THE UNIVERSITY
OF BIRMINGHAM

Department of English

Centre for English Language Studies

MA TEFL / TESL / Translation Studies, Open Distance Learning

Name	John Larson 665728
Country where registered	Japan
Dissertation title	PREDICTIVE SKILLS AND PROFICIENCY: an Investigation of General Language Proficiency and Its Relation to Pragmatic Expectancy
Submission date	September 2008
Submission	First Submission
Name of supervisor	Dr. Brent Wolter

DECLARATION

I declare:

a) that this submission is my own work;

b) that this is written in my own words; and

c) that all quotations from published or unpublished work are acknowledged with quotation marks and references to the work in question.

d) that this dissertation consists of approximately 12,436 words, excluding footnotes, references, figures, tables appendices & long quotations.

Signature: John Larson

Date: 2008 / 08 / 31

PREDICTIVE SKILLS AND PROFICIENCY:
an Investigation of General Language Proficiency and Its Relation to Pragmatic Expectancy

by

John Larson

A dissertation submitted to

The School of Humanities of the University of Birmingham

in part fulfillment of the requirements for the degree of

Master of the Arts in Teaching English as a Foreign / Second Language

Supervised by Dr. Brent Wolter

This dissertation consists of 12,436 words

Centre for English Language Studies
Department of English
University of Birmingham
Edgbaston
Birmingham B15 2TT
England

Submitted September 2008

ABSTRACT

John Oller's Pragmatic Expectancy (PE) is a unitary competency hypothesis which posits predictive skill as the foundation of language proficiency. Oller's evidence for this theory is a strong correlation between skills-based proficiency tests and predictive-skill-based pragmatic tests such as cloze tests. While there are strong arguments against PE, there may also be reason to reconsider Oller's theory. The effectiveness of pragmatic language tests and their as yet unexplained correlation with orthodox proficiency tests are two areas explored by this experiment.

Using a novel type of pragmatic test which uses corpus data as assessment criteria (the OpeC test), this experiment reproduced Oller's original findings, showing strong correlation between pragmatic tests and skills-based proficiency tests. By performing this experiment first-hand it was hoped that the underlying causes of this correlation could be closely examined.

The results validated this experiment's hypothesis by showing a significant correlation between scores on the OpeC and a skills-based proficiency test (called the GTEC). In addition, in-depth analyses led to many insights into the underlying causes of this correlation. Also, this correlation suggested further research is needed in the both the fields of predictive language skills and of the validity of corpus-based pragmatic tests.

ACKNOWLEDGEMENTS

I am most grateful to my supervisor Dr. Brent Wolter for all the help and encouragement.

I would like to thank the students and teachers of my school who participated in this experiment.

I am also thankful to Jonathan Skinner for the use of his students at his school in New Zealand.

I would like to thank my wife Kaori for her unceasing support throughout this course.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

This dissertation examines the results of a novel type of test which uses corpus data as assessment criteria to evaluate predictive language skills. The purpose of this study is to provide empirical evidence that supports the idea that proficiency can be effectively measured through testing designs that require test-takers to use predictive rather than isolated language skills.

Predictive skill testing, or pragmatic testing, was suggested by John Oller in the late 1970s. Oller's theory stemmed from very strong correlations between pragmatic tests and general skills-based proficiency tests. Citing this correlation as evidence, Oller proposed that pragmatic expectancy was the single underlying ability to general proficiency. The term pragmatic expectancy is used to signify the ability to predict subsequent linguistic structures based on previous structures. Because he saw language proficiency as "consisting of such an expectancy generating system" (1979: 16), Oller argued for a radical change in language instruction and assessment. Instead of focusing on separate language skills such as listening, reading, writing and speaking, Oller supported a move towards education and assessment which focused on predictive skills – the pragmatic expectancy – of language learners.

This dissertation focuses on the results of an experiment which resembles those supporting Oller's theory of Pragmatic Expectancy. In this experiment the results of a skills-based orthodox proficiency test were compared with the scores of a pragmatic test, as was done in Oller's original studies. This was done in order to more closely examine the correlations Oller and others found

between pragmatic and orthodox proficiency tests. By performing this experiment firsthand, it was hoped that the relationship between predictive skills and general proficiency could be better understood. Instead of using cloze tests as in Oller's original experiments, this study made use of a pragmatic test created specifically for this study called the OpeC. The OpeC makes use of data taken from the 450-million word Bank of English corpus (jointly owned by Harper Collins Publishers and the University of Birmingham). This particular type of test was used to avoid problems with cloze assessment procedures, as well as to investigate the effectiveness of corpus-based tests.

It was hypothesized that in this experiment a strong correlation would be found between the scores of the OpeC and the scores of an orthodox proficiency test. It was hoped that investigation into any correlation found between these two tests would shed light on the results which form the foundations of Oller's PE, as well give insight into the possibilities and practicalities of corpus-based testing.

CHAPTER 2 JOHN OLLER'S THEORY OF PRAGMATIC EXPECTANCY

Oller believed that the ability to predict subsequent language structures using previous structures is the one skill which underlies all other language skills. This theory, and the predictive skill on which it is based, are both called pragmatic expectancy. However, to avoid confusion, throughout this dissertation I will use the term 'Pragmatic Expectancy' (PE) to indicate Oller's theory and the term 'pragmatic expectancy' to refer to the predictive skill upon which his theory is based.

The aspects of Oller's PE relevant to this study can be summarized into three main focus points. In Chapter 2, the first section explores Oller's Pragmatic Expectancy in-depth. The second section takes a brief look at the evidence both supporting and undermining PE. The third section surveys the different kinds of tests which were first used by Oller to support his ideas. These points will be summarized in sections 2.1, 2.2, and 2.3. In section 2.4, I will discuss the rationale for further inquiry into PE.

2.1 Pragmatic Expectancy and its ramifications

Fluent communication is a complex and demanding task, and yet it seems a task which any normal mind can achieve. Consider briefly the intricacy involved in communication. First it must be noted that the entirety of any moment or experience can never fully be communicated through language. The mind knows more than it can say in words. Even so, it has more words than it can easily process in a given moment, by some accounts over 150,000 in an

average person's mental lexicon (Aitchison, 2003: 5). The rules by which those words are strung together and the number of possible combinations of words those rules allow is often said to approach infinity (Chomsky, 1965: 8). Yet, somehow, everyone everywhere accomplishes communication with ease, more often than not while concentrating on 'more demanding' things such as driving, shopping and dancing. People somehow make the complex task of communication seem easy. In order to accomplish such a demanding task with such apparent ease, there is little choice but to assume the mind has some means of extremely efficient language processing.

John Oller's Pragmatic Expectancy (PE) is one explanation for how such processing is achieved. PE is fundamentally a theoretical description of how the mind goes about processing language. In PE, prediction is essential to language use. In order to quickly process language, the mind must predict words, phrases or other elements of language which are likely to follow after the words or phrases before. As the mind then expects what is to follow, it can significantly reduce the amount of potential words or phrases it must discern between. Oller theorized that these calculations happen both consciously and unconsciously, and on many different levels of language complexity. (Oller, 1979: 25)

For instance, the subject of this chapter is John Oller's Pragmatic Expectancy, so a reader takes for granted the fact that every segment here relates to PE. It is highly unlikely that the next paragraph will be about theoretical particle physics. In the improbable event that the next paragraph indeed turns out to be

about physics, the rational reader would look for some connection, whether through metaphor or example, to PE. The same kind of prediction happens inside paragraphs and across sentences as well (Oller, 1979: 42). Without this, the sentence you are currently reading would be difficult to understand, as the second word ‘this’ refers to the method of language prediction that has been discussed over the past couple hundred words. This predictive skill lets a reader of a text decipher words that might be left out. Another example of this predictive proficiency is the ability of capable language users to disregard mistakes in spelling, pronunciation and grammar.

It is perhaps easy to understand how this kind of predictive skill can help receptive language tasks such as reading and listening. However, Oller applied this skill even more broadly by suggesting that it forms the basis for all areas of language proficiency. He proposed that even speaking and writing skills are derived from a language user’s pragmatic expectancy, helping fluency by predicting words that are likely to follow and making those words more readily available for use.

As Oller’s hypothesis saw all language skills stemming from a user’s pragmatic expectancy, his proposal later came to be known as a “Unitary Competence Hypothesis” (McNamara, 2000: 15). Oller’s hypothesis was supported by research which showed that language students taking one sort of test tended to perform similarly when taking a different sort of test. In short, Oller claimed that, by and large, students who do well on listening tests will also do well on writing tests; students who are poor at verb transformation also lack the ability

to use articles correctly, and so on. “Why,” Oller asked rhetorically, “should tests that look so different in terms of what they require people to do correlate so highly?” (1979: 60).

Oller expanded his claims even further to encompass other forms of learning. Taking as evidence data which correlates IQ scores and language proficiency (Oller & Perkins, 1978), Oller linked pragmatic expectancy and general intelligence. His logic posited that, just as the basis of language learning rests on the skill of pattern prediction, the same or a similar expectancy device is at work that, in essence, determines how smart people are.

Oller claimed that pragmatic expectancy is (at least) the cornerstone of language proficiency, and he used this claim as rationale for a radical shift in language education and assessment. Oller maintained that forms of language education and testing which focus on separate language skills, usage or forms are intrinsically flawed. Instead, argued Oller, language education would do better to focus its energies on testing and development of the underlying expectancy device – the language learner’s pragmatic expectancy.

While the basis for Oller’s Unitary Competence Hypothesis may seem reasonable in some respects, many of his contemporaries did not agree to the lengths to which he took his ideas. In the next section, some arguments for and against Pragmatic Expectancy are explored.

2.2 Oller’s PE, data for and against

Oller’s hypothesis was supported by data which showed strong correlations between tests of widely varying language skills and language aspects; in

particular, the tendency for students to show similar levels of achievement on the different sections of skills-based proficiency tests. In the 1970s when Oller first proposed PE, most proficiency tests currently in use were skills-based tests. This same type of test is still widely used today. Skills-based tests assess language ability by testing some combination of the four skills regularly associated with language use – reading, writing, speaking and listening. Oller categorized skills-based proficiency tests as discrete point tests, as their aim is to separate out aspects of language for testing, rather than attempting to test language proficiency holistically. These skills-based, discrete point tests rest on the assumption that by testing each language skill separately, and then by combining the scores of the separate sections of the test, the participant's overall language proficiency can be known.

Oller noticed that students tended to perform similarly on each separate section: students who did well on the listening section tended to also do well on reading, and so on. According to Oller, the reason behind the scores' similarity was two-fold. Firstly, Oller argued that skills-based discrete point tests were in reality unable to completely segregate each skill, and were therefore unable to actually test each skill separately. More to the point, Oller maintained that all of these skills are based on the same underlying predictive ability – pragmatic expectancy - so even if all four skills could be isolated, any such test results would be necessarily similar.

In addition to comparing different discrete point tests against each other to show correspondence, Oller also supported his claims by comparing scores of

discrete point tests with those he considered to be pragmatic tests. Briefly, pragmatic tests are tests designed to assess a participant's pragmatic expectancy. (They are discussed in detail in section 2.3.) Oller showed that discrete point tests and pragmatic tests also showed strong correlations - evidence which he used to further support PE.

Oller not only used his own studies to support PE, but also referenced independently collected data. The earliest such independent study showed strong correlations between dictation test results and the results of various listening and writing scores of college-aged FFL students (Valette, 1964). Alongside cloze tests, Oller saw dictation as a valid method of evaluating pragmatic expectancy. He suggested that the correlation between dictation test results and typical listening and writing test results supported his idea of pragmatic expectancy as the foundation of language proficiency. Three separate experiments similar to Valette (and run by Oller in conjunction with various colleagues; Oller, 1970, Oller and Streiff, 1975, Oller and Conrad, 1971) had the same results showing, "dictation as accounting for more of the total variance … than any other part" of the language tests run in these experiments (Oller, 1979: 58).

Oller also extensively used experiments involving cloze tests to support his claims. A University of Colorado study (Darnell, 1968) showed that when compared with separate sections of the skills-based TOEFL, cloze tests scores correlated most strongly with the listening section, giving a correlation score of .73. Because these cloze tests did not correlate as strongly with the

vocabulary and reading comprehension sections they closely resemble, Oller believed this to be evidence that the cloze tests were in fact measuring a proficiency factor more fundamental than listening or reading, etc. The same study showed further evidence of this. The correlation between the same cloze test results and the aggregate of all sections of the TOEFL was .85. Again, similar tests performed by Oller achieved similar results; correlating cloze tests scores with more orthodox proficiency testing methods (Oller, 1972).

Lastly, experiments showing very high correlation between cloze and dictation tests seemed to confirm to Oller that these two tests were evaluating the same underlying language skill. While strong correlations between cloze test scores and conventional proficiency test scores were found in Oller's 1972 study, these were overshadowed by the correlation found between cloze and dictation scores. Oller replicated these results later that same year, finding correlations of .74, .84, and .85 between cloze and dictation scores.

Essentially, Oller used these data to support two separate, but closely related claims. Firstly, he claimed the extremely high correlation between such seemingly diverse tests such as cloze and dictation showed that in fact they were testing a language skill which underlies listening, writing, vocabulary knowledge and etc. Secondly, when cloze and dictation tests correlated higher with the sum of the parts of a skills-based proficiency test than they correlated with those parts individually, it meant that the skill being tested by the cloze and dictation tests was more fundamental than isolated skills such as listening, reading and the like. Using these claims as bases, Oller further postulated that

the widespread practice of assessing language proficiency by separately testing different language skills was unnecessarily complicated and as such could easily produce misleading results.

Unfortunately, Oller's claims, as well as his supporting research data, have been attacked on many fronts. Cummins (1980) argued that all native speakers, regardless of IQ, develop basic interpersonal communication skills. This outlook would seem to oppose a view of language ability as stemming from IQ. Cummins, alongside Krashen (1978) and many others, maintained that the skills used in linguistically challenging tasks, such as writing a dissertation, are quite different from the skills used in interpersonal communication. This difference was evident in discrepancies in the results of tests which evaluate language manipulation and communicative competence respectively (Cummins, 2001). This discrepancy led them to the conclusion that linguistically complicated tasks were not governed by the same cognitive systems as common communicative tasks. While this research did not definitively disprove Oller's theory, it did lend credence to the belief that there is more than one significant factor relating to language proficiency.

Were these the only types of criticisms of Oller's view, he would undoubtedly be a much better known linguist than he is. Unfortunately, the research that Oller used to support his hypothesis was also heavily criticized. Ellis (2003: 281) called his results, "inconclusive and even plain wrong". Vollmer and Sang (1983) used Oller's same data but more complex statistical analyses to arrive at different conclusions, also putting Oller's original claims in doubt. After

Vollmer and Sang's analysis was published, under criticism of both his research practices as well as the coherence of his theory, Oller admitted that his theory Pragmatic Expectancy was perhaps overstated.

And perhaps it was. Most of the evidence supporting Oller's PE can be seen as circumstantial. Correlations between test scores do not necessarily imply a cause and effect relationship between the skills being tested. While the existence of a global language proficiency factor would explain the correlation between different skills-based discrete point tests, it certainly is not the only explanation possible. Another factor, IQ for example, could well have been responsible for the scores' correlations. Even if there exists a single, underlying language proficiency factor, Oller gives no definitive proof that this factor is pragmatic expectancy as he describes it.

Furthermore, Oller's arguments are inconsistent when taken to their logical conclusions. He argued that pragmatic expectancy was a predictive skill underlying all other language skills, and at the same time he argued against the validity of skills-based and other discrete point assessment techniques. Yet, if pragmatic expectancy does in fact underlie all other language skills, then it would stand to reason that any skills-based test would necessarily be assessing pragmatic expectancy, albeit indirectly. It is likely that these inconsistencies, coupled with the drastic extents to which Oller (mis?)applied them, ultimately doomed his theory of PE to obscurity.

What is more, Oller's theories at the least called for considerable revisions to current teaching and testing practices, while at worst they heralded the

restructuring of foundational beliefs in education and learning. Any theory which calls for such drastic changes to established practices in language testing, teaching, and education in general is bound to be strictly scrutinized.

However, radical declarations and lack of decisive data do not take away from Oller's initial observation that discrete point test scores often coincide with one another. Nor do they subtract credibility from Oller's observations that skills-based test results regularly correlate with pragmatic test results. Nor even do they take away from the observation that successful language manipulation most likely involves skilled predictions about what words and phrases are likely to come next in the flow of communication. Even if one dismisses PE outright – which perhaps is itself a rash declaration - the relationship between pragmatic tests and discrete point tests, and what connections this relationship may have with a predictive language faculty, bears further examination. In the next sections these different types of test and examples of each are more fully introduced.

2.3 Pragmatic and Discrete Point Tests

In his book *Language Tests at School*, Oller spent most of his time lauding tests which are capable of evaluating language learners' pragmatic expectancy. These he dubbed 'pragmatic tests', with a pragmatic test being

> "any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate

> sequences of linguistic elements via pragmatic mappings to extralinguistic content" (p.38).

In simpler terms, a pragmatic test is one which causes participants to use their pragmatic expectancy to understand and produce meaningful language. Pragmatic tests compel participants to use learned grammatical and idiomatical language structures as well as contextual clues to predict meaningful language items, thereby utilizing their pragmatic expectancy. (p.70)

In *Language Tests at School*, Oller set pragmatic tests in opposition to discrete point tests. Discrete point tests are those which focus on a single part or multiple parts of language use or usage. Tests which focus on, for example, certain grammatical transformations, certain vocabulary items, or certain skills are defined as discrete point tests. Oller took issue with the tendency for language teachers to use these sorts of tests to evaluate student proficiency, as he felt they fail to accurately test the predictive pragmatic ability that underlies all language skills. Pragmatic tests, on the other hand, assess the underlying expectancy grammar of participants and are therefore, according to Oller, much more reliable proficiency tests.

2.3.1 Example of a pragmatic test: the cloze test

One pragmatic test on which Oller particularly concentrated in *Language Tests at School* is the cloze test. In general, a cloze test has blanks which must be filled using the language knowledge of the participant. There are many different types of cloze test that in turn use many different types of language structures. Take, for example, the variety of the following questions (Oller, 1979: 341)

(1) one, t___, t___, f___, _ive, ___x, _____n

(2) Four _____ and seven _____ ago _____ _____...

(3) After the mad dog had bitten several people he was finally sxghtxd nxxr thx xdgx xf txwn xnd shxt bx a local farmer.

(4) It is true that persons _____ view the treatment of mental _____ from a clinical perspective tend _____ explain socioeconomic and ethnic differences _____ biological terms.

A typical cloze test starts with a text in the target language. Words have been removed from the text and blanks inserted in their place. The test taker must then use his or her knowledge of grammatical language usage, common phrases and idioms, and contextual clues to fill in the blanks.

Words removed from a cloze test are generally removed either randomly, or systematically – i.e. every n^{th} word. Answers generally can be judged in two ways. One scoring method simply uses the original text as a key for deciding right or wrong answers. A more lenient method allows more than one correct answer, provided the word in question makes sense grammatically and contextually. Yet each of these methods is imperfect. In either case, there is a necessary compromise that must be made when evaluating the answers.

The first method of evaluation is straightforward, but extremely restrictive. In this method the original text is used as a key, and any answer which differs from the original text is counted as a mistake. Take for example question 4 above. The original text dictates that the answer to the blank after mental is

illness. However, what of the responses *instability*, or *retardation*, or even *illnesses*? Using this strict method of evaluation, the above three responses are just as wrong as if the participant had left the answer blank.

The second method of evaluation is more open, but has a high potential to be ambiguous. In this second method, various answers are allowed as long as they are appropriate to the context. Determining what answers are appropriate and which are unnatural almost necessarily involves arbitrary decisions on the part of the assessor. Are the above three alternative answers appropriate? If so, what is to be made of less plausible answers such as *sickness* or *disease*? The phrase *mental illnes*s may sound more natural than *mental sickness,* so is it worth more points? How many more points? Intricacies such as these are seemingly endless and are extremely difficult for the assessor to untangle.

While there are no easy answers to the scoring predicaments noted above, it is important to remember that Oller uncompromisingly supported the use of pragmatic proficiency tests in general and cloze tests in particular. The tests Oller took issue with were discrete point proficiency tests, which see wide use in language education today.

2.3.2 Example of a discrete point test: the GTEC

A large portion of *Language Tests at School* is a diatribe against skills-based proficiency tests. Skills-based proficiency tests are discrete point tests. Each section - reading, writing, listening and speaking - is in itself a separate discrete point test as it focuses on testing one separate skill. These kinds of tests were fervently argued against in *Language Tests at School*. According to Oller they are

ineffective in their goals and give redundant results. Nonetheless, makers of these kinds of test operate under the logical assumption that because these four skills make up the majority of language usage, then compiling the proficiencies of these skills should give an accurate indication of a language user's overall proficiency. One example of the countless skills-based proficiency tests currently in extensive use in Japan (where the present study was conducted) is the Global Test of English Communication or GTEC. As the GTEC played an important role in the following experiment, a closer look is taken at its contents in the second section of Chapter 3.

2.4 Rationale for the present study

Despite some inconsistencies and a lack of decisive supporting evidence, John Oller's PE is nonetheless an intriguing idea. Were it widely accepted, a Unitary Competence Hypothesis such as PE would change language teaching and learning considerably. Even if PE as described by Oller was overstated, questions raised by Oller's preliminary observations that led to his theory still have validity. Why do tests that purport to assess entirely different skills produce similar results? Do tests which require participants to use their predictive abilities correlate strongly with skills-based proficiency tests? Most importantly, if PE is not the correct explanation, what else can explain the consistency of the correlations described by Oller?

In order to more fully understand these questions, it was necessary to reevaluate Oller's 1972 study. By examining correlations between the skills-based proficiency test GTEC and a modified cloze test, I have attempted

to re-evaluate Oller's initial findings. This was done in order to accomplish two goals. First, since the conclusions reached using Oller's original supporting evidence were strongly questioned, an independent test, coupled with a careful examination of the data may help to alleviate doubts as to the interpretation of any correlations between test results. Second, in modifying the cloze test procedure I have attempted to eliminate the aforementioned compromises involved in the scoring procedures associated with cloze tests. The means by which this has been accomplished is the main topic of Chapter 3.

CHAPTER 3 EXPERIMENT

The goal of this experiment was to provide empirical evidence that supports Oller's idea that proficiency can be effectively measured through testing designs that require test-takers to use predictive rather than isolated skills. In order to achieve this, a novel type of cloze test was created. The test is called the OpeC, which stands for Open-ended Cloze test. It was created especially for this experiment and uses the Bank of English corpus as criteria in its assessment. In utilizing a corpus, I hoped to address problems associated with cloze test scoring procedures. As I mentioned previously, scorers of cloze tests often face a difficult dilemma. On the one hand, they can choose to disregard any answer at variance with the original text. This process may lead to better reliability in scoring but may also result in valid answers being marked wrong. On the other hand, scorers can instead choose to allow ‘acceptable’, ‘natural-sounding’ or ‘grammatical’ answers, but in doing so they must undertake the extremely precarious task of defining ‘acceptable’, ‘natural-sounding’ and the like. Instead, using a corpus to arbitrate what answers are acceptable permits the test maker to accept many various answers without subjectivity.

As stated earlier, in Oller’s research cloze tests results correlated with skills-based proficiency exam results. The purpose of this experiment then was to test the hypothesis that scores on a pragmatic test (the OpeC: section 3.2.1) would significantly correlate with the scores of a skills-based proficiency test (the GTEC: section 3.2.4). It was conjectured that Oller’s original supporting

evidence would be upheld, and the results of a corpus-based pragmatic test would closely correlate with the results from an orthodox skills-based proficiency test. Although the interpretation of Oller's original supporting evidence was disputed, the correlation he found between pragmatic and skills-based test scores was not. Therefore, it was assumed that whatever connection there may be between the skills used to complete these tests would also be found in this experiment. It was hoped that by re-evaluating this connection, the basis of this relationship could be more thoroughly examined.

3.1 History and Rationale behind the OpeC

Originally, the OpeC was modeled after a word association test. Word association tests are sometimes used to evaluate the links between words in the mind. In a very simple form of word association test, a single word prompt is given, and the participant then answers with the first word that comes to him or her. During a previous research project using such a test (Larson, unpublished) it was noticed that the majority of responses were related to the prompts semantically, rather than grammatically or idiomatically. While semantically related responses cannot be considered invalid, they convey little information about the target language ability of the participant. One example from that study was the common response of 'white' to the prompt 'milk', which proved generally uninformative (see also Meara, 1983). Milk is white regardless of the language being used.

In order to combat this problem, the prompts given in the OpeC were phrases instead of words, and the participants were instructed to complete sentences

rather than write one word answers. It was reasoned that by doing this, participants would be compelled to use any grammatical and/or idiomatical skills they may have in the target language. This in turn was expected to produce results which more clearly reflected the structure of the participant's pragmatic expectancy. Even in its beginning stages, the aim of the OpeC was to compare participant-produced structures with native-produced structures.

Oller's PE and his use of cloze tests are in close accord with the usage of the OpeC as detailed above. Certainly, the OpeC can be considered a pragmatic test. It requires participants to use their pragmatic expectancy to formulate sequences of linguistic elements, while the corpus used as an assessment tool determines how well those sequences conform to the normal constraints of native produced English. Like cloze tests it assesses participants' ability to complete sentences appropriately. Unlike cloze tests however, the OpeC is not constrained to an extremely limited range of possible answers. The OpeC allows the participant great freedom when answering and yet avoids the arbitrariness often encountered in cloze tests.

Anyone who attempts to allow more than one correct answer in a cloze test faces many problems relating to the test's reliability. As stated before, decisions about what answers are "acceptable" or "natural" or "merely grammatical" are always difficult and often arbitrary. Added to this is the question of whether to quantify these answers depending on the answer's 'naturalness'. Take again the example from section 2.3.1:

"…the treatment of mental _____ from a clinical perspective…" (Oller, 1979: 341)

The original text dictates that the answer for the blank after mental is *illness*. Other potential answers are *instability* or *retardation*. If these answers are acceptable, it then begs the question of whether they are as acceptable as the answer in the original text. Some other possible answers are *ills* or *sickness* or even *badness*. If these answers are unacceptable, it then begs the question of whether they are as unacceptable as completely ungrammatical or nonsensical answers. It seems reasonable that there are some answers which are less acceptable than others, and yet not completely unacceptable. To count these as either entirely wrong or entirely right would limit a test's reliability and sensitivity.

In order to combat these problems associated with cloze tests, the answers to the prompts of the OpeC were scored using a sliding scale developed using corpus data from the 450-million word Bank of English corpus. By accessing such a large corpus, I hoped to make the test sensitive, and yet avoid arbitrary and subjective decisions as to what constitutes 'acceptable' language usage.

In these two ways, the OpeC builds on the strengths of both cloze and word association tests while discarding many of their limitations. The OpeC is not only a more well-controlled word association test, it is also an updated version of Oller's pragmatic cloze test. The OpeC shares its chief goal with these two tests: to bring into focus the grammatical and idiomatical aptitude of the participants. Whether this aptitude is synonymous with Oller's pragmatic

expectancy, and whether that itself is fundamental to language proficiency is beyond the scope of this study. Rather, this experiment was concerned primarily with the correlation between the scores of a corpus-based pragmatic test and an orthodox skills-based proficiency test. If such a correlation could be found, the secondary goal of this experiment was to determine more definitively where said correlation comes from, and if this correlation indeed signifies Oller's PE as a valid premise.

3.2 Materials

Three things were necessary in order to search for this correlation: a corpus-based pragmatic test, an orthodox skills-based proficiency test, and participants to pilot these two tests. Most significant perhaps is the discussion describing the pragmatic test the OpeC. As it is the most important tool used in this experiment, discussion of the OpeC takes up the first three subsections of this section. The last two subsections then deal with the GTEC and the participants.

3.2.1 The Open-ended Cloze Test (OpeC)

The Open-ended Cloze Test (OpeC) is based on the cloze testing procedure. However, instead of single words or phrases removed from a text, participants are given the beginnings of sentences which they then must complete. In completing these sentences, participants must use grammatical and idiomatical skills as well as any contextual clues included in the prompts. The skills required to successfully finish a sentence are in many ways the same skills used to complete cloze test questions. However, as these open-ended sentences are

not connected to a longer text, participants are free to give an extremely wide range of answers. In this respect, this style of prompt allows more freedom of choice, and thereby more closely resembles natural language creation.

The entire test consists of thirty prompts and brief directions. English learners can expect to complete the test in less than 20 minutes; native speakers can expect to finish in much less time. The first few questions are reproduced below in Table 1, while the entirety of the test is shown in Appendix 1.

Directions: Please finish the sentences. (The above directions reproduced in the student's native language.)
1) Give me a_____
2) Are they_____
3) I think we_____
4) I get the_____
5) Until it_____
6) It happened_____
7) When you get_____

Table 1: the first seven questions of the OpeC

3.2.2 Prompt development

There were two different but related guidelines that were followed when choosing what prompt phrases to include in the test. First and most importantly, the words had to be easily understood by participants of many different levels of English proficiency. Thus, all words in the 30 prompts were among the 1000

most frequently used English words according to the frequency listings contained in the Bank of English. The second guideline was that there needed to be a large enough number of instances of the prompt phrase in the Bank of English corpus. If the occurrences of the phrase were too few, then the chances were greater that a given answer would score zero points, which in turn would have made the test less sensitive. In this case, all prompt phrases appeared over 700 times in the corpus, with some phrases appearing over 2000 times.

The use of phrases rather than words allowed participants to use their grammatical, idiomatical and contextual predictive skills, thereby allowing these skills to be assessed. This allowed us to avoid the overwhelming preponderance of semantically related (and thus largely uninformative) responses usually observed in normal word association tests.

3.2.3 Evaluation

This style of test required a novel scoring method. In order to cope with the wide range of answers participants may give, a corpus was used to evaluate the validity of the answers. Essentially, collocation was used to establish point values – the stronger a given answer collocated with the prompt, the higher the point value of that answer.

Participants were expected to complete sentences using language which occurred to them spontaneously. These responses were then compared with data from the corpus and points were assigned depending on how well the answers collocated. As such, there were really no ‘right’ or answers, but there were ‘wrong’ answers that did not result in any points being awarded.

Participants were free to use as many words as they thought necessary when completing the sentences started by the prompt phrases, but only the first four responses were given point values for reasons explained below.

As this test aimed to analyze the grammatical and idiomatical skills of the test-takers, it used collocation data collected in a novel way. Generally, collocation data is gathered by focusing on one word at a time. This word is called the node. Words that commonly appear before or after the node are called collocates. Researchers can choose how far from the node they wish the computer to search for these collocates. Generally speaking, the span of words appearing four places before and behind the node word is considered to be its "relevant verbal environment" (Sinclair, 1991: 175). More or fewer appearances of the collocate within this span of words before and after the node are linked to stronger or weaker collocation. Also taken into account is the collocate's relative frequency in the corpus at large. Simple collocation is a function of the two frequencies: the frequency of the collocate in the corpus at large and the frequency of the collocate in the span of words before and after the node.

The first difference between standard collocation data collection procedures and those used in the OpeC is obvious. As OpeC prompts are the beginning of sentences, nodes consist of strings of words rather than single words. This has the effect of limiting the number of examples of these nodes in the corpus. A common word like 'because' occurs in the Bank of English corpus approximately half a million times. However, when paired with another common word the number of occurrences drops dramatically. The phrase

‘because it’ appears less than 50,000 times in the corpus – more than an order of magnitude less frequently than ‘because’ and two orders of magnitude less frequently than the single word ‘it’.

The second difference between standard collocation data collection and that used in the OpeC is that collocate data is collected separately for each place after the node phrase. This is necessary if the evaluation data is to properly reflect native produced English language structures. In order to explain this more clearly, the first five collocates and t-scores of the last prompt, “Ask her” are given below in Table 2.

Rank	1st word after prompt phrase		2nd word after prompt phrase		3rd word after prompt phrase		4th word after prompt phrase	
1st	to	19.7081	she	16.6946	She	7.8674	she	7.0674
2nd	if	12.3373	her	5.8854	Her	5.9691	her	4.5624
3rd	about	10.4736	come	5.5713	Me	4.6952	you	3.8336
4th	what	10.4441	do	5.0851	You	4.4995	like	3.4962
5th	out	8.0527	go	3.9135	Can	4.1402	about	3.368

Table 2: the first five collates and t-scores of the prompt “Ask her”

The point values for each response were calculated differently depending on the distance from the last word from the prompt phrase. This is evident in the chart above. Note that the first collocate for the second, third and fourth response are the same word: ‘she’. This may seem odd at first, but consider the following sentences:

Ask her what she said to him.

Ask her about who she likes.

Ask her about that thing she bought.

Without considering placement, the word ‘she’ correlates very highly with the prompt phrase ‘Ask her’. However, ‘she’ given as the first response would have a score of zero because the corpus lacks examples of ‘Ask her she…’ - and rightfully so, as it is grammatically dubious.

Notice also that the point values for the responses tend to decrease the further removed they are from the prompt. This is logical in that the more words a sentence has, the more grammatically possible permutations there are.

In order to more fully understand this scoring process, it is perhaps necessary to score the following model answer:

Ask her if she can go to the party.

Using Sinclair’s idea of relative verbal environment, the first four words results an overall score of around 34.8 (a very high score) as laid out below:

if(12.3373) + she(16.6946) + can(4.1402) + go(1.6116) = 34.7837

This novel style of evaluation has succeeded in eliminating the otherwise unavoidable compromise scorers face when scoring typical cloze tests. Furthermore, by assigning various point values to different answers depending on the strength of their correlation, it is hoped that the OpeC has been made more sensitive than would normally be possible with a simple right/wrong style of assessment.

3.2.4 GTEC

GTEC stands for Global Test of English Communication and is the brainchild of the Benesse Corporation and Berlitz International (Benesse, 2008).

Although there are many different variations of GTEC, the version used in this experiment was a paper-and-pencil multiple choice test called GTEC for Students. The proposed aim of GTEC for Students is to assess the language proficiency of Japanese junior high school, high school, and college students. (Benesse, 2008)

GTEC is administered in two consecutive sessions of 45 minutes each. The first half contains the reading section, and is worth a total of 250 points. The second half contains both the listening and writing sections. The former is worth 250 points and the latter worth 160. Thus the entire test has a maximum possible score of 660. The test is further broken-down into more detailed sections. Point values for each of the separate sections are detailed in the translated table below.

Test Contents	# of Qs	Points
Reading / Marksheet test	36	250
A - Vocabulary and idioms	12	
B – Information Scanning and general comprehension	12	
C – Summarization	12	
Listening / Marksheet test	40	250
A – Illustration description	10	
B – Conversational Q and A	10	
C – Problem Solving	10	

D – Summarization	10	
Writing / Free Writing	1	160
Expressing and Developing Opinions *3 points: grammar, vocabulary, and organizational development will each be ranked on an 8 rank scale.	1	
Totals	77	660

Table 3: Question type and point breakdown of GTEC

Test sections are further broken down by type; however point values for these subsections are not set. Instead, individual question point values are calculated as a function of the percentage of test-takers answering each individual question correctly.

The GTEC tests three skills separately: listening, writing and reading, which when added together give a test-taker's total score. In using this scoring method, Benesse seems to imply that the GTEC can measure a test taker's language proficiency by combining these three separate skills-based tests.

3.2.5 Participants

The participants involved with the experiment consisted of three groups. Two of these groups were Japanese high school EFL students: 70 second-year students (16-17 year olds) and 71 third-year students (17-18 year olds). The third group consisted of 48 ESL high school students studying in New Zealand and ranging in ages from 15-18 years old.

The ESL group members had various countries of origin, and their time spent in New Zealand varied from three months to thirteen years. The mean time living abroad for this group of students was two years, with a standard deviation of 2

and a half years. Despite these wide variations, it is almost certain that this ESL group is on the whole at a much higher level of English proficiency than either EFL2 or EFL3. There are two reasons that support this. First, the communication skills necessary to survive in an English-speaking country suggests that their exposure and motivation to learn English are deeper than those of EFL students. Secondly, at the time of the test they were all students in good standing at an ordinary prep school, where lessons are carried out entirely in English.

Both groups EFL2 and EFL3 consisted of students attending the same Japanese high school. These participants had therefore been exposed to the same teaching materials and methodologies. The Japanese educational practice of separating high school students according to standardized examination scores also insured these students all had relatively the same academic ability. These students were chosen because of these similarities to eliminate as many unknown variables as possible. The only difference between EFL2 and EFL3 is essentially that participants in EFL3 were one year older and had been exposed to one extra year of EFL teaching in Japan. Other than that, EFL2 and EFL3 were quite similar. As these two groups had reasonably comparable English language abilities, it was assumed that any test which could reliably assess their different aptitudes could be called a successful one.

3.3 Testing methods

All participants in groups EFL2 and EFL3 were given the GTEC in a classroom setting as part of a school-wide academic assessment. Neither group received any special instruction in preparation for the GTEC. Before taking the GTEC,

they were told that this test would not have any effect on their class scores or grades. Three weeks later, both groups were given the OpeC under the same circumstances. Students were not given any special directions prior to the OpeC other than those printed on the OpeC sheet for test proctors (see Appendix I). It was unfortunately not possible for participants in group ESL to take the GTEC. Group ESL therefore was given only the OpeC in a classroom situation during normal school hours.

Scores of these tests were then compared statistically to show if any correlation existed between the results. As in Oller's original supporting evidence, individual test scores of the OpeC and GTEC were compared with each other. It was hypothesized that a significant correlation would be found between these two sets of scores. The discovery of such a correlation, if present, was the primary goal of this experiment, and any validation or refutation of the hypothesis rests primarily on this correlation.

As a further investigation, comparisons were also made between the average scores of each group. Average scores of the OpeC for each EFL group were compared to average scores of the GTEC. It was hoped that the OpeC would prove capable of differentiating groups EFL2 and EFL3 in the same way as the GTEC. If found, such a connection would serve to corroborate any correlation found between individual scores of each test.

As a separate investigation into the efficacy of the OpeC as a language proficiency test, the average OpeC test scores of the group ESL were compared alongside the OpeC scores of EFL2 and EFL3. It was conjectured that because of

their different methods and motivations for studying English, the ESL group's average score would be significantly higher than that of group EFL3. This difference between the average OpeC scores of ESL and EFL3 was predicted to be larger than any difference found between the average scores of EFL3 and EFL2.

CHAPTER 4 RESULTS

In Oller's research, correlations were seen between scores of pragmatic tests (such as cloze tests) and scores of skills-based proficiency tests. Strong correlations were obtained, despite the different skills utilized in completing these tests. Here too, despite the different skills utilized in completing the OpeC and the GTEC, it was hypothesized that a significant correlation between the scores of these two tests would be found.

In the next two sections of this chapter the scores of these tests and any correlations between those scores will be examined in two ways. First and foremost, individual scores of the GTEC and OpeC will be compared to determine any correlation between them. In addition, the mean scores from all three groups of participants on the OpeC will be examined in order to determine if there were any significant differentiation between them. If the group differentiation realized by the OpeC coincided with that of the GTEC, then this would in turn support any correlation found between the individual scores of each test.

4.1 Correlations between individual scores of the OpeC and GTEC

The individual participant scores of the OpeC and those of the GTEC were compared using a Pearson correlation test. This was done in order to determine any relationship between individual scores of the OpeC and GTEC. This resulted in correlation coefficient of $r = 0.65$ ($p < .001$). This data shows that individual OpeC scores correlated fairly well with individual scores on the

GTEC. This significant correlation between individual scores of the GTEC and OpeC validated the hypothesis which conjectured a correlation between these test scores.

4.2 Differentiation between average group scores of the OpeC

In the next two subsections, the mean group scores of each test were compared in two ways. First, mean OpeC and GTEC scores for both groups EFL2 and EFL3 were compared in order to investigate any similarities in the ways these two tests differentiated these two groups. This was done in the hopes that if such a similarity was found, it could serve as support for the correlation found between individual test scores. Second, as a separate inquiry, mean OpeC scores for all three groups were compared against each other.

4.2.1 Comparisons of mean GTEC and OpeC scores of groups EFL2 and EFL3

For the two Japan-based EFL groups, the mean scores for the GTEC were EFL3: 450 (with a standard deviation of 81.6), and EFL: 368 (76.0). This showed that EFL3 performed, as expected, better on the whole than EFL2. It is worth noting that the point gap between the two groups is fairly large. As there is only one year of institutional English education separating these two groups, a large point gap such as the one found here indicated that this test was a sensitive one. A t-test performed to determine the equality of means produced a score of $t(139) = 6.17$, $p < .001$.

The mean scores on the OpeC for these two groups were EFL3: 214 (62.2) and EFL2:165 (79.2). This likewise shows that the third year Japanese high school students performed on average better than the second-year students. As this

test was not created specifically with Japanese high school students in mind (unlike the GTEC), it was surprising to see such a large point gap between EFL2 and EFL3. Again, a t-test was run to determine the equality of means and produced a score of $t(139) = 4.16$, $p < .001$. This showed that the OpeC was effective at distinguishing between the two groups, but not quite as effective as the GTEC. More importantly, the similarity of the average test scores pointed to a connection between the OpeC and GTEC, and therefore supported the individual score evidence presented in section 4.1.

4.2.2 Comparisons of the mean OpeC scores from all three groups

It was also assumed that mean OpeC scores for each of the three participant groups would be differentiated in the following order: group ESL would score highest, EFL3 next highest, and EFL2 lowest. This order is indeed what was observed. The mean average OpeC scores for each group along with their standard deviations are listed in the table below.

Group	Mean OpeC Score	Standard Deviation
ESL	364	62.1
EFL3	214	62.2
EFL2	165	79.2

Table 4: Mean OpeC scores and Standard Deviations for each group

The mean OpeC scores for each group were clearly differentiated in the order hypothesized. A one-way analysis of variance between groups (ANOVA) returned values of $f(2, 188) = 123.9$, $p < .001$. This shows that the OpeC was able

to significantly differentiate the average scores of these three groups from one another.

CHAPTER 5 DISCUSSION

In Chapter 5, the results will be examined in order to study the connection between these tests, as well as what may have caused the correlation between them, more closely. Then these conclusions will be related back to Oller and his Pragmatic Expectancy theory. Lastly, some comments will be given on corpus-based testing and predictive ability in general.

5.1 The hypothesis is supported

It is important at this point to restate the main hypothesis of this study which was first put forward in the beginning of Chapter 3. It was hypothesized that the results of the corpus-based pragmatic OpeC would closely correlate with the results from the orthodox skills-based GTEC.

All of the numerical data collected point to the conclusion that the hypothesis has been supported. The most persuasive evidence comes in the form of individual correlation scores that are in line with Oller's original findings. Individual scores of the OpeC were shown to significantly correlate with those of the GTEC. The correlation in this experiment of $r = 0.65$ is perhaps not as impressive as the $r = 0.84$ correlation score in Oller's 1972 experiment. Nevertheless it indicates a positive and significant connection between the scores of an orthodox proficiency test and a novel pragmatic test. Further support was seen in the differences between the average scores of group EFL2 and EFL3. The differentiations between the average scores of these groups were quite similar for the GTEC and the OpeC.

5.2 A closer look at the scores

In the same way as Oller's original evidence, this experiment's data provides some indirect support for his unitary competence hypothesis PE. If there does exist an expectancy skill which underlies all other language skills, then results such as these are to be expected. It is worth pointing out again that statistical data of the sort collected in this experiment, as well as in Oller's original experiments, by no means proves PE to be valid. Instead it simply supports PE in that the data collected match the results one would expect were PE valid. In this way, the results of various experiments cited by Oller, as well as of this experiment, all support Oller's PE.

However, when one investigates deeper into the results – into the causes behind the numbers – one finds that things are not as simple as the statistics above might suggest. Close inspection of individual student responses revealed that there seemed to be three factors which accounted for the majority of the test scores' disparities. These factors were: 1) the ability for students to understand the language contained in the prompt, 2) the length of the answer given, and 3) the extent to which the answer given matches frequent collocations in the corpus. In the next three sections, these three factors are discussed briefly. In some instances the rolls these factors played in the differentiation of the OpeC scores has been measured.

Before exploring these points, it is important to point out that any discussion of the causes behind participants' particular answers is conjecture. Although it is impossible to determine unequivocally the thought processes that led certain

participants to certain answers, it is nevertheless irresponsible to completely ignore such potentially useful data. As the creator and proctor of the OpeC, and as the regular instructor of all the participants in EFL2 and EFL3 groups, I have reason to be confident in my speculation. The conclusions reached in the following subsections should be treated as well-informed conjecture.

5.2.1 Participant (in)ability to understand the prompt

Despite sincere efforts to make the prompts of the OpeC easily understandable, there were still some participants who seemed to fail to understand all the prompts correctly. This inability to understand the language contained in the prompt materialized itself in two ways.

The first way this failure to understand presented itself was simple misunderstanding. Because of their inability to understand the prompt properly, participants answered with words not represented in the corpus. This is not to be confused with simply choosing the wrong word because one does not understand English well enough to know which word should come next. Instead, these wrong answers seemed to be brought on because of an inability to read and comprehend the prompt correctly. Unfortunately, simply by studying the answers to the prompts it is frequently impossible to discriminate between wrong answers brought on because of a mistaken reading, and wrong answers brought on through a lack of English collocative ability. Therefore, while an unscientific survey of individual answers gives the impression that this error method was quite common, it is impossible to measure accurately.

In order to understand better the difficulties measuring this error method, two examples are detailed below. Both examples stem from the prompt ‘Turn on the…’. The first answer to be examined was written

‘Turn on the paper.’

This mistake was almost certainly brought on by the participant’s inability to comprehend the prompt correctly. Were the prompt ‘Turn over the’ or even ‘Turn in the’, the answer ‘paper’ would be acceptable. This participant has probably mistaken the verb phrase ‘turn on’ for one of the other alternatives mentioned above. Another fact which supports this conclusion is that the prompt in question is number 18 of 30, which put it on the flipside of the test paper. The participant had literally minutes before turned over her paper.

Another mistake made from the same prompt was

‘Turn on the merry go round.’

It is less clear what mistake the participant is making in this case. However, it seems plausible that she mistook the word ‘Turn’ for the word ‘Spin.’ Nonetheless, the possibility that she is instead ordering someone to activate the merry go round is hard to dismiss totally.

These two examples illustrate how it is sometimes difficult to discern between non-scoring answers made because of a misunderstanding of the prompt, or a general lack of collocational competence.

The second (and more easily quantifiable) way this failure to understand the prompt may have materialized itself was in prompts to which participants

failed to write any answer whatsoever, leaving the response space completely blank. Although it is difficult to be completely certain that a blank answer indicates a failure to understand the prompt in every circumstance, it seems reasonable to assume that incomprehension is by far the most likely reason for blank answers. The number of blanks on each test sheet was compared to the individual test scores using a Pearson correlation test. This was done in order to determine how much effect blank answers had on test point totals. The test showed a correlation of $r = -0.58$, $p < .001$. This suggests that participants who completed more prompts scored higher than those who left blank answers. A correlation such as this was expected as the number of blanks and total OpeC score constitute a part / whole relationship. It is also likely that this behavior was instrumental in bringing about the close correlation between the GTEC and OpeC, as the inability to understand a three or four word prompt would certainly hamper a participants ability to score well on a general proficiency test of which reading is a substantial part. A comparison between the number of blank answers on the OpeC with total scores of the GTEC bears this out. These two had a correlation of $r = -0.38$, $p < .001$, which shows that to some extent at least, students who left more answers blank in the OpeC tended not to do as well on the GTEC.

It is easy perhaps to incorrectly attribute the inability to understand the prompt to poor pragmatic expectancy skill. Indeed, if Oller is correct in assuming that pragmatic expectancy is the root of all language proficiency then misunderstandings of this sort can have no other explanation. However, inability to understand the prompt and inability to appropriately finish the

sentence that is started by the prompt are two very different assessment criterions. Assessment based on the understanding of the prompt would constitute a discrete point test; inability to answer only indicates that the participant does not understand the vocabulary contained in the prompt. On the other hand, assessment based on the collocative appropriateness of the answer (which is OpeC's intended manner of assessment) evaluates a much wider and more fundamental swath of language knowledge and language skill(s). Despite the fact that it contributes to the overall correlation with GTEC results, an apparent inability to understand the prompt must be seen as a small, but unfortunate failure of the OpeC.

5.2.2 Answer length (brevity)

The second way in which scores seemed to differ from one another was in respect to the length of each individual answer. Answer length was compared with overall individual scores of the OpeC to search for correlations. This was done to determine what effect, if any, answer length had on score differentiation. As the first four answers to each prompt were scored, answers with less than four words were in effect treated as partially blank answers.

If answers of zero length – i.e. blank answers – were found to contribute to lower scores than answers of one word or more, then it stood to reason that answers of three or four words would tend to yield higher scores than answers of one or two words. This indeed was what was found. Answer length, discounting answers of zero length, was compared with individual OpeC point totals and the two values were found to correlate at a figure of $r = 0.41$, $p < .001$.

Here again, this correlation was expected because of the part / whole relationship shared between these two figures.

Unlike participants who left answers completely blank, these participants seemed to understand the prompt and simply felt the sentence could be finished by one or two words. An unscientific perusal of individual test answers turned up some instances where shorter, subjectively better answers scored fewer points than longer, but less native-like answers. Take for example the following three answers to the first prompt:

Give me a hand	Score: 6.7528
Give me a cup of tea	Score: 11.3135
Give me a money so I can	Score: 17.4945

Of these three examples, it seems clear that the third answer, ‘Give me a money so I can’, is the least native-like. However because of its relative length, (or instead perhaps because of the other two answer’s relative brevity) it scored considerably more points than the other two. One could also argue that the first answer demonstrates a richer understanding of English that the second. Here again though the length of the answer carries the most weight.

Therefore, it seemed that in some instances the OpeC unfortunately ended up discriminating against otherwise perfectly valid answers simply because of their brevity. These subjective observations aside, this factor surprisingly seemed to contribute to the overall correlation between the OpeC and GTEC. The number of words used by each individual participant, when compared

with their individual GTEC scores showed a correlation of $r = 0.37$, $p < .001$. This shows that participants who chose to answer OpeC questions with more words scored somewhat better on the GTEC.

5.2.3 Extent to which the answers collocate

Lastly, the scores were differentiated by how well answers matched frequent collocates found in the corpus. As this was the original aim of the OpeC, it was fervently hoped that this factor was the most influential in deciding score values. In order to determine whether this was the case, it was necessary to somehow set aside the two previously mentioned factors of prompt comprehension and answer length. This was accomplished by calculating each individual participant's average score per written word. In essence, each participant's total score was divided by the number of words that participant used to answer all the prompts. By doing this it was possible to ignore blank answers as well as disregard any negative effects caused by shorter answers. Each participant's average score per written word value was then compared with each participant's total OpeC score. As was hoped (and expected due to the part/whole relationship), the total OpeC scores correlated very strongly with the average score per written word, scoring a Pearson correlation of $r = 0.875$, $p < .001$. Through this it can be surmised that, compared with answer length, a great deal more score differentiation is coming from the extent to which individual answers match collocates found in the corpus.

Unfortunately, this fact in itself is deceiving. It is easy to assume that if a word scores few or no points, it is because that word is inappropriate – i.e. it simply

doesn't match the way native speakers use English. However, there are many more ways, other than simple inappropriateness, in which answers failed to score points. I will discuss these in more detail below.

5.2.3.1 Grammaticality

One way answers failed to scores points was grammaticality. The first prompt is an excellent example of how grammaticality played an important part in deciding whether answers matched collocates in the corpus or not. The prompt, 'Give me a…' was originally made to test specifically the idioms 'Give me a hand' or 'Give me a break', but instead it often ended up testing participants' grammatical knowledge (or lack thereof) for the article 'a'. Participants often answered with uncountable nouns such as 'money' or 'help', and answers such as these are of course not represented in the corpus and therefore scored zero. Others answered with countable nouns, yet chose nouns that began with vowels such as "apple", and as persnickety as it may seem, these words were scored zero as well.

This must be seen, to a certain degree at least, a failure on part of the OpeC. As the OpeC was created as a pragmatic test, questions such as these which focus primarily on a single grammar point should have been avoided. As the Japanese language has no article system, learners from Japan often exclude or misuse English articles. Were the OpeC to be also taken by learners whose first languages had article systems similar to English, Japanese learners would be at a disadvantage. It is not unreasonable to assume that this question was in part responsible for the disparate average scores of ESL and EFL groups.

5.2.3.2 Specificity

Participants who used specific words, especially proper nouns, were discriminated against because of the makeup of the corpus. The prompt ‘Tell me if…’ is quite open-ended. The answer ‘Tell me if he is okay.’

he(2.4481) + is(4.6622) + okay(1.3908) = 8.5011

garners over eight and a half points.

However, the answer ‘Tell me if John is well.’

Tell me if John(0) + is(4.6622) + well(0) scores less than five. The two answers can hardly be called semantically or grammatically different. Instead, one is rewarded for using general language, and the other is punished for using specific words.

Proper nouns were not the only problem. Articles proved problematic as well in this respect. For the most part, students choosing more specific answers were penalized. ‘Put it into a box’ scored about eight points, ‘Put it into the box’ scored a bit fewer points, and ‘Put it into that box’ failed to score any points at all.

5.2.3.3 Spelling

Lastly, misspellings also proved to be a problem, though a rare one. In almost all cases of misspellings, the intended word was quite clear, however in some it was not. Words that could not be recognized at all (such as ‘opapi’) were left as is. In yet other situations, words were seemingly spelled correctly, yet could be misspellings of other words as well. A common example is the answer ‘Turn on

the right.' 'Right' is a possible answer, albeit a not very good one as the phrase 'turn on' is much more frequently used to mean 'activate' than it is to mean 'rotate'. 'Light' on the other hand is an excellent answer, and given the Japanese tendency to confuse R and L, it is more than likely that the latter was intended.

For the sake of consistency, words that were not misspelled were not changed, no matter how compelling the evidence that any correctly spelled word was intended to be a different word.

5.2.4 Lessons learned through the OpeC

The OpeC was used in this experiment in order to avoid certain perceived problems with cloze tests. Specifically, the use of a corpus allowed the OpeC to avoid subjective decisions on the part of the grader while allowing test takers freedom of expression more closely mirroring natural communication. The OpeC was not perfect however. In the previous three subsections various failures of the OpeC have been detailed, but these failures are not without value. Instead, these failures can be used to shed light on inherent pragmatic assessment in general, and Oller's cloze tests in particular. Issues such as these are discussed in detail in the next section.

5.3 What these results mean for pragmatic assessment

To sum up the previous section, even though the OpeC was designed to be a test of pragmatic expectancy by specifically testing participants' predictive ability, the scores instead seem to stem from three major factors: reading comprehension, answer length, and (most importantly) agreement with the corpus.

It is reasonable to believe that something similar occurred in Oller's own cloze test data. Of course, without Oller's original data it is impossible to say for sure, but it is likely that Oller's cloze test also inadvertently ended up assessing proficiency in some of the four recognized language skills.

Almost certainly reading was assessed, as the type of cloze test used in Oller's experiments require both contextual and structural knowledge in order to logically fill in the blanks. This knowledge is gained by means of reading skill, and a lack of reading skill would necessarily preclude answering most cloze questions satisfactorily. Grammatical knowledge as well was probably necessary to answer correctly. Otherwise mistaken verb tenses and other purely grammatical mistakes would prevent correct answers.

Using different reasoning however, it is a forgone conclusion that Oller's pragmatic tests must necessarily be contaminated by inadvertently including orthodox skills proficiency into its scores. The reason this is necessarily so is because, however skillfully devised, any language test must necessarily be comprehended and completed by the participants through one or more of the four basic skills. The four basic skills are 'where the rubber meets the road' in language proficiency, and thereby whatever test is given, and no matter what it purports to be assessing, it must first either be filtered through the lenses of reading or listening to be comprehended by the test taker and/or be filtered again through the lenses of speaking or writing in order to be comprehended by the test grader. No global proficiency factor such as pragmatic expectancy can be assessed purely in and of itself without at least some influence of orthodox

language skills, as these skills are the only such places where a unitary language competency device intersects with reality.

So while the raw data in this evidence seemed at first to support Oller's claims, it in fact shows how difficult any unitary competence device is to isolate. Thus it undermines Oller's original evidence by showing a very likely possibility of how that original data could be intrinsically flawed.

5.4 What these results mean for corpus-based testing

As was suggested in Chapter 1 and in other sections throughout, the creation of the OpeC had a secondary purpose in addition to examining Oller's Pragmatic Expectancy. This was to evaluate the possible effectiveness of the OpeC as a proficiency test. In so doing it was hoped that both the viability of corpus-based testing in general, and the OpeC specifically could be explored.

To a great extent, this exploration can be carried out using the same statistical calculations that were used to analyze the OpeC's effectiveness as an assessment of pragmatic expectancy. To do this we must first assume that the GTEC is indeed a valid proficiency measuring instrument. By assuming the GTEC a valid proficiency test, and then by comparing the results of the OpeC with the results of the GTEC, this last section will investigate the use of the OpeC and/or other corpus-based tests as valid language proficiency testing instruments.

5.4.1 Thoughts on the OpeC

Considering the comparatively small amounts of time and energy used to create, take and score the OpeC, it performed admirably. The r = 0.65 level of correlation found between individual scores of the GTEC and OpeC and the p < .001 level of significance show a positive and significant correlation between the results of these two tests. Further, the group averages when compared against each other were in agreement with the hypothesized outcome. ESL high school students scored significantly higher than third year Japanese EFL students, who in turn scored significantly higher than their second year counterparts.

As was detailed in section 5.2, OpeC's inclusion of various language skills (such as reading and grammatical knowledge) was partly responsible for its correlation with GTEC scores. While these inclusions were detrimental to an uncontaminated assessment of pragmatic expectancy, they ended up facilitating the substantiation of the OpeC as a proficiency assessment tool. While far from irrefutable evidence, this experiment has shown that the OpeC can at the very least be used to as a 'quick and dirty' proficiency assessment tool.

More importantly, the OpeC has shown how corpora can be utilized as assessment tools. Given the correct framework, native-produced language usage contained within corpora can be a powerful tool to evaluate language proficiency. Given the numerous types of corpora, and considering the successes corpora have already shown in the areas of language instruction and material construction, the possibilities for corpus-based assessment are promising.

5.4.2 Thoughts on the future of PE

An important connection which was made from the results of this experiment stems from two observations which have already been discussed. The first observation is the strong and significant correlation between scores of the OpeC and those of the GTEC. The second is that the vast majority of differentiation between scores of the OpeC was attributed to predictive ability. These two observations taken together give strong evidence supporting the important idea that predictive ability and general proficiency are in some way connected.

These are perhaps not connected in the way that Oller first envisioned when he proposed PE as the one underlying global proficiency factor, but nonetheless predictive ability and general proficiency seem to be associated with each other in some fashion. Unfortunately, the basis of this connection cannot be discerned with the OpeC in its current arrangement, and an investigation of this sort is beyond the scope of this dissertation.

The fact remains, however, that the results of the OpeC linking general proficiency to predictive ability give sufficient grounds to continue the analysis of this important, and heretofore overlooked language skill. At the very least, this experiment has set the stage for future investigation into the connections between predictive ability and general language proficiency.

CHAPTER 6 CONCLUSION

As a whole, this experiment has been a success. The main goal was to attempt to reproduce Oller's original evidence supporting PE. By finding a significant correlation between the OpeC and the GTEC, this goal has been met.

By performing this experiment firsthand it was also possible to more fully understand this correlation and its origins. Despite attempts to the contrary, it seemed that the OpeC was not completely able to isolate the pragmatic expectancy of the participants. From the results it seemed clear that both reading proficiency and grammar knowledge both played significant parts in deciding the scores. It is likely, though not completely certain, that Oller's original cloze tests suffered the same outcome. In this way, what can be perceived as unintended flaws in the OpeC have nevertheless been enlightening.

Despite these 'flaws' it was also shown that most of the correlation between the GTEC and OpeC was brought about by predictive skill assessment. This shows that ability to anticipate what words or phrases are likely to occur next (call this ability Pragmatic Expectancy, predictive skill, collocative ability or what have you) is closely related to general language proficiency. So while Oller's call to uproot discrete point language assessment in favor of pragmatic testing may have been premature, his underlying premise that language proficiency and pragmatic expectancy are connected seems to have merit.

Lastly, by showing a positive correlation to an orthodox general proficiency test, the OpeC has demonstrated the as of yet untapped potential of corpus-based

testing. It seems only logical that corpora, which have proven invaluable as research resources and teaching tools for decades, can aid assessment as well. The impartiality and efficiency achieved thanks to the use of corpora all but ensure their ever-increasing presence in the future of language assessment.

APPENDIX 1　THE OPEC

Directions: Please finish the sentences.

(次の表現ではじめて、英文を完成しなさい。)

1) Give me a_____

2) Are they_____

3) I think we_____

4) I get the_____

5) Until it_____

6) It happened_____

7) When you get_____

8) That's what I_____

9) I went to the_____

10) Put it into_____

11) It was found_____

12) Write down_____

13) He will be_____

14) Tell me if_____

15) Is there any_____

16) If you get_____

17) I enjoy_____

18) Turn on the_____

19) Anyone can_____

20) I am going to_____

21) What do you think_____

22) I need to_____

23) That sounds_____

24) She makes_____

25) Let's not_____

26) You look like_____

27) How about_____

28) Keep it_____

29) I can't remember_____

30) Ask her_____

Notes for proctors:

Students should be instructed as written below before the test is handed out:

- There are many acceptable answers. Please write only one.
- Students should not think too much about the right answer
- Just write the first words that come to mind
- Spelling is not an issue as long as the word intended is clear

監督先生方へ

ここに書いたとおり、プリントを配る前に受験生に指導よろしく

ー正しい答がいくつもあります。ひとつだけ書きなさい。

ー受験生はあんまり考えて、悩まないほうがいい。

ー思いついたものを書きなさい。

ーつづりあんまりこだわらなくても良い。意味が分かればよい。

REFERENCES

Aitchison, J. (2003). Words in the Mind. Oxford: Blackwell.

Baker, C. & N. H. Hornberget. (Eds.). (2001) An Introductory Reader to the Writings of Jim Cummins. Clevedon: Multilingual Matters Ltd.

Benesse (2008) Global Test of English Communication GTEC for STUDENTS. Benesse. <http://gtec.for-students.jp/product/product.htm>

Calfee, R. C & S. W. Freedman. 'Understanding and Comprehending'. (1984) Written Communication, Vol. 1, No. 4, 459-490.

Chomsky N. (1965). Aspects of the Theory of Syntax. Cambridge: MIT.

Cummins, J. (1980). 'The entry and exit fallacy in bilingual education'. NABE: The Journal for the National Association for Bilingual Education, 4 (3), 25-29.

Cummins J. (2001) Language, Power and Pedagogy: Bilingual Children in the Crossfire. Multilingual Matters Ltd.

Darnell, D. K. (1968). 'The development of an English language proficiency test of foreign students using a clozenthropy procedure'. Boulder, CO: Department of Speech and Drama. University of Colorado.

Dewey, J. (1910). How We Think. Lexington: D.C. Heath.

Ellis, R. (2003) Task-based Language Learning and Teaching. Oxford: OUP.

Krashen, S. D. (1981) Second Language Acquisition and Second Language Learning. Oxford: Pergamon Press.

McNamara, T. (2000). Language testing. Oxford: OUP.

Meara, P. M. (1983) 'Word associations in a foreign language'. Nottingham Linguistics Circular 11 (1983), 28-38.

Oller, J.W., Jr. (1970). 'Dictation as a device for testing foreign language proficiency'. English Language Teaching 25 (1971), 254-259.

Oller, J.W.,Jr. and Conrad, C. (1971). 'The cloze procedure and ESL proficiency'. Language Learning 21, 183-196.

Oller, J.W., Jr. (1972). 'Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language'. Modern Language Journal 56, 151-158.

Oller, J.W., Jr. and V. Streiff (1975). 'Dictation: A Test of Grammar-Based Expectancies'. ELT Journal 1975 XXX(1), 25-36

Oller, J.W., Jr. and Perkins, K, (Eds.), (1978). Language in Education: Testing the Tests. Rowley, MA: Newbury House.

Oller, J.W., Jr. (1979). Language tests at school. London: Longman.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: OUP.

Valette, R. (1964). 'The use of dictee in the French language classroom'. Modern Language Journal 39, 431-434.

Vollmer, H.J. and Sang, F. (1983). 'Competing hypotheses about second language ability: a plea for caution'. In J.W. Oller, Jr., (Ed.), Issues in Language Testing Research (pp. 29-79). Rowley, MA: Newbury House.