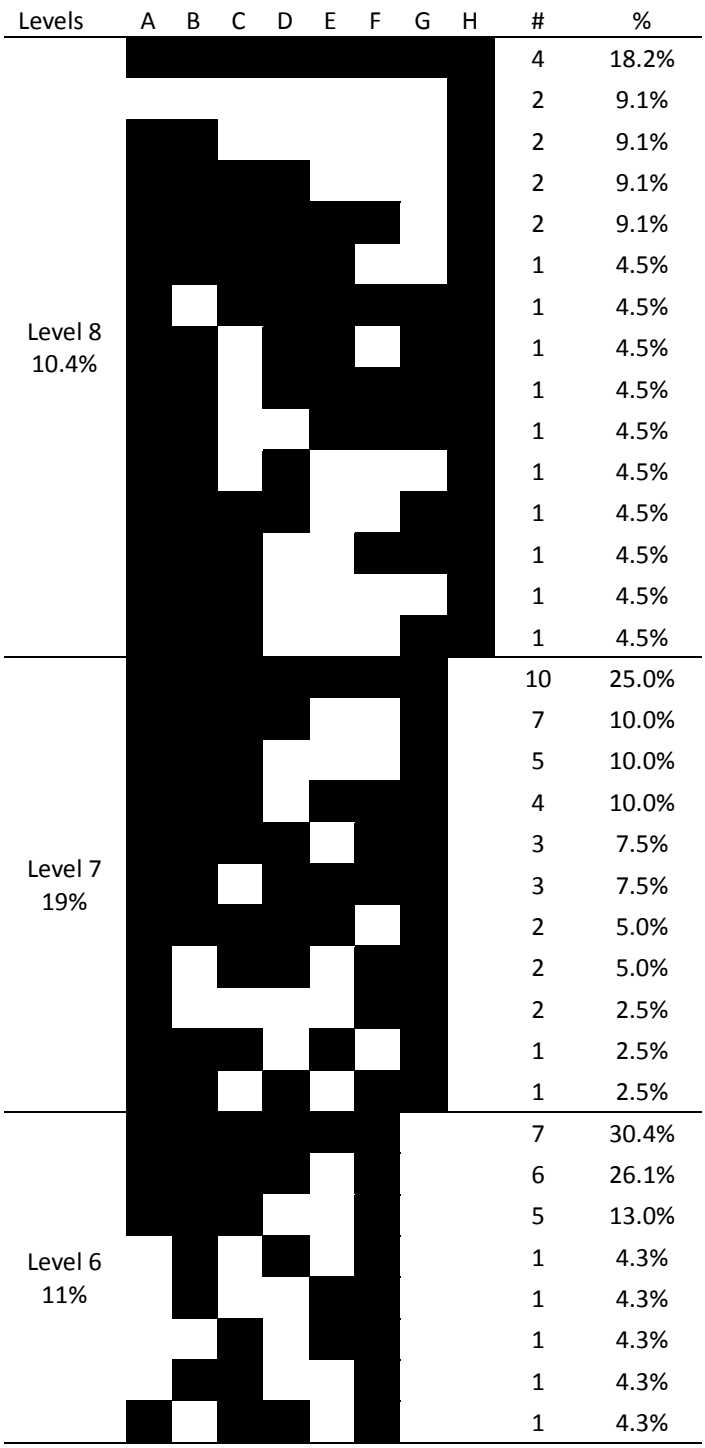
	Det1	Det2	Det3	Det4	Det5	NC1	NC2	NC3	NC4	NC5	PlurN1	PlurN2	PlurN3	PlurN4	PlurN5	Past1	Past2	Past3	Past4	Past5	PrCont 1	PrCont 2	PrCont 3	PrCont 4	PrCont 5
Det1	1.00																								
Det2	-0.07	1.00																							
Det3	-0.03	0.21	1.00																						
Det4	0.04	0.17	-0.03	1.00																					
Det5	0.14	0.14	0.23	0.13	1.00																				
NC1	0.09	-0.08	-0.05	-0.10	-0.05	1.00																			
NC2	0.10	0.01	-0.03	-0.21	0.07	0.22	1.00																		
NC3	0.34	-0.09	-0.06	-0.07	0.02	0.26	0.22	1.00																	
NC4	0.25	-0.09	-0.03	-0.09	0.05	0.17	0.20	0.82	1.00																
NC5	0.06	-0.19	-0.05	-0.21	-0.15	0.29	0.17	0.25	0.23	1.00															
PlurN1	0.08	0.14	0.03	0.06	0.14	0.03	0.00	0.16	0.19	-0.06	1.00														
PlurN2	0.13	0.18	0.04	0.12	0.19	0.08	0.04	0.13	0.15	0.02	0.61	1.00													
PlurN3	0.12	0.09	-0.02	0.08	0.08	0.12	0.01	0.18	0.20	0.04	0.66	0.62	1.00												
PlurN4	0.10	0.04	0.03	-0.05	0.01	0.07	0.03	0.19	0.21	0.10	0.55	0.49	0.68	1.00											
PlurN5	0.19	0.15	0.01	0.10	0.15	0.05	0.06	0.19	0.24	0.12	0.58	0.62	0.64	0.59	1.00										
Past1	-0.02	0.07	0.01	-0.01	0.00	0.03	0.03	0.15	0.19	-0.02	0.20	0.23	0.27	0.35	0.29	1.00									
Past2	0.09	0.03	-0.01	-0.11	-0.02	-0.01	0.12	0.18	0.19	0.02	0.18	0.13	0.18	0.27	0.16	0.37	1.00								
Past3	0.03	0.04	0.03	-0.05	0.01	0.01	0.06	0.17	0.19	-0.07	0.26	0.19	0.23	0.38	0.32	0.72	0.44	1.00							
Past4	0.15	-0.02	0.02	-0.08	-0.01	0.07	0.18	0.23	0.26	0.05	0.19	0.18	0.20	0.26	0.22	0.42	0.72	0.55	1.00						
Past5	0.08	0.10	0.01	0.02	0.01	0.10	0.13	0.17	0.19	-0.07	0.24	0.21	0.29	0.31	0.29	0.62	0.49	0.66	0.44	1.00					
PrCont1	0.30	-0.01	-0.03	-0.07	-0.02	0.13	0.10	0.30	0.25	0.13	0.08	0.12	0.21	0.14	0.30	0.30	0.03	0.24	0.16	0.17	1.00				
PrCont2	0.32	-0.05	0.03	-0.05	0.07	0.04	0.10	0.33	0.31	0.13	0.14	0.17	0.21	0.15	0.31	0.23	0.37	0.22	0.43	0.21	0.61	1.00			
PrCont3	0.38	-0.04	0.04	-0.04	0.08	0.12	0.11	0.41	0.37	0.15	0.14	0.18	0.24	0.23	0.34	0.26	0.32	0.25	0.40	0.27	0.61	0.87	1.00		
PrCont4	0.26	0.00	-0.03	-0.07	0.04	0.12	0.11	0.30	0.28	0.10	0.17	0.20	0.26	0.25	0.38	0.29	0.10	0.22	0.19	0.25	0.75	0.71	0.76	1.00	
PrCont5	0.27	-0.02	-0.02	-0.04	0.07	0.06	0.04	0.25	0.21	0.12	0.13	0.18	0.19	0.20	0.31	0.26	0.14	0.20	0.22	0.21	0.63	0.67	0.70	0.83	1.00
SVSing1	0.34	0.01	0.05	-0.03	0.14	0.14	0.10	0.29	0.28	0.09	0.17	0.15	0.20	0.24	0.17	0.23	0.30	0.17	0.31	0.27	0.18	0.35	0.41	0.29	0.26
SVSing2	0.26	0.05	0.07	-0.07	0.10	0.12	0.13	0.32	0.34	0.04	0.23	0.18	0.22	0.26	0.19	0.31	0.31	0.26	0.34	0.32	0.22	0.34	0.38	0.32	0.27
SVSing3	0.29	0.02	0.07	-0.08	0.11	0.14	0.12	0.38	0.39	0.11	0.21	0.17	0.27	0.31	0.27	0.31	0.34	0.26	0.35	0.36	0.22	0.34	0.41	0.30	0.28
SVSing4	0.36	-0.04	0.07	-0.14	0.09	0.11	0.17	0.45	0.42	0.14	0.16	0.17	0.16	0.23	0.19	0.26	0.28	0.21	0.34	0.24	0.23	0.37	0.41	0.34	0.35
SVSing5	0.32	0.02	0.08	-0.01	0.09	0.06	0.15	0.35	0.29	0.09	0.12	0.15	0.12	0.20	0.19	0.27	0.25	0.18	0.35	0.19	0.24	0.36	0.39	0.30	0.34
SVSing6	0.31	0.00	0.15	-0.14	0.09	0.12	0.16	0.36	0.37	0.12	0.15	0.19	0.16	0.23	0.19	0.21	0.27	0.18	0.33	0.21	0.17	0.34	0.40	0.30	0.31
SVPlur1	0.15	0.05	0.15	0.03	0.24	0.08	0.09	0.22	0.24	0.05	0.15	0.18	0.14	0.14	0.19	0.14	0.28	0.20	0.22	0.12	0.18	0.26	0.27	0.19	0.16
SVPlur2	0.09	-0.03	0.08	0.02	0.16	0.03	0.08	0.20	0.19	-0.05	0.12	0.20	0.24	0.18	0.20	0.09	0.23	0.12	0.19	0.15	0.17	0.21	0.22	0.18	0.13
SVPlur3	0.13	0.05	0.05	0.04	0.12	0.07	0.08	0.15	0.10	-0.10	0.34	0.33	0.30	0.37	0.27	0.22	0.28	0.24	0.22	0.17	0.16	0.14	0.15	0.14	0.09
SVPlur4	0.12	-0.05	0.02	0.00	0.13	0.01	0.07	0.19	0.16	0.00	0.26	0.27	0.27	0.28	0.28	0.19	0.34	0.28	0.28	0.26	0.17	0.18	0.22	0.20	0.15
Prep1	0.13	-0.02	-0.14	-0.05	0.04	0.13	0.13	0.25	0.22	0.17	0.07	0.14	0.16	0.17	0.19	0.03	0.10	0.12	0.17	0.03	0.08	0.15	0.17	0.12	0.13
Prep2	0.42	0.07	0.00	-0.05	0.15	0.12	0.16	0.38	0.28	0.17	0.19	0.21	0.20	0.26	0.31	0.17	0.31	0.23	0.36	0.19	0.38	0.49	0.51	0.37	0.43
Prep3	0.29	-0.13	-0.07	-0.02	0.13	0.15	0.16	0.24	0.21	0.11	0.06	0.01	0.03	0.07	0.11	0.08	0.10	0.17	0.20	0.04	0.19	0.20	0.20	0.19	0.20
Prep4	0.19	-0.19	-0.20	-0.01	-0.01	0.19	0.10	0.28	0.20	0.16	0.13	0.08	0.17	0.22	0.21	0.10	0.16	0.20	0.26	0.10	0.22	0.25	0.26	0.26	0.28
Prep5	0.39	-0.02	-0.10	-0.02	0.16	0.10	0.12	0.29	0.28	0.12	0.19	0.18	0.22	0.25	0.27	0.23	0.30	0.25	0.36	0.22	0.29	0.45	0.43	0.36	0.36
SCA1	0.23	-0.04	0.00	-0.02	0.11	0.04	0.04	0.20	0.17	0.06	0.11	0.11	0.14	0.14	0.17	0.01	0.14	0.03	0.17	0.07	0.15	0.23	0.25	0.16	0.19
SCA2	0.21	-0.08	-0.02	-0.14	0.11	0.04	0.12	0.25	0.27	0.10	0.03	0.07	0.09	0.09	0.14	0.11	0.27	0.07	0.25	0.15	0.17	0.30	0.29	0.21	0.23
SCA3	0.33	0.06	0.04	-0.03	0.14	0.13	0.17	0.37	0.28	0.06	0.14	0.19	0.19	0.21	0.29	0.13	0.32	0.17	0.35	0.23	0.32	0.41	0.45	0.35	0.33
SCA4	0.20	-0.04	0.03	-0.09	0.15	0.06	0.17	0.23	0.24	0.08	0.02	0.03	0.07	0.14	0.10	0.10	0.15	0.08	0.18	0.13	0.14	0.23	0.24	0.16	0.18
SCB1	0.18	-0.02	0.03	-0.02	0.17	0.02	0.08	0.27	0.32	0.07	0.14	0.18	0.15	0.21	0.23	0.19	0.19	0.25	0.31	0.22	0.16	0.15	0.19	0.17	0.15
SCB2	0.23	0.07	0.06	-0.03	0.09	0.07	0.13	0.24	0.26	0.02	0.13	0.21	0.10	0.14	0.20	0.16	0.21	0.21	0.29	0.21	0.14	0.21	0.22	0.18	0.17
SCB3	0.27	0.02	-0.06	0.06	0.16	0.05	0.20	0.38	0.34	0.14	0.09	0.18	0.07	0.11	0.20	0.14	0.27	0.16	0.31	0.21	0.18	0.25	0.30	0.18	0.21
SCB4	0.11	-0.01	-0.05	-0.06	0.09	0.12	0.15	0.09	0.16	0.09	-0.04	0.07	0.05	0.05	0.08	0.06	0.10	0.09	0.11	0.08	0.15	0.15	0.14	0.09	0.12
SCC1	0.10	-0.05	0.08	0.00	0.06	0.07	0.08	0.16	0.06	-0.02	0.06	0.09	0.08	0.10	0.10	0.05	0.01	0.04	0.12	0.01	0.14	0.12	0.16	0.17	0.19
SCC2	0.21	-0.02	0.04	-0.07	0.08	0.07	0.11	0.11	0.07	0.06	0.08	-0.01	0.10	0.12	0.07	0.03	0.06	0.07	0.16	0.05	0.11	0.10	0.14	0.09	0.16
SCC3	0.20	-0.04	0.02	-0.12	-0.01	0.11	0.12	0.18	0.18	0.22	-0.01	0.09	0.08	0.05	0.10	0.06	0.07	0.05	0.17	0.02	0.15	0.18	0.22	0.14	0.20
SCC4	0.16	-0.05	0.08	-0.18	0.10	0.06	0.14	0.19	0.22	0.17	0.05	0.14	0.08	0.12	0.12	0.14	0.18	0.12	0.23	0.12	0.15	0.19	0.22	0.13	0.16

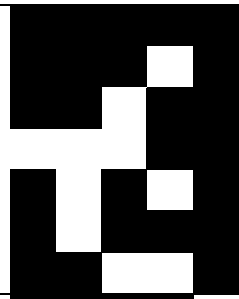
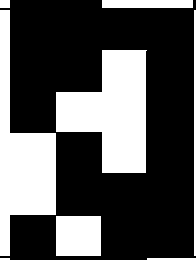
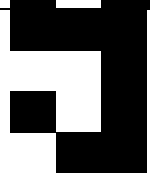
Inter-item Correlation Matrix

SVSing1	SVSing2	SVSing3	SVSing4	SVSing5	SVSing6	SVPlur1	SVPlur2	SVPlur3	SVPlur4	Prep1	Prep2	Prep3	Prep4	Prep5	SCA1	SCA2	SCA3	SCA4	SCB1	SCB2	SCB3	SCB4	SCC1	SCC2	SCC3	SCC4
1.00																										
0.86	1.00																									
0.79	0.85	1.00																								
0.60	0.55	0.61	1.00																							
0.51	0.48	0.53	0.82	1.00																						
0.51	0.52	0.58	0.81	0.70	1.00																					
0.26	0.28	0.20	0.19	0.16	0.16	1.00																				
0.16	0.11	0.11	0.14	0.10	0.11	0.52	1.00																			
0.12	0.12	0.05	0.12	0.12	0.07	0.33	0.31	1.00																		
0.24	0.21	0.20	0.19	0.17	0.14	0.41	0.29	0.56	1.00																	
0.17	0.20	0.22	0.16	0.14	0.25	0.02	0.10	0.10	0.09	1.00																
0.34	0.32	0.37	0.48	0.42	0.39	0.24	0.17	0.22	0.21	0.26	1.00															
0.25	0.20	0.22	0.28	0.19	0.21	0.13	0.06	0.18	0.12	0.37	0.38	1.00														
0.22	0.19	0.21	0.20	0.15	0.15	0.15	0.12	0.14	0.11	0.26	0.30	0.45	1.00													
0.33	0.26	0.30	0.42	0.37	0.33	0.17	0.10	0.13	0.14	0.16	0.62	0.43	0.37	1.00												
0.18	0.18	0.22	0.22	0.19	0.23	0.13	0.08	0.09	0.12	0.19	0.35	0.13	0.10	0.27	1.00											
0.24	0.24	0.27	0.32	0.23	0.25	0.14	0.09	0.07	0.13	0.23	0.37	0.26	0.22	0.37	0.26	1.00										
0.32	0.29	0.29	0.38	0.28	0.31	0.24	0.24	0.28	0.27	0.19	0.51	0.23	0.24	0.40	0.31	0.42	1.00									
0.21	0.21	0.26	0.23	0.15	0.21	0.10	0.04	0.06	0.08	0.24	0.29	0.26	0.15	0.32	0.24	0.75	0.36	1.00								
0.23	0.23	0.27	0.34	0.25	0.28	0.13	0.02	0.17	0.21	0.22	0.24	0.24	0.17	0.33	0.26	0.34	0.36	0.30	1.00							
0.28	0.24	0.29	0.37	0.24	0.33	0.15	0.11	0.19	0.15	0.23	0.36	0.34	0.24	0.34	0.24	0.41	0.34	0.42	0.56	1.00						
0.24	0.20	0.27	0.35	0.26	0.29	0.15	0.07	0.16	0.18	0.21	0.51	0.36	0.28	0.38	0.30	0.38	0.42	0.34	0.43	0.61	1.00					
0.07	0.08	0.10	0.09	0.06	0.09	0.13	0.14	-0.01	0.03	0.18	0.09	0.21	0.31	0.20	0.18	0.26	0.03	0.19	0.18	0.24	0.28	1.00				
0.10	0.10	0.17	0.14	0.13	0.14	-0.02	-0.01	0.09	0.09	0.12	0.21	0.08	0.01	0.13	0.14	0.25	0.26	0.32	0.08	0.17	0.27	0.03	1.00			
0.21	0.25	0.28	0.26	0.19	0.19	0.02	-0.01	0.01	0.01	0.08	0.27	0.22	0.25	0.23	0.03	0.36	0.20	0.28	0.13	0.22	0.30	0.15	0.34	1.00		
0.21	0.21	0.23	0.21	0.15	0.21	0.10	0.06	-0.08	-0.02	0.23	0.22	0.10	0.23	0.15	0.27	0.33	0.23	0.24	0.19	0.24	0.24	0.36	0.29	0.32	1.00	
0.29	0.23	0.30	0.36	0.30	0.34	0.16	0.14	0.11	0.15	0.15	0.25	0.16	0.27	0.24	0.17	0.37	0.24	0.27	0.18	0.35	0.25	0.26	0.16	0.30	0.29	1.00

Appendix E
Implicational Hierarchy for the Full Proposed Grammar

Grammar Test:



Level 5 10%		11	52.4%
		4	19.0%
		2	9.5%
		1	4.8%
		1	4.8%
		1	4.8%
		1	4.8%
Level 4 12.9%		14	51.9%
		7	25.9%
		3	11.1%
		1	3.7%
		1	3.7%
		1	3.7%
Level 3 22.9%		27	56.3%
		9	18.8%
		8	16.7%
		4	8.3%
Level 2 8.6%		15	83.3
		3	16.7
Level 1 2.4%		5	100.0%
Level 0 2.9%		6	100.0%

Writing Test:

Levels	A C D E F G H I	#	%
Level 8 0%		0	0.0%
Level 7 0%		0	0.0%
Level 6 0%		0	0.0%
Level 5 0%		0	0.0%
Level 4 1.4%		0 3	0.0% 100.0%
Level 3 0.9%		0 2	0.0% 100.0%
Level 2 2.3%		4 1	80.0% 20.0%
Level 1 49.8%		109	100.0%
0 Levels 45.7%		100	100.0%

Appendix F
Examples of Feedback

Paragraph Style:

CH108 has acquired determiners with count nouns. CH108 has partially acquired making the plural form of count nouns, present continuous tense, plural subject verb agreement, and the prepositions "before", "after", and "during", though there was insufficient evidence in the writing sample to confirm full acquisition. CH108 has partially acquired the past tense, though CH108 sometimes gets confused about which tense to use in various situations. Focused worksheets and writing exercises should help with this. CH108 has not acquired non-count nouns, tending to use determiners when they are not needed, especially "the". CH108 has not acquired 3rd person singular subject/verb agreement, tending to drop the "s" on the verb. CH108 has not acquired subordinate clauses using "before", "after", and "while"; CH108 gets confused about which conjunction to use, tends to overuse "when", and also has problems with tense. Focused worksheets should help with these grammar points.

Report Card Style:

IN231					
Grammar	Acquired	Part Acquired	Not Acquired	Problem	Recommendation
Determiners + count nouns	✓				
Non-count nouns			✓	Uses determiners when not needed, especially "a"	Worksheets
Plural form of count nouns		✓		Not enough writing	Writing practice
Past tense	✓				
Present continuous		✓		Not enough writing	Writing practice
3rd person singular s/v			✓	Does not put "s" on the verb	Worksheets
Plural subject/verb agreement		✓		Not enough writing	Writing practice
Prepositions: before, after, during			✓	Sometimes incorrect word order	Writing practice
Subordinate clauses: before, after, while		✓		Sometimes unsure of conjunction; avoids in writing	Student can write other complex forms (when); needs to practice more variety in writing.

국문초록

처리가능성 이론에 기반을 둔 진단적 영문법 평가 시험 개발 연구

서울대학교 대학원
영어영문학과
로살리 허쉬
(Rosalie Hirch)

진단 평가를 개발하기 위해서 가장 중요한 고려사항은 진단평가 시험을특정 교수학습 과정에 연계시킬 것인지 아니면 습득이론에 기반을 둔 시험으로 개발할 것인가의 여부이다. 최근 평가학자들과 제 2언어 학습 연구자들 사이에서 가장 많은 관심을 끌고 있는 문법 습득 이론은 Manfred Pienemann(1989)의 처리가능성이론(Processability Theory)이다. 이 이론에 바탕을 둔 Rapid Profile 진단평가가 이미 개발되어 현재 사용 중이나, 이러한 진단평가의 큰 한계점 중의 하나는 말하기 평가 시험 형태를 띠고 있어서 외국에서는 적용하기가 어려운 점이 있다. 게다가 이러한 진단평가는 수험자에게 원어민 또는 원어민에 가까운 능숙도를 요구함으로 한국, 중국, 일본과 같은 나라에서는 제한적으로만 사용할 수 있을 것이다.

그동안 여러 연구자들이 처리가능성 이론에 기반을 둔 다른 여러러 과제들을 활용한 시험들을 개발하려고 시도해 왔는데 그 중 가장 주목목할 만한 것으로는 Norris(2005)와 Chapelle et al.(2010)의 연구에서 개발되되어 사용된 과제들이다. 이 두 평가도구는 모두 컴퓨터에 기반한 대학 수수준의 배치평가시험을 개발하는데 처리가능성 이론을 적용하였다. 이러러한 과제들은 좀 더 세분화된 형태를 취할 수도 있고 학생들에게 주어지지게 되는 문맥에 대한 좀 더 강한 통제를 가능하게 해줄 수 있다. 과제의

제의 이러한 측면은 처리가능성 이론에도 좀 더 적합하기도 하고 원어민 화자들이 많지 않은 상황에 일하고 있는 외국의 영어 교수자들에게도 접근성을 높여준다고 할 수 있다. 반면에, 평가 항목들이 평가의 취지에 부합하는 바를 측정하지 못하여 실제로 습득되지 않은 문법 항목을 완벽히 학습했다고 잘못 분류하는 긍정오류(false positive)를 유발할 수도 있는 단점이 있다. 또 다른 가능성은 과제의 문맥 자체가 특정 문법항목의 습득 여부를 보여주기에 불충분한 상황(예: 수험자가 과제의 소재에 생소한 경우)이 조성될 수 있다는 점이다. 이러한 경우에는 문법 항목이 실제로 습득되었음에도 불구하고 습득이 이루어지지 않았다고 잘 못 분류하는 부부정오류(false negative)를 범하게 된다.

본 연구에서 개발된 평가시험은 쓰기 과제들과 혼합형 과제들을 포함하고 있다. 쓰기 과제는 6개의 연결된 그림을 기반으로 하고 있고 혼합형 문법시험의 후반부에서 다루는 문법항목들과 동일한 문법항목들을 이끌어내도록 고안된 스토리텔링(story-telling) 과제이다. 본 시험에서 사용된 혼합형 문법시험은, 시험의 길이를 적정하게 유지하면서 동시에 여러 문맥적 요구사항을 시험에 반영하기 위해서 좀 더 적은 수의 문법항목들이 사용했다는 점을 제외하면, Chapelle et al. (2010) 시험에서 평가한 문법항목들과 거의 비슷한 문법항목들을 포함하고 있다. 본 시험은 한국의 중학생들을 대상으로 두 번의 예비 평가를 거쳤으며, 특히 최종 평가시험이 완성되기 전에 앞서 실시한 예비평가의 결과들에 기반해서 필요한 수정이 이루어졌다. 총 200여명의 한국인 중학생들이 본 연구에 참여했고 학생과 교사들이 시험 결과에 기반해서 진단적 피드백을 제공받았다. 두 과제 유형과 이 두 과업이 제공하는 진단적 정보 사이에 공통점과 차이점을 분석하기 위해서 양적 및 질적 분석이 이루어졌다. 분석 결과에 따르면 두 과업은 비슷한 결과를 보였고 처리가능성 이론에서 제시하고 있는 함축적인 위계(implicational scale)를 보여주었다. 하지만 혼합형 문법 시험의 경우에는 상대적으로 낮은 생산적 경향성을 보여주었다.

Keywords : 진단적 언어평가, 처리가능성 이론, 쓰기 평가, 생산적 기능, 생산적 과제, 배치 평가

Student Number: 2009-23815