Grammatical Complexity as a Predictor of Difficulty of Grammar Items in an English Test

Suthathip Thirakunkovit

suthathip.thi@mahidol.edu

Department of Applied Linguistics, Faculty of Liberal Arts, Mahidol University, Thailand

Seongha Rhee

srhee@hufs.ac.kr

Department of English Linguistics, Hankuk University of Foreign Studies, South Korea

---

Abstract

This study explores the extent to which the difficulty levels of grammar items in an English test can be predicted by the complexity of grammatical structures. The researchers carried out two sets of analyses. In the first analysis, the item facility and item discrimination indices of 175 multiple-choice items were examined. In the second analysis, the researchers asked a group of experienced English instructors to rate the difficulty level of 116 English grammatical features. The results from both analyses were compared with the developmental stages of grammar acquisition of Pienemann's Processability Theory (1998). The results revealed that some English grammar features were more difficult, and others less difficult, for ESL learners than the English instructors thought or believed. Moreover, the difficulty levels of grammatical structures derived from this study was not perfectly in accord with the developmental progression order proposed by Pienemann (1998). This result may suggest that Processability Theory might not adequately take account of the wide variety of grammatical structures in the standardized testing context. It is hoped that the findings of this study will be beneficial, not only to item writers and test developers, but also to English instructors, course syllabus designers, and developers of teaching materials.

Keywords: grammatical complexity, test item performance, Pienemann's Processability Theory

Introduction

Many language tests consist of grammar test items focusing on different grammatical structures of varying difficulty. For simple grammar items, finding the correct answer can be relatively easy for low-proficiency test takers, while difficult items require a higher level of grammatical knowledge. Therefore,

item difficulty or the percentage of students who answer an item correctly is a crucial parameter for creating a new set of test items for an English test because, in order to be fair to all test takers, all test forms must be similar in terms of the overall difficulty of the test items that they contain.

In classical test theory, the item difficulty is usually estimated mathematically by calculating the percentage of the test takers who answered the item correctly. Technically, this value should be calculated empirically by a pilot study before an actual test administration occurs. However, this could cause great concern about test security if such test items are to be used in a high-stakes test. Therefore, in order to make reliable estimates of item difficulty, the test writers must closely investigate the factors that might affect the difficulty of such items.

On a theoretical level, the findings of this study may provide new insights into one of the key factors that can distinguish between easy and difficult grammar questions: the complexity of grammatical structures. On a practical level, the results of this study may not only provide guidelines for item writers to more precisely generate grammar test items with the required levels of difficulty, but also offer pedagogical recommendations for English teachers who teach grammar structures.

Therefore, in this study, we have examined a variety of grammatical structures that are usually included in multiple-choice grammar tests in order to obtain a better understanding of the receptive order of grammar acquisition, and to establish the difficulty levels based on the test takers' test performance. This improved understanding may enable language testers to more effectively design reliable grammar tests with the desired levels of difficulty. We believe that language experts, by using this empirically validated order together with the experts' own holistic difficulty judgements, can write test items and develop tests that avoid relying on intuition or testing experience when choosing grammatical structures. Furthermore, we believe that English teachers will be able to plan more efficient curricula by selecting task types that can help learners with mastery of complex grammatical structures.

Theoretical Framework

In the field of second language acquisition (SLA), the receptive acquisition of L2 grammar knowledge is usually studied within the Processability Theory (PT) proposed by Manfred Pienemann in 1998. Even though the PT is mainly based on natural speech data, the researchers believe that this theory is still applicable to any grammar tests whose purpose is to measure communicative language competence.

In the original work, Pienemann (1998) proposed 6 stages of acquisition that could predict or explain the acquisition order of a variety of English grammatical structures:

1. No procedure (one single lemma, e.g., "yes" or "no");
2. Category procedure (lexical morpheme markings, e.g., past tense *-ed*, possessive *'s*, and plural *-s*);
3. Noun phrase procedure (number agreement between modifiers, e.g., "two persons", and nouns

corresponding to the phrasal procedure, e.g., “my house”);

4. Verb phrase procedure (adding an adverb into the verb phrase, e.g., “I ran quickly” and “Quickly, I ran”);
5. Sentence procedure (involving the unification between a verb phrase and the person’s number, e.g., “This girl loves cats”, and do-fronting, e.g., “Do you like cats?”); and
6. Subordinate clause procedure (involving sub-clausal word order, e.g., the use of inversion or the subjunctive in subordinate clauses).

Pienemann (1998) argued that every procedure is a prerequisite for the next procedure; for example, learners cannot assemble phrases without the knowledge of nouns, verbs, adverbs, adjectives (i.e., knowledge of English words or lemmas). Similarly, they cannot assemble words to make correct sentences without the knowledge of nouns and verb phrases and so forth.

Regarding empirical findings, there have been several studies conducted in the field of SLA within the Processability Framework to investigate whether the same or similar order could be observed. Even though several studies suggested a language development that could somewhat conform to the predictions of the PT (Baten, 2011; Buyl & Housen, 2015; Spinner, 2013; Yamaguchi & Kawaguchi, 2016), stages in the acquisition of a variety of grammatical structures were less clear or even unclear, and the learners in some of those studies were shown to skip some of PT stages, depending on their L1 backgrounds (Peker & Toprak-Celen, 2020). Therefore, we believe that more empirical evidence from different contexts of language use is absolutely necessary.

Even though several studies have investigated the accuracy of the Processability Theory when predicting data derived from naturally occurring language use and production, the literature still lacks extensive reviews of the relevance of PT in the context of standardized testing. Therefore, in order to fill in this gap in the literature, we would like to provide a starting point for the development of comprehensive guidelines for use in the selection and writing of grammar items in test design. We also include additional information about grammatical complexity that can be applicable to SLA scenarios.

Considering the nature of learning a second language, researchers need to emphasize the fact that linguistic competence is a multi-stage cognitive capability, which cannot be fully captured by simple, compartmentalized models, e.g., the Processability Hierarchy of Pienemann (1998). Instead, language competence might not develop in a linear fashion across the six stages. For instance, language has diverse domains, e.g., sound, lexicon, grammar, and discourse, and the constituents in each of these domains exhibit intricate interaction. In the simplest conceptualization of the bottom-up model, the knowledge pertaining to nouns (Stage 3 of the Hierarchy), by definition, is easier to acquire than that pertaining to verbs (Stage 4 of the Hierarchy). This conceptualization in fact masks the great complexity involved in language acquisition.

Given the challenges inherent in determining the levels of complexity in language, and, accordingly, the levels of difficulty for test-takers, it is imperative that test developers and item writers be

informed by much finer-grained approaches than as proposed in the idealized hierarchy. Therefore, this study is intended to contribute to the development of a more comprehensive and detailed test-frame of grammatical features, based on the item analyses of the tests that have been administered, the test-takers' performance, and expert judgments on a wide variety of grammatical structures.

The Review of Literature on Grammatical Complexity

In the field of linguistics, the term *Grammatical Complexity* has been used in different ways. For example, in psycholinguistics, it refers to processing complexity that can be measured by the amount of time required to understand a linguistic structure (Newmeyer & Preston, 2014). In SLA, it is defined as the complexity of grammatical structures that directly arise from the number of linguistic elements and their interrelationships within the sentence (Pallotti, 2015). In typological linguistics, the term is tied to the complexity of the phonemic and morphological structures of a language. Therefore, if a language (or variant) has more phonological or morphological distinctions, that language tends to be more difficult to acquire than languages that have fewer distinctions. Within the field of second language writing, Biber et al. (2011) stated in their study that learners would produce more advanced grammatical structures as they progressed in their language proficiencies. This is seen in the results of many studies that investigated writing development in relation to grammatical complexity (Grant & Ginther, 2000).

A number of studies have investigated several grammatical features in order to observe the existence of a natural order for the acquisition of English grammar because it is argued that different grammatical features have different levels of complexity and acquisition difficulty. However, many early studies in the field of SLA tended to focus on the order of morpheme acquisition. One of the seminal studies was conducted by Larsen-Freeman in 1975. She investigated 24 adults from four native-language backgrounds (Arabic, Japanese, Persian, and Spanish), acquiring 10 different morphemes, e.g., the acquisition of progressive (*-ing)*, progressive auxiliary, short plural, long plural (*-es*), third person singular, present tense singular, regular past, and irregular past, through different language tasks. She found that the orders across all tasks were remarkably similar to the proposed order across different L1 groups.

Apart from the investigation of morphemes, some researchers (e.g., Cancino et al., 1975; Wode, 1977) investigated the grammatical rules governing negatives. Both studies showed that early negative utterances usually contained a lexical particle "no" when the learners expressed denial, rejection, refusal, or correction, e.g., "No play here," and "No go." However, the full target rule of negation, e.g., "She doesn't want it.", would be acquired later after the learners were able to inflect for tense and number.

Other studies have investigated the difficulty levels of English structures, but most of these focused on proposing various hypotheses rather than investigating the acquisition order *per se*. One of these studies was conducted by Ioup in 1983. She examined L2 learners' acquisition order of various types of subordinate

clauses—e.g., adverbial clauses, infinitive complements, gerundive complements, and participle phrases—and three part-of-speech categories, which are adjective, adverb, and preposition. The participants were 166 adult speakers of Arabic in Egypt. In this study, a sentence completion format was used. The results led Ioup to a set of hypotheses relating to relative difficulty or ease across and within parts of speech. These hypotheses, however, could not be ordered in a single, consolidated scale since the difficulty levels could not be quantified in a reliable way.

In a more recent study, Biber et al. (2011) have taken a corpus-based and empirically-grounded approach to investigate the order of grammar acquisition. By examining 28 grammatical structures that could best capture the complexity of L2 learners' production, they have proposed a set of five developmental stages for grammatical complexity features. Biber et al. (2011) explained that, generally, L2 learners progress from finite complement clauses controlled by common verbs, e.g., "I know how to do it", and *that*-complement clauses, e.g., "I believe that you should go", through intermediate stages of non-finite dependent clauses, e.g., "I love watching TV with my family", to the final stage requiring dense use of different kinds of phrasal (non-clausal) dependent structures that function as constituents in noun phrases and clauses, e.g., appositive, prepositional phrases as post-modifiers, participial phrases, and relative clauses.

Even though these two hypotheses—Ioup (1983) and Biber et al. (2011)—seem to be complementary to one another, they should be viewed as tentative and derived from empirical studies of L2 language development. Another empirical study, conducted by Parkinson and Musgrave (2014), specifically investigated the development of noun phrase complexity in the academic writing of different groups of graduate L2 writers. They found that the more proficient learners used significantly higher numbers of noun modifiers, while less proficient learners relied heavily on attributive adjectives, a modifier hypothesised as being acquired early, e.g., *that*-complement clauses. The findings of this study could confirm Biber et al.'s developmental stages in the sense that the less proficient group also relied heavily on attributive adjectives.

Motivated by the order of acquisition of levels of grammatical complexity shown in previous studies, we would like to explore the extent to which the levels of item difficulty of different grammatical structures obtained from a grammar test are consistent with those hypotheses mentioned in this section.

Research Objectives

The objectives of this study are: a) to examine the extent the level of test difficulty of a grammatical test item can be predicted by the characteristics of English structures defined as easy or difficult to acquire for second language learners, and b) to examine the difficulty levels of multiple-choice items with the order proposed by the PT.

These objectives lead to the following research questions.

1) To what extent can different grammatical structures predict the difficulty of grammatical items?
2) To what extent can the levels of difficulty of different grammar structures be aligned with the order predicted by a Processability Theory (PT) framework (Pienemann, 1998)?
3) To what extent does the difficulty level revealed by test results align with the judgments of language experts?

Ultimately, we hope that the results of this study will contribute to the fields of SLA, grammar test development, and grammar teaching.

Research Methodology

Overview of the English Test

The English test used for the analysis in this study is an exit test developed at a large public university in Thailand. It is an English proficiency test designed to evaluate the general English language proficiency of the test takers in educational, business/workplace, public, and personal contexts. The test content reflects everyday interactions in an English-speaking environment. The test is aimed at the reading and listening skills at the B2 (Independent user) level of the Common European Framework of Reference (CEFR). From the beginning of the 2017 academic year, university regulations require all new undergraduate students to have a score of 84 or higher (out of 150) to fulfill the English language requirement prior to their graduation. Therefore, the test can be considered a high-stakes test because the test results are used to determine an important outcome, i.e., whether a student will be able to graduate or not.

The Test Format

The test is a paper-and-pencil test that contains 150 multiple-choice questions in two major sections: Listening and Reading. Grammar and vocabulary are assessed within the reading section.

The listening section takes approximately 35 minutes to complete. This section consists of 75 questions assessing the ability to understand conversations and talks in a variety of contexts. This section is divided into four parts: question-response, short conversations, long conversations, and talks. Upon completion of the first section, the test takers are given 45 minutes to complete the reading section. This second section also consists of 75 questions assessing the ability to understand the texts in a variety of contexts, and is divided into four parts: grammar, vocabulary, short reading passages, and long reading passages.

The grammar section, which is the main focus of this study, consists of 25 incomplete sentences. Each incomplete sentence is followed by a choice of four words or phrases which may be used to complete the sentence. Only one choice can make the sentence grammatically correct. Each statement ranges between eight and 25 words, including the answer. The grammar points cover a wide range of grammatical features. The test takers are on the verge of graduation; therefore, all of the grammar items on the test are aimed at learners from low-intermediate to high-intermediate levels, and a passing test score is intended to indicate a good mastery of English grammar. The content of each test item reflects language which is used naturally in different settings, e.g., universities, offices, social gatherings, and personal conversations.

Research Data and Analytical Procedures

The English structures of all grammar test items were initially analyzed descriptively, based largely on the grammar classification schemes of Quirk et al. (1985) with some modifications (See Appendix 1). The researchers worked independently when coding the data, thereafter, cross-checking three times to ensure maximum accuracy and agreement. Any coding disagreements were resolved during discussion.

Next, we analyzed the test performance of approximately 8,500 undergraduate students on the grammar section of the seven test forms of the test. Each test set comprises 25 four-choice grammar items. The seven test forms were administered between the academic years of 2017 and 2020. The internal consistency values of all test forms could be comparable because they ranged between 0.92-0.95. A total of 175 items were used in the item analysis process. Each correct answer was given one point and each wrong answer received zero. All test takers were Thai undergraduate students from various majors. They all had studied English as a foreign language for at least six years in junior and senior high schools. Then, the test responses obtained from the 8,500 test takers were analyzed for estimates of test item difficulty using the item facility index[1], and for item discrimination[2] using the point biserial correlation coefficient (r_{pb}). All the test items were then ranked according to their mean difficulty. Even though our data were based on a large dataset, the results of this study should be interpreted with caution because they came from only one grammatical test developed locally.

In common practice, a mean score of 0.5 represents the average difficulty of all items on the test. Therefore, if the mean difficulty of a particular item is 0.75 or above, the item is considered very easy; if the index is between 0.25 and 0.74, it is considered moderate, and if the index is less than 0.25, it is considered very difficult. However, as the item difficulty for the moderate items has a relatively wide range,

[1] Item facility index (IF) refers to the percentage of a group of test takers that answer a test item correctly.
[2] Item discrimination index (ID) refers to a measure that identifies how well a test item can discriminate between test takers who perform well and those who perform poorly.

it was further classified into two categories—moderately easy and moderately difficult—in order to have a clearer picture of the item difficulty.

Table 1

Evaluation of Item Difficulty (IF)

Item Facility Index	Item Evaluation
0.75 or above	Very easy item
Between 0.50 and 0.74	Moderately easy item
Between 0.25 and 0.49	Moderately difficult item
0.24 or below	Very difficult item

Note. Adapted from *Introduction to Measurement Theory* by M. Allen and W. Yen, 1979, Brooks/Cole.

The item discrimination index indicates the ability of an item to differentiate test takers on the basis of how well they do the test. In common practice, if the value of the discrimination index is above 0.40, it is considered excellent. If it is between 0.30 and 0.39, it is considered fair or reasonably good. If it is between 0.10 and 0.29, it should be subject to improvement. If it is below 0.09 or has a negative value, it is considered poor and should be either rejected or improved by further revision.

Table 2

Evaluation of Item Discrimination (ID)

Discrimination Index	Item Evaluation
Above 0.40	Excellent
Between 0.30 and 0.39	Fair or reasonably good
Between 0.10 and 0.29	Marginal items, usually needing and subject to improvement
Below 0.09	Poor items, usually to be rejected or improved by revision

Note. Adapted from the relationship of the reliability of multiple-choice test to the distribution of item Difficulties by F. M. Lord, 1952, *Psychometrika, 18*, 181-194.

Moreover, in an attempt to enhance the robustness of item difficulty estimation, which is believed to be one of the crucial factors ensuring the equivalency of high-stakes tests, we asked 21 university instructors of English to rate the difficulty of 116 grammar items based on a scale of 1-10 (1 = *very easy*, 10 = *very difficult*). All instructors selected to participate in this study had extensive experience in both teaching grammar and writing grammar test items. Before they began rating, they were all instructed to briefly scan through all the questions. The ratings obtained from the language experts were statistically analyzed in terms of mean, mode, and standard deviation. For the ratings received from 21 university instructors, we classified the mean value between 1 and 2.5 as "easy"; "moderately easy" if it was between 2.6 and 5.0; "moderately difficult" if it was between 5.1 and 7.5; "difficult" if it was between 7.6 and 10.

Results

Item Difficulty of English Grammar Features

From the data gathered from the test takers' test responses, it was found that only 4% of the test items were considered easy, 33% were considered moderately easy, 52% were considered moderately difficult, and only 11% were considered difficult. Below are the structures, divided into four categories based on their item facility indices. The results are presented in the order of IF values.

Table 3

Mean Difficulty Estimates of Very Easy Items (IF = 0.75 - 1.00)

Grammar Area	Example	Estimates of IF	Estimates of ID
Simple subordinator	John loves to travel around Bangkok <u>because</u> he loves street food.	0.86 - 0.95	0.28 - 0.4
Comparative adjective	Difficult experiences teach you to be a <u>stronger</u> person.	0.84 - 0. 89	0.36 - 0.37
There-construction	The busiest time for stores is before Christmas because <u>there are</u> shoppers everywhere.	0.8	0.34
Adverb modification	This report is solely prepared by the secretary.	0.8	0.49
Reflexive pronoun	The customers have to go to the bank <u>themselves</u> to open a bank account.	0.77	0.51

Please note that this table should be read with care. The indices in Columns 3 and 4 indicate the ranges of item facility and discrimination values of two or more items, so the difficulty levels of some items may overlap. As seen from Table 3, simple subordinator (particularly, the use of causative and time conjunctions), comparative adjective, adverb modification, and *there*-structure were easier than other grammatical structures in the test. These features were quite easy for our participants. One explanation is that their structures are based on a limited set of rules that students can easily memorize and understand. However, we have observed that many simple coordinator and subordinator items can have their levels of difficulty increased by increasing the level of vocabulary difficulty. An example is given below.

Thomas has been afraid of planes ……. his father died in a plane accident when he was young. (IF = 0.86; ID = 0.4)

*(a) since (b) while (c) during (d) due to

Hundreds of private nursery schools around the country have closed down ……. the government began offering free childcare to parents. (IF = 0.41; ID = 0.24)

(a) so *(b) since (c) during (d) although

Another possible explanation is the existence of the features in the first language. Many of these structures, e.g., *there*-construction, simple subordinators, and comparative adjectives are quite similar in the test takers' mother tongue, as seen from the example below.

Our hotel will make you feel as ……. as you are at home. (IF = 0.71; ID = 0.35)

(a) comfort (b) comforting *(c) comfortable (d) comfortably

(Thai) โรงแรม ของเรา จะ ทำให้ คุณ รู้สึก สบาย เหมือน อยู่ บ้าน

hotel our will make you feel comfortable like at home

Table 4

Mean Difficulty Estimates of Moderately Easy Items (IF = 0.50 - 0.74)

Grammar Area	Example	Estimates of IF	Estimates of ID
If-clause	The class would be more interesting if the professor incorporated some media into his lecture.	0.56 - 0.65	0.38 - 0.52
Determiners and quantifier	May I talk to you just for a few minutes?	0.51 - 0.56	0.37 - 0.4
Gerund	Having no time for a family can destroy a marriage.	0.51 - 0.68	0.47 - 0.6
Verb complement	Teachers usually do not let the students read scripts during their presentations.	0.51 - 0.73	0.06 - 0.65
Simple and progressive tenses	Jeffrey has been working in the sales department since he came back from sabbatical leave.	0.51 - 0.73	0.29 - 0.51
Relativizer	Foie Gras, which is one of the most popular elements of French cuisine, is seen as a luxury food.	0.5 - 0.69	0.29 - 0. 55
Subject-verb agreement	A wireless mouse, along with headphones, is provided for everyone in the meeting.	0.5 - 0.74	0.26 - 0.45

Based on the obtained results, it was revealed that even though some similarities and overlaps in the difficulty levels and item discrimination levels could be seen among relativizer, subject-verb agreement, tenses (simple and progressive), determiners and quantifiers, gerund, *if*-clause, and verb complementation, they are moderately easy items. This might be because the concepts of these structures are relatively easy to understand, and the semantic relations between clauses are simple and straightforward.

Table 5

Mean Difficulty Estimates of Moderately Difficult Items (IF = 0.25 - 0.49)

Grammar Area	Example	Estimates of IF	Estimates of ID
Preposition complement	I am sure that you will get accustomed to walking with these crutches very soon.	0.31 - 0.41	0.34 - 0.48
Adjective complement	Jane is responsible for organizing all events in this department.	0.27 - 0.47	0.22 -0.46
Mood (subjective)	It is necessary that Mary see the physician as soon as possible.	0.26 - 0.45	0.21 - 0.51
Pre/post noun modifications	Many investors are hesitating to invest in our country due to the fluctuating market.	0.25 - 0.47	-0.17 - 0.55
Relative clauses	The manufacturer offers replacements only for that which is under warranty.	0.25 - 0.46	0.09 - 0.48
Preposition	Due to a large backlog of orders, we are unable to process your shipment.	0.25 - 0.44	0.34 - 0.67
Subordination (Participle)	Aiming for the university to become a world-class university, the president has created several international projects.	0.36 - 0.41	0.18 - 0.43
Passive	In this exam room, all students are required to present their student ID cards.	0.3 - 0.47	0.16 - 0.53
Verb complementation	One practical way to prevent dehydration is to drink 1.5 liters of water a day.	0.26 - 0.47	0.17 - 0.58
Tenses	It is predicted that many graduates will be working with small companies rather than big ones.	0.29 - 0.48	0.07 - 0.39

Many areas of complements (e.g., verbs, adjectives, prepositions, infinitives) and pre/post modifications were found to be relatively difficult for the majority of our participants. Moreover, passives, relative clauses, and mood subjunctive were shown to be at the same level of difficulty with similar mean scores. The difficulty may be due to the complexity of the rules. For example, counterfactual conditionals

(mood subjunctive) and relative clauses could be difficult to interiorize by a number of L2 learners, even by relatively proficient ones, because the patterns typically involve complex constructions, and they might not be frequently used in everyday speech. Moreover, the absence of some of these features in the test takers' mother tongue, e.g., tense agreement, active-passive distinction, and participles, may add to the level of difficulty.

When investigating the items that test the use of preposition complements closely, it is interesting to note that the use of the same prepositions could have different levels of difficulty, depending on the level of vocabulary difficulty used in the sentence and internal structural complexity involving multiple subordinate clauses. A comparison between two test items is given below.

The price of gasoline went up last year ……. the political issues in the country.
(IF = 0.44; ID = 0.39)
(a) despite *(b) due to (c) besides (d) because

Economists anticipated that the economy would be at risk because residents who left the area ……. the hurricane last month could possibly never return. (IF = 0.25; ID = 0.35)
*(a) due to (b) in spite of (c) caused by (d) result from

Table 6

Mean Difficulty Estimates of Very Difficulty Items (IF = 0 - 0.24)

Grammar Area	Example	Estimates of IF	Estimates of ID
Pair conjunctions	There was so much smoke that we could not see anything across the hallway.	0.08 - 0.16	0.18 - 0.19
Noun nominalization	The murder of the man at the gas station was very tragic.	0.09 - 0.39	-0.02 - 0.34
Pre/post noun modifications	Our university focuses on forming graduates equipped with knowledge and morals.	0.16 - 0.25	0.1 - 0.28
Mood subjunctive	It's important that he do exercise every day. I wish you were here yesterday.	0.19 - 0.26	0.12 - 0.37
Passive	All of the luxury cars in this showroom were imported from Europe.	0.22 - 0.37	0.17 - 0.53
Relativizer (Possessive)	Customers whose deliveries do not arrive on time should contact our customer services immediately.	0.21 - 0.41	0.46 - 0.09
Complementation (participle)	All students will be kept informed of any new changes.	0.14 - 0.41	0.37 - 0.44
Aspect tenses	I do not think that the secretary will not have finished my report by tomorrow.	0.21 - 0.48	0.16 - 0.41

Table 6 has shown that pair conjunctions is the most difficult feature on the test. Noun nominalization was shown to be the second most difficult feature for our participants, followed by pre/post modifications, mood subjunctive, passive, relativizer (possessive), participial complementation, and aspect tenses.

Mood subjunctives seem to be a difficult feature for a number of test takers, as reflected in item facility indices as low as 0.19. As seen from the example below, subjunctive items can be complex and very difficult for our participants.

It is advisable that university ……. a great place to get an education. (IF = 0.2; ID = 0.12)

(a) is *(b) be (c) been (d) being

The doctor recommended that he ……. cold showers and avoid processed sugar for a month. (IF = 0.26; ID = 0.37)

(a) took *(b) take (c) takes (d) taking

One explanation may be that the correct usage of this kind of subjunctive is uncommon in informal speech, and it is even rarely seen in modern English writing. Therefore, some students might have a hard time correctly using this particular usage.

It is interesting to note that a number of features, e.g., passive voice and participial construction items, can become more difficult if more than one single structure is measured within the same test item. A comparison between two test items is given below. In the second test item, the test takers are required to know the structure of the imperative sentence in the passive voice form.

In this exam room, all students ……. to present their student ID cards to the invigilators. (IF = 0.37; ID = 0.53)

(a) require (b) are requiring

*(c) are required (d) will be requiring

Please …… that the next meeting will be held this Thursday at 2 pm. (IF = 0.21; ID = 0.28)

(a) inform *(b) be informed

(c) have been informed (d) be informing

Moreover, we found that pronoun relativizer items (pronouns used in relative clauses) were spread out along the continuum in terms of difficulty. The items in which relativizers or relative pronouns are in the subject positions seemed to be easier than the ones in which the relativizers were in the object position, e.g., ‘whose,’ ‘whom’, and ‘where’, or in which the relativizers were omitted.

Foie Gras, ……. is one of the most popular and well-known elements of French cuisine, has been a controversial issue recently. (IF = 0.51; ID = 0.55)

(a) who (b) whose *(c) which (d) where

The company ……. John is working offers really good salaries. (IF = 0.35; ID = 0.15)

(a) which (b) when (c) whose *(d) where

Another feature whose difficulty levels vary is tenses. When comparing different tenses in English, tenses that are commonly used in everyday conversations (e.g., simple present and past, and present and past continuous tenses) are quite easy, while complex tenses, e.g., aspects, and combinations of different tenses, e.g., progressive aspects, are relatively more difficult for our participants. One possible explanation is that complex tenses might not be clear for L2 learners. These results, at the same time, are in line with Bardovi-Harlig's (2000) finding that aspect tenses are complex features for L2 learners.

Instructors' Perceptions of the Item Difficulty of English Grammatical Features

The level of item difficulty of English grammar features obtained from the test was cross-checked with English instructors' perceptions in order to investigate the extent English instructors perceive grammar difficulty. The data from both sets were compared to find possible similarities and differences.

Table 7

Instructors' perceptions and test takers' performance of relatively easy items

Instructors' perceptions	**Results from test takers' performance**
Simple subordinator	Simple subordinator
Comparative adjective	Comparative adjective
If-clause	*If*-clause
Simple and progressive tenses	Simple and progressive tenses
Determiner and quantifier	Determiner and quantifier
Subject-verb agreement	Subject-verb agreement
Definite and indefinite pronouns	Reflexive pronoun
Adverb modification	Adverb modification
Active voice	*There*-construction
Auxiliary verbs	Gerund
Verb infinitive	Verb complement
Modal verbs	
Aspect tense	
Noun (countable/uncountable, singular/plural)	
Preposition	
Verb, adjective, and preposition complement	

Table 8

Instructors' perceptions and test takers' performance of relatively difficult items

Instructors' perceptions	Results from test takers' performance
Mood (subjunctive)	Mood (subjective)
Subordination (adverbial clause, participial)	Subordination (participial)
Relative clause	Relative clause
Noun, preposition, verb, adjective complements	Noun, preposition, verb, adjective complements
Nominalization	Nominalization
Passive	Passive
Special construction, e.g. fronting and inversion	Pair conjunctions
Pre/post modification	Pre/post modification
	Preposition
	Relativizer (Possessive)
	Aspect tenses

Overall, the rating results indicate that most of our language experts perceived most of the English structures as moderately easy and moderately difficult (with the average = 5.33). Notwithstanding, the ranking results of the experts' rating seem to suggest that their perceptions of grammatical difficulty are based on whether the rules to describe the formation of language features are easy or difficult to understand, and based on the frequency with which those structures can be encountered by the learners on a daily basis. In particular, we have found some areas, e.g., mood subjunctive, passive, nominalization, pre/post modifications, some types of relativizers, and participle complements, were on top of both lists of difficult items. At the same time, some other areas, e.g., simple tenses, coordinating and simple subordinators, determiners, comparative adjectives and adverb forms, were perceived as comparatively easy. Certain structures, e.g., the use of passive, aspect tenses, and prepositions, were clearly perceived to be more difficult by Thai test takers than by the instructors of English. For the remaining structures that are not mentioned specifically, the similarities or differences cannot be clearly identified.

Further Investigation of the PT Based on the Results

The present study set out to investigate the extent PT, which has been established for understanding L2 grammar acquisition, can predict the difficulty level of different English grammatical structures. The data obtained from PT, the test takers' performance, and the perceptions of English instructors were

converted in order to make data from different sources comparable on a scale of 1-10. The following adjustments were made.

The rank based on the Processability Theory

All values from PT were averaged (e.g., Level 5 and Level 6 = 5.5), and then the obtained values were converted to the 1-10 scale by using the formula *(1+((a-1)*1.8))*. Therefore, the PT level of 5.5 is equal to 9.1 on the scale of 1-10.

The difficulty values obtained from the English instructors' perception

The min and max values obtained from the English instructors' perception (min =1; max = 8) were converted to the 1-10 scale by using the formula *1+(a-1)*1.285*. Therefore, the level of difficulty of the English instructors' perception of 8 is equal to 10.

The IF values obtained from the test takers' performance

In order to convert test takers' IF values between 0.0 and 1.0 to a 0.0-10.0 scale, the IF values were multiplied by 10 (*a*10)*. For example, an IF value of 0.12 is equal to 1.2. Then, in order to express values in terms of difficulty, the values for ease were inversed using the formula *10-a*. For example, the value of 1.2 became 8.8. Finally, in order to convert from the 0-10 scale to the 1-10 scale, the values were modified using the formula *(a*0.9)+1*. Therefore, the value of 8.8 became 8.92. This value indicates a very difficult test item.

Category 1: PT and English instructors' perception

Comparison between the PT scale and the scale of the English instructors' perception shows that the prediction of PT is not well-aligned with the perception of difficulty of the English instructors. For example:

The chairperson suggested that it is better for the team to talk to its staff individually than ……. the issue in the next meeting.

(a) discuss (b) discusses *(c) to discuss (d) will discuss

The test item involves *to*-phrase complementation and clausal parallelism, which PT would place at Level 4 (verb phrase procedure), Level 5 (sentence procedure), and Level 6 (subordinate clause procedure), which altogether can translate into the modified value of level 8.2. The assessors, on the other hand, place this test item at a very low level of difficulty, i.e., at the average level of 2.3, departing from

PT prediction level by -5.9. The English instructors' judgment is largely based on the assumption that test-takers would apply a simple grammatical structure of [A (rather) than B], in which A and B are structurally comparable or parallel. Thus, the prediction is not a step-by-step progression of morphosyntactic complexity, which is the basis of PT. Thus, for the English instructors, this test item is merely at the level of idiomatic structures bordering on the lexical level. Incidentally, test-takers' performance, at the inversed IF 6.31 (IF 3.69), is not consonant with the distance either from PT at +1.89 and from the English instructors at -4.01. In other words, PT's prediction of the difficulty level is too high, while the English instructors' prediction is too low.

A similar phenomenon can be seen from the following item. PT projects it at the high-difficulty level or late stage acquisition, but the English instructors perceived it as an easy item.

His marketing plan includes strategies to raise the visibility of the products, as well as direct distribution ……. the target customers.

(a) of *(b) to (c) at (d) as

The sentence structure of this test item involves noun phrases and post-modifying prepositional phrases. For this item, PT would place it at Level 3 (noun phrase procedure) and Level 5 (sentence procedure), which can together translate into the modified value of level 8.2. The English instructors, on the other hand, put the item at a very low level of difficulty, i.e., at an average level of 2.3, departing from PT prediction level by -5.9. The English instructors' reasoning is that the noun phrase is a nominalization of the phrasal verb "distribute x to y" into "distribution to y", and thus the syntactic paraphrase would be very easy. The test-taker performance indicates a moderate level of difficulty at the inversed IF value of 4.96, again departing from the predictions of PT as well as of the English instructors.

On the other hand, the opposite polarity can be observed in the following example, where PT predicts it to be at the low-difficulty level or easy to acquire, while the English instructors expect it to be a difficult item.

We are terribly sorry to say that the items ……. are temporarily out of stock.

*(a) ordered (b) ordering (c) order (d) to order

The test item involves post modification of a noun phrase by means of a participle, which PT would place at Level 3 (noun phrase procedure), which translates into the modified value of 4.6. The English instructors, on the other hand, put it at a considerably higher level of difficulty, at an average of 8.7, departing from PT by +4.1. The English instructors' reasoning is that this type of participial post

modification is the product of ellipsis of a full-fledged relative clause, and, therefore, that the test-takers are required to have a high level of understanding of syntactic operations, e.g., relative clause complementation, relative clause, and ellipsis—a highly theoretically motivated analysis. However, the test-takers' performance indicates that the question was only of moderate difficulty with the inversed IF of 4.96, thus rather close to the prediction by PT (with +0.36). This strongly suggests that the test takers may be simply using the technique of differentiating noun phrase post modification types among *to*-infinitive, present participle, and past participle, largely prompted by the cues provided by the answer choices.

Category 2: The English instructors' perception and test takers' performance

The next observation involves the English instructors' perception and the test takers' performance. The English instructors think the following example is a difficult item, while the test takers' performance indicates that it is an easy item:

Some people believe that there ……. always ample opportunities if we try to be more optimistic.

*(a) are (b) is (c) be (d) have been

The question involves a special construction (*there*-construction) and the present tense. The English instructors predicted that this would be very difficult for the test takers (with a value of 8.7), just by virtue of involving a special construction. The English instructors, all of whom are grammar instructors, believe that special constructions always need special attention and instruction because they involve apparent grammatical anomaly in one way or another. In the case of the *there*-construction, English learners need to understand that it involves interpolated word order, insertion of the dummy subject *there*, number-agreement with the semantic subject contra the structural subject, and definiteness constraint, among others. The test-takers, however, found this question very easy (with the inversed IF of 2.8). The test-takers seem to have solved the problem only by way of identifying it as a *there*-construction and the number agreement, which makes the solution very easy. Once again, the test-takers may not go through all syntactically-motivated analytic processes. Incidentally, the PT prediction of Level 5 for this item is also widely deviant from the test-takers' performance, with a distance of +5.4.

On the other hand, the following example shows an observation in the opposite direction. The English instructors predicted the structure to be an easy item, while the test takers' performance indicates a high level of difficulty.

Laura always ……. museums when she spent her summer in her grandmother's hometown.

(a) visit (b) visits (c) visiting *(d) visited

The question involves the past tense of a sentence, seemingly a very easy task, and the English instructors expected that it would be at the lowest difficulty level (1). The performance result of the test takers, in contrast, shows that it is a very difficult item (7.39), creating the widest gap in the dataset (-6.39). A closer look at the question reveals the presence of the distractor, i.e., the adverb of frequency *always*, which would have tempted the test-takers to opt for the present tense or the present participial form as a part of the present progressive. The fact that the cue for the temporal reference is the subordinate clause verb in the suppletive form *spent* must have played a role in further making the selection less straightforward. Incidentally, the prediction by PT also exhibits a moderate level of difference from the test-takers' performance, i.e., +1.8.

Category 3: PT and test takers' performance

Another discrepancy can be seen in the comparison between PT and the test takers' performance. In this first example, PT predicts that this item should be placed at a high level of acquisition; however, the test takers' performance shows the opposite.

Martin often uses his days off to travel in downtown Bangkok ……. he loves street food.

(a) despite *(b) because (c) unless (d) while

The test item involves a subordinate clause with a simple subordinator, which PT would unequivocally place at the level of subordinate clause procedure, the highest level of processability difficulty (i.e., 10). The test-takers, however, found the question to be at a considerably lower level of difficulty, i.e., an inversed IF of 1.45, thus creating the longest distance between PT prediction and the test takers' performance, with a distance index of 8.55. The ease on the part of the test-takers seems to be attributable to the fact that the question is straightforwardly making reference to the semantic function, i.e., the relationship between the two clauses, and, further, that the causal connection is also quite plainly visible. Attribution of causal relation to events is among the fundamental reasoning processes of humans, and occurs frequently, activated by the juxtaposition of even remotely related events. Thus, the test item is reduced to the identification of the conjunction meanings, similar to the lexical level question. This strongly suggests that for second language learners the difficulty level is more strongly tied to the differential ease

or difficulty at the cognitive and perceptual levels and that the mechanical calculation of the syntactic complexity, as found in PT, does not provide an adequate prediction of the empirical difficulty a language learner experiences.

Differences in the opposite direction can be observed in the following example, which PT placed at the lower end of the continuum, whereas the test takers found it difficult.

Of the two textbooks selected, the …… is better in many aspects.

(a) later *(b) latter (c) latest (d) last

The test item involves a lexical form and pro-form substitution for a nominal category, which PT would place at Level 1 (no procedure), Level 2 (category procedure), and Level 3 (noun procedure), which altogether translate to the composite index of 2.8, a very low level of difficulty. The test-takers, however, found this item very difficult (i.e., inversed IF of 8.38), thus creating a wide divergence from the PT index, with a difference of -5.58. This suggests that even though the test question may seem to belong to the nominal category level or lower, as is predicted by PT, the question, in fact, requires far more extensive knowledge than the learning of a lexical item, i.e., knowledge about the substitutive pro-forms, selection of demonstratives, and specification of deictic contrasts, among others.

Discussion and Conclusion

Even though PT has grown exponentially in the field of language acquisition for more than two decades, the current study revealed that PT seems to be inadequate to account for the maturation stages of grammatical competence in SLA. Taking a closer look at the results of this study, we have found two major flaws in the PT framework.

First, PT is too simplistic or coarse-grained, simply classifying language development into six broad stages: 1) lemma, 2) category procedure, 3) noun phrase procedure, 4) verb phrase procedure, 5) S-procedure, and 6) subordinate clause procedure. As our data shows, there are a number of discrepancies between the stages of language development proposed by PT and the language performance of the L2 test takers. For example, some rules of subordinate clauses are shown to be easy and acquired relatively early in the development process because their logical relationships are straightforward and easy to memorize, while other kinds of subordinate clauses, e.g., relative clauses, can be quite difficult for a number of L2 learners. Moreover, not in accord with the order predicted by processability theory, our data show that noun phrases and different kinds of complements are among the most difficult English structures for Thai learners. Therefore, the development stages proposed by Pienemann might not correctly predict what features are more or less difficult for L2 learners.

Second, PT is too static to be able to capture intra- and inter- stages of language development. PT simply pre-arranges the grammatical features under the investigation of natural speech data, instead of arranging them based on their order of acquisition by L2 learners. When a student has a strong foundation in their first language, the students' L1 can be used as a bridge to facilitate L2 learning. Even though PT is argued to universally predict which grammatical structures an L2 learner can process at a given level of development, it fails to describe the diverse range of interlanguage developments of L2 learners. Therefore, the results of this study suggest that it is possible that the test takers might have used both their implicit and explicit knowledge when choosing the correct answer.

To put it simply, we believe that PT might not be able to offer the basis for detailed understanding of the complex grammatical development of L2 learners' interlanguage. The difficulty level of a number of grammatical structures derived from PT is not in accord with the difficulty order derived from the test takers' performance. Statistical analysis reveals that the correlation between PT and the test takers' performance is extremely low (Pearson's correlation coefficient (r) = 0.22), which suggests that there is no significant relationship between them regarding the ordering of grammar items. This means that not all grammatical features of a given stage need to have emerged in the L2 learners' interlanguage before they move onto the next stages of L2 learning.

Moreover, when looking at the comparison between the test takers' performance and English instructors' perception of difficulty, our findings show that the test takers seemed to look at the items differently than instructors, and what they notice on the test and how they interpret what they notice strongly influences item difficulty. The correlation between the test takers' performance and the English instructors' perception is also low (r = 0.32). English instructors might be more analytic by relying on their linguistic knowledge; thus, they ignore more practical problem-solving techniques used by L2 learners. Therefore, our final note on the implications of this study is that grammar instructors or test writers need to be familiar with analyses of test takers' performance. That is, they need to be informed with the performance patterns of their students.

Finally, we would like to say that it can be a real challenge to write items that produce similar difficulty estimates based on our linguistic knowledge. Our intuition and experience might not be the best guide for writing good grammar test items. Therefore, we strongly recommend that test developers and grammar teachers pilot their test items to get statistical information about item functioning, and conduct interviews or a questionnaire to investigate what determines the relative difficulty of their test items.

References

Allen, M., & Yen, W. (1979). *Introduction to Measurement Theory*. Brooks/Cole.

Bardovi-Harlig, K. (2000). *Tense and Aspect in Second Language Acquisition: Form, Meaning and Use*. Blackwell.

Baten, K. (2011). Processability Theory and German case acquisition. *Language Learning*, *61*(2), 455-505.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5-35.

Buyl, A., & Housen, A. (2015). Developmental stages in receptive grammar acquisition: A Processability Theory account. *Second Language Research, 31*(4), 523-550.

Cancino, H., Rosansky, E. J., & Schumann, J. H. (1975). The acquisition of the English auxiliary by native Spanish speakers. *TESOL Quarterly, 9*(4), 421-430.

Freeman, D. E. L. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly, 9*(4), 409-419.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123-145.

Ioup, G. (1983). Acquiring complex sentences in English. In K. Bailey, M. Long, & S. Peck (Eds.), *Second Language Acquisition Studies* (pp. 41-55). Newbury House.

Lord, F. M. (1952). The relationship of the reliability of multiple-choice test to the distribution of item difficulties, *Psychometrika, 18*, 181-194.

Newmeyer, F. J., & Preston, L. B. (Eds.). (2014). *Measuring Grammatical Complexity*. Oxford University Press.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117-134.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for academic purposes students. *Journal of English for Academic Purposes*, *14*, 48-59.

Peker, H., & Toprak-Celen, E. (2020). A critical review on the components of Processability Theory: Identifying the limitations. *ELT Research Journal*, *9*(1), 71-89.

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

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985), *A Comprehensive Grammar of the English Language*. Longman.

Spinner, P. (2013). Language production and reception: A Processability Theory study. *Language Learning, 63*(4), 704-739.

Wode, H. (1977). *On the systematicity of L1 transfer in L2 acquisition*. ERIC: ED176585

Yamaguchi, Y., & Kawaguchi, S. (2016). Development of relative clause constructions in English L2. *International Journal of Applied Linguistics and English Literature*, *5*(1), 83-93.

Appendix A

Classification summary based on Quirk, Greenbuam, Leech and Svartvik (1985)

Verb	Main Verb
Auxiliary	Primary Verb, Modal, Subjunctive Mood, Active/Passive Voice
Verb Phrase	Tense, Aspect
Noun	Noun class, Proper Noun, Number, Gender, Case
Determiner	Predeterminer, Postdeterminer, Quantifier, Article
Pronoun	Gender, Case, Reflexive, Interrogative, Demonstrative
Numeral Adjective	Attributive/Predicative, Postpositive, Supplementive, Contingent, Exclamatory, Comparative, Superlative
Adverb	Comparative, Superlative
Preposition	Simple, Complex
Sentence	Concord, Coordination, Quasi-coordination, Negation
Sentence-Type	Question, Directive, Exclamative, Echo, Irregular sentence, Formulae, Interjection
Pro-Form	Substitution
Ellipsis	Ellipsis, Reduction
Coordination	Coordinator, Correlative
Subordination	Subordinator, Direct/Indirect Speech, Nominal clause, Adverbial clause, Sentential relative clause
Verb Complementation	Multi-word verbs, Infinitive/Participle/Wh-/That... complement
Adjective Complementation	PP-/That-/Wh-/Than-/To-... complement
Noun Phrase	Postmodification (Relative/Participle/Infinitive/PP), Nominalization, Apposition, Premodification (Adj/Participle/Adv...)
Info Processing	Focus, Fronting, Inversion, Postponement, Emphatic, Reinforcement

About the Authors

Suthathip Thirakunkovit, the first and corresponding author of this study, is an assistant professor in the Department of Applied Linguistics, Faculty of Liberal Arts, Mahidol University, Thailand. Her current research interests cover language test development, test validation, assessment literacy, corpus linguistics, and second language writing.

Seongha Rhee is a professor at Hankuk University of Foreign Studies, Seoul, Korea. He received his Ph.D. in Linguistics from the University of Texas at Austin in 1996. His primary research interest is to identify cognitive and discursive mechanisms that enable language change from the crosslinguistic and typological perspectives.