1-1-2010

Teacher Evaluation of Item Formats for an English Language Proficiency Assessment

Jose Luis Perea-Hernandez
Portland State University

Let us know how access to this document benefits you.
Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Recommended Citation

10.15760/etd.436

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Teacher Evaluation of Item Formats for an
English Language Proficiency
Assessment

by
José Luis Perea-Hernández

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Arts
in
Teaching English to Speakers of Other Languages

Thesis Committee:
Kimberley Brown, Chair
Lynn Santelmann
Dannelle Stevens

Portland State University
©2010
ABSTRACT

Language testing plays a significant role in the collection of systematic information about English Language Learners’ ability and progress in K-12 public schools in Oregon. Therefore, the following thesis reports on a process of the development of this assessment as I was at some point one of the members in the test development team and then drew upon teachers’ opinions and judgments to compensate for a gap in this process.

For example, despite the fact that Oregon item-writers produced test items in summer 2004, they identified some deficiencies as they followed the process. Therefore, the need for more examples of item formats in order to develop better items, a proper identification of materials for reading and listening, and mainly the lack of detailed item test specifications was imminent.

Unfortunately, item writers did not possess a guiding document that would allow them to generate items. The item writing process was not ‘spec-driven’. Hence, these drawbacks in this test (larger project), provide the basis and rationale for this thesis. Thus, the aim of the present study was to identify what item formats worked best based on professional teacher judgment. Then, once identified these were used to detect what items could potentially be reverse-engineered into specifications for future studies or developmental stages of this assessment.

Therefore, the purpose of this thesis was to analyze through professional teacher-judgment the suitability of item formats, test items and the construct in a way that would be helpful for future item-writing sessions, but most importantly, to the better understanding of teachers devising items for this test. Teachers determined what items and item formats were more suitable in a testing situation for the diverse English language learners according to their relationship to the construct of the test, practicality, cognitive level, appropriateness, and similar tasks as used in classroom instruction.
Finally, the outcomes based on teacher feedback allowed the required data to establish a process to reverse-engineer test items and item formats into item test specifications. Thus, based on this feedback a more systematic process was illustrated to demonstrate how item formats can be the basis to reverse-engineer item test specifications.
Dedication

Having completed a project such as this has never been more challenging, yet rewarding in my academic life.

Thus, the energy, fulfillment and care that went into the completion of this study is dedicated first of all to my Lord and Savior Jesus, who gave me the strength to keep going in the most difficult time in my life, while also trying to complete this project.

Second of all, a dedication to my son Mateo as this is the best present I can give.

Lastly, to my immediate family (my mom, brother and sister) and my Professors for the continuing support, and to all those who always believed...

Thank you.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii
LIST OF FIGURES ............................................................ ix

CHAPTER 1: INTRODUCTION TO THE STUDY ........................................... 1
Project Background ............................................................... 6
Glossary .......................................................... 14

CHAPTER 2: LITERATURE REVIEW ......................................................... 17
Evaluation, Assessment & Testing ................................................ 17
Measurement & Types of Language Testing/Assessment .................... 19
NCLB, AMAO, Title I & Title III Mandates in Relation to Oregon Education for ELLs .............................................................. 21
NCLB and its Repercussion on Large Scale Assessment ................. 26
Ethical Considerations for Large Scale Assessment Including Language Assessment .................................................... 28
Assessment/Testing for Language Minority Students ....................... 32
Test Construction/Development & Test Specifications .................... 33
Reverse-Engineering .............................................................. 36
Conclusion ..................................................................... 40
Research questions .............................................................. 41

CHAPTER 3: METHODOLOGY ............................................................... 42
Introduction .................................................................. 42
Research Design & Paradigm ................................................ 43
Setting ......................................................................... 46
Participants ...................................................................... 47
Teacher Participants (for Item Format Evaluation) ......................... 47
Sampling ......................................................................... 51
Data Gathering Instruments and Procedures ................................ 56
Procedures ...................................................................... 56
Instruments ...................................................................... 59
1st Instrument: Constructed Response Method Questionnaire in the Evaluation of item formats vis-à-vis Individual Test Items ......................... 61
2nd Instrument: Likert Scale Method Questionnaire in the Evaluation of Item Formats vis-à-vis Individual Test Items ........................................ 62
Writing ........................................................................ 66
Data Analysis Procedures .................................................. 68
Establishing Validity for Research Findings ................................ 71
CHAPTER 4: RESULTS

Quantitative Results (Yes/No Questionnaire Open-Ended Results)

Item Format Relationship with the Construct

Writing Item Formats

Item Format Difficulty for Grade Level (IF)

Item Format & Test Item Relationship (IF/TI↔Relationship)

Item Format Comprehensibility/Generative to Similar Test Items (IF↔Comprehensibility)

Item Format Appropriateness to ESL Students (IF↔ESL Sts)

Item Format & Test Appropriateness to Proficiency Level (IF↔Prof Lvl)

Task from Item Format & Test Item vs. Regular Instructional Tasks (IF/TI↔ClassTask)

Need to Pre-teach/Train Students on Selected Task/Item Format (IF/TI↔PreTeach)

Need to Pre-teach/Train Students Test Taking Procedure on Selected Task/Item Format (IF↔PreTeach Proc)

Grade Groups Most Appropriate per Item Format

Quantitative Results (Inferential Statistics)

Reliability of the Likert-Scale Questionnaire

Reliability of the Writing Likert Scale

Likert Scale Results for the Relationship between the Test and the Construct

Likert Scale Results for the Remaining Variables in the Questionnaire

Qualitative Results

Final Writing Data Display

Conclusion

CHAPTER 5: DISCUSSION

Putting it All Together (Quantitative and Qualitative Data Results)

Addressing the Research Questions

Writing Modality

The Construct of Grammar in Writing

The Construct of Vocabulary in Writing

The Construct of Morphology in Writing

The Construct of Representational Language Functions in Writing

The Construct of Manipulative Language Functions in Writing

The Construct of Heuristic Language Functions in Writing

The Construct of Imaginative Language Functions in Writing

The Construct of Cohesion & Coherence in Writing

The Remaining Variables (Discussion)

Research Questions Addressed

Grade Groups Most Appropriate for All Item Formats

Item Format Difficulty for Grade Level (IF)

Item Format & Test Item Relationship (IF/TI↔Relationship)
LIST OF TABLES

Table 1. Takala’s (1998) types of testing/assessment........................................20
Table 2. Hughes’ Procedures for Test Construction........................................33
Table 3. Hughes framework for crafting the content of a test..............................34
Table 4. Demographics of All the Participants in All Modalities
    Based On the Teachers’ Background Checklist........................................50
Table 5. Target Distribution of Participants per Modality and Grade Level..........58
Table 6. Final distribution of participants per modality....................................59
Table 7. Proposed Coding System for Qualitative Data Analysis.........................69
Table 8. Coding Format for each of the Variables found
    in the Open Ended Questionnaire.........................................................90
Table 9. ELP Test Item Formats (4th Draft) after facilitators’ consensus
    input, item analysis and categorization. Perea (2004)...............................141
Table 10. Criteria Extracted from Test Items in order to RE Specifications..........142
Table 11. Item Format Combination in the Modality of Writing........................144
Table 12. Item Format Combination in the Modality of Reading........................145
Table 13. Item Format Combination in the Modality of Listening.......................145
Table 14. Item Format Combination in the Modality of Speaking.......................146
LIST OF FIGURES

Figure 1. Bachman’s model of language competence (1990) .............................................2
Figure 2. Overall scope of the thesis project .....................................................................13
Figure 3. The role of test specifications in stages of test development, from Davidson and Lynch (2002) .................................................................35
Figure 4. Mixed Research Methods ................................................................................57
Figure 5. Metaphors to item writing process ..................................................................136
Figure 6. Spec Reverse-Engineering Process for an English Language Proficiency Assessment ............................................................153
Figure 7. Specplate (Item Format Blue Print) – Combination of Item Formats 161
Figure 8. Specplate (Item Format Blue Print) – Selection of Best Item Formats based on Teacher Input .................................................................162
CHAPTER 1
INTRODUCTION TO THE STUDY

Due to the policies of the No Child Left Behind Act and the AMAO Mandate (Annual Measurable Achievement Objectives), states throughout the U.S. need to comply for their ELL (English Language Learners) population in terms of academic achievement. Hence, language testing plays an important role in terms of reporting English Language Proficiency for English Language Learners. No Child Left Behind (NCLB) requires schools to annually measure and report acquisition by all English Language Learners in the public school system (K-12). States must also develop English Language Proficiency (ELP) standards that are linked to their English Language Arts Content Standards. At this time, standards and assessments are being developed throughout states in four domains, listening, speaking, reading and writing and will report on student progress in five domains with comprehension as the fifth domain. Due to this, standards and assessment must be aligned.

Currently, the state of Oregon is in the process of developing its own English Language Proficiency (ELP) Test with the Oregon Department of Education (ODE) as principal organization to devise this instrument. Such a measurement instrument is intended to be linked to the Oregon ELA (English Language Arts) standards as well as to be aligned to the Oregon ELP (English Language Proficiency) standards. The construct for the Oregon ELP test is English Language Proficiency. Overall, language proficiency (also called communicative proficiency) is defined as: "Language use, or performance. It is the ability to communicate, to interact, to express, to interpret and
to negotiate meaning, and to create discourse in a variety of social contexts and situations." (Pawlikowska-Smith, 2002, p. 6).

Many scholars such as Bachman (1990) and Savignon (1972) postulate that language proficiency is a skill in communication that should be measured without reference to a particular curriculum (as cited in Pawlikowska-Smith, 2002, p. 6). According to them it includes communicative competence & language ability (organizational and pragmatic language competence, or what the learner knows, and strategic competence activated in a situation of language use), and performance, which is observable and measurable. The five components of language proficiency (as part of the Oregon ELP Test construct) are: linguistic competence, textual competence, functional competence, socio-cultural competence, & strategic competence (Pawlikowska-Smith, 2002; Scott & Nave, 2004).

Figure 1: Bachman's Model of Language Competence (1990)
Linguistic competence includes grammatical competence that has been defined by Canale and Swain (1980) as a type of competence that is understood to include knowledge of lexical items and rules of morphology, syntax, semantics, and phonology. Textual competence refers to knowledge to achieve textual unity (coherence and cohesion). Cohesion refers to how utterances are linked structurally to facilitate the interpretation of the text. Coherence concerns the relationship among the different meanings in a text. (Canale & Swain, 1980; Halliday, & Hassan, 1976; Celce-Murcia, Dörnyei, & Thurrell, 1995, p .13). Textual competence is also referred as discourse competence, which is the knowledge of using grammatical forms and meanings to create harmonious spoken or written texts in different genres such as narrative, argumentative essay, scientific report or business letter. Additionally, functional competence is the knowledge and ability to understand what the real intent of the speaker or written text is, beyond the literal meaning of words. It is what Bachman (1990) calls "illocutionary competence." It is what "enables us to interpret relationships between utterances or sentences and texts and the intentions of language users." (Bachman & Palmer, 1996, p. 69).

The Oregon ELP Test construct is based on functional competence and uses a taxonomy of macro-functions by Bachman (1990, pp. 92-94) and, Dutro and Moran (2003). These functions are classified as follow: ideational or representational, manipulative, heuristic and imaginative functions. Therefore, representational functions are used as mean to convey information and to present, describe and share our experience of reality. Manipulative functions are devised to get things done with
language as they are divided into instrumental, regulatory and interpersonal. Heuristic language functions allow learning things about language, to extend our knowledge, to learn, to problem-solve. Finally, imaginative functions are those linked to the enjoyment of language, for instance metaphors, jokes, etc.

Socio-cultural competence is the component made up of two sets of rules: socio-cultural rules of use & appropriateness and rules of discourse. Socio-cultural competence can also include what is called strategic competence; this component consists of verbal and nonverbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence (Canale & Swain, 1980, p. 29).

As part of the ELP Test construct, language forms are also critical since they represent the parts of words called morphology. Forms (syntactic and morphological components of language) are the anatomy of a language since they deal with the internal grammatical structure of words (Viking Coe, U of Houston, 1994).

45 Oregon ESL teachers served as item writers and produced over 3000 items for this test. After that, this process allowed to define, expose, and operationalize the components of the ELP Test construct. Test items were created for four domains: speaking, reading, listening, and writing across grade bands: K-1, 2-3, 4-5, 6-8 and 9-12 and taking into consideration the following proficiency levels: beginning, early intermediate, intermediate, early advanced, advanced, and proficient. Such proficiency levels are correlated to the Oregon ELP Standards.

In order to produce accurate test items or tasks, item writers need to devise
those from test specifications. Davidson and Lynch (2002) postulate that a well-written test specification should generate many equivalent test tasks (p. 4). In addition, they claim that a clear "spec" should mirror the purpose of the test, as they point out this emphasis must be in test purpose and test content. Since the ELP Test is based on an IRT (Item Response Theory) scale, and it measures language proficiency, specifications should convey the notion of ‘absolute decisions’ and place the creation of the test at the CRM (Criterion Reference Measurement) end of the continuum. Thus, the instrument should be devised to make decisions about the test-takers in relation to a specific behavioral domain or criterion (Davidson, 2002).

Given the fact that ODE did not possess detailed specs at the time of item writing, this study will allow the researcher to conduct a more teacher-oriented analysis of item formats, test items and their suitability to the test construct to obtain more detailed information on their structure and tasks presented. Eventually, these results will establish a better framework to reverse-engineer item specifications in future developmental stages of this test. According to Davidson (2003), "Reverse engineering refers to the production of test specification when none exist: all we have are the test items or tasks, and from those we induce what the blueprint might be" (Davidson, 2003, p.1).

Due to ODE’s last experience on item writing, teachers writing items needed to have explicit examples of each target item and well as congruent instructions on how to create the items. Despite the fact that item writers (teachers) received training on the test construct, the ELP standards and the projected proficiency levels (as
represented in the standards) yet specifications were not provided. Test specifications
are designed to provide this type of guidance.

**Project Background**

To describe the background from the overarching project that lead into the proposal for this thesis project, I will explain my role in this larger project. First, my role in the completion of the ELP Test project evolved as I started in the Northwest Regional Educational Laboratory (NWREL) as an intern in order to do research on language testing.

I was part of a collaborative process among the assessment team members as this larger project has evolved across time. In order to start thinking about the process of constructing the test, team members did not start first crafting the test prior to having a strong theoretical foundation; instead both linguistic and testing perspectives were approached. Thus, the first step consisted of a careful revision of what constitutes the testing of English language proficiency for second language learners in Oregon public schools at K-12 levels. After having identified the construct of the test, the next step was to align the K-12 Oregon English Language Proficiency Standards to the construct of the test. Then, the following step consisted of operationalizing the construct in order to device a testing framework.

At that time, the Oregon Department of Education (ODE) lacked enough representative item formats in order to start thinking about what item formats may have been appropriate and what was available for second language testing. After
having compiled a taxonomy of item formats and submitted those to ODE, the next step was to work on the alignment of the State Standards and the ELP Test Construct.

After having a strong mind set in terms of the projected types of tasks and items the team wanted, the next step was to conduct a session of item writing and train over 45 ESL K-12 teachers in general terms of second language acquisition, communicative competence (to define language proficiency vis-à-vis the test construct) and to provide a general framework in terms of the testing instrument we were intended to pursue.

Basically, the construct reflects the components of communicative competence including linguistic, textual, grammatical, and functional competence. Under functional competence, the language functions represented are representational or ideational, heuristic, imaginative, and manipulative (Halliday, 1976, 1985a, 1985b, 1985), which were taken as reference from the Theoretical Framework of the Canadian Language Benchmarks (Pawlikowska-Smith, 2002) and Scarcella’s conceptual framework of academic English (2003). At that stage of the process, the Oregon Department of Education identified a taxonomy of language functions presented by Dutro and Moran (2003) as highly important in a classroom setting.

Considering the construct of the test, language functions and forms are also critical in the operationalization of the test. Hargett (2003) declares that LEP students need to be able to demonstrate English proficiency to communicate concepts by means of using rules of syntax to vocabulary unique to English, which represent these concepts embedded through phonological systems or conventions of writing.
After these functions were identified the assessment team agreed that ELL students needed to perform all of these 23 linguistic functions (Dutro & Moran, 2003) in order to be competent in an academic setting.

As a consequence, once the macro-functions (representational, imaginative, heuristic, & manipulative) were clearly identified, the next step was to be able to align them with the Oregon ELP Standards since the law requires that the ELP Test be aligned with the pertinent standards form each state (Hargett, 2005; Oregon English Language Proficiency Standards, 2004). Then, a panel of teachers (content panel) was selected in order to identify those standards that could be assessed in a testing situation. The panel looked at over 1100 standards originally coming from the Oregon Language Arts framework and then selected those standards for assessability purposes. Interestingly at that point in the process, the standards analyzed by the panel were coded under Functional Competence by using only the 23 Representational Language Functions as devised by Dutro and Moran (2003).

Once this process was completed, a more detailed coding took place for these standards by Perea and Zaheed (2004) at NWREL. The coding format included the macro-functions mentioned above and even going through a more detailed layer in terms of competence areas. Thus, grammatical competence included phonology, morphology, lexicon, semantics, and syntax. For textual competence, cohesion and coherence were indicated and for functional competence, manipulative, heuristic, imaginative and representational functions were identified.

In this stage of the project, the purpose of the coding was to align the
competence areas to the state standards. In addition, this coding was used as reference to create a process of sampling in order to select standards for item writing. This system of sampling was intended to reflect the totality of the 1100 Oregon ELP standards.

This process reflected three stages: a) a process of elimination, b) a process of categorization, and c) a process of consolidation. For the process of elimination, 1100 standards were reduced to 476, so standards deemed un-assessable by the content panel and double-checked by independent consultant were eliminated. In addition, those standards that were part of English Language Arts Content Areas and too broad were also discarded. For the process of categorization, the same categories used for Grammatical, Textual and Functional Competence were used.

Thus, in order to make sense of this process of consolidation, some standards across domains (listening, speaking, reading, & writing) overlapped among grade groups expressed as rationales. Consequently, in order to make item writing more tangible and convenient for item writers and educators, some standards reflected ‘unique’ academic concepts and rationales that could happen in one grade only but more than one Common Curriculum Goal (CCG) as observed in the standards. In the same way ‘consolidated’ standards that represented a specific concept/rationale covering many grades and CCGs were collapsed. Again, the purpose for doing this collapsing was to allow a more manageable use of standards for the teachers writing items. Due to the complex process mentioned above, we will define both of these standards (ELA vs. ELP) in their different roles since the ELP standards are a by
product of the ELA standards as:

English-Language Arts (ELA) Content Standards define what all Oregon students, including students who are speakers of a language other than English, are expected to know and be able to do. The ELP Standards are designed to supplement the ELA standards to ensure that LEP students develop proficiency in both the English language and the concepts and skills contained in the ELA standards (2004, Oregon Department of Education, pp. 2).

As it is observed in the previous definition of the content standards, it is also critical to reference that the resulting assessable ELP standards are only to be used for assessment purposes. This is referenced in the official standards since they will be used to develop the Oregon English Language proficiency Tests.

Finally, the item writing session was held in summer 2004 with a combination of members at the NWREL, ODE and NWEA (North West Evaluation Association). 45 teacher-item writers were recruited in order to develop test items in the following categories or Item Types: (MCh) Multiple Choice, 1 point; (SA1) Short Answer, 1 point; (SA2) Short Answer, 2 points; (ER) Extended Response, 4 points.

In terms of the training provided to the 45 item writers I served as a facilitator in order to help teachers in their item writing process, I provided training on the standards alignment and revision of items as they were developed. Overall, teachers were delivered a general description of the ELP test construct, standards alignment, language functions, developmental grade grouping, subjects to use (Math, SS. etc.), social language in academic setting, ELP standards consolidated, ELP item plan (blue
print), ELP item types and some formats, ELP proficiency level criteria, a language forms overview and some examples.

To conclude this introduction, the process in which I was myself involved allowed me to fully understand the different stages in the development of the Oregon ELP test. Having experienced the process of standard consolidation into assessable standards all the way to item writing, was essential to identify the gap that will be the focal point of this thesis.

Regarding the scope of this thesis, the initial focus of this study included all the modalities of the Oregon ELPA (reading, writing, speaking and listening), which indeed the researcher was able to collect both quantitative and qualitative data results. Nevertheless, due to space restrictions and the overwhelming amount of information, I will only be reporting on a subsection of the results. Therefore, the results on the modality of writing will be brought to discussion on the results and conclusions chapters of this thesis.

In order to allow the reader realize of the larger scope of this project and to provide some context to see where this piece of the pay fits (writing modality), the following figure 2 illustrates the entire process and results achieved from the original intentions of this thesis. I will also make reference to some of the results collected from the teacher background checklist used in the present study to collect data. Such results will be brought to light to make an argument on the importance of teachers in large scale assessment development.
Concurrently, in the conclusions chapter I emphasize on possible applications to other test development situations based on data from the other modalities (listening, reading and speaking) and from my own experiential situations.
Figure 2. Overall Scope of Thesis Project

- **Positivist Paradigm**
  - True exploration of test spec design and teacher reaction

- **Mixed Research Method**
  - Connected contributions approach
  - QUANT → Qualit → QUANT → Qualit → QUANT
  - Pilot survey Teachers
  - Redefine Pilot survey Teachers
  - Revised survey (item formats)
  - Analyze data from survey (data displays) & (statistical procedures)

- **Target Distribution of Participants**
  - Modalities: reading, writing, speaking, listening
  - Grade groups: K-1, 2-3, 4-5, 6-8, 9-12 = 30 teachers per modality

- **Item Formats for Evaluation**
  - Listening - 8 item formats & sample test items for evaluation
  - Speaking - 10 item formats & sample test items for evaluation
  - Reading - 6 item formats & sample test items for evaluation
  - Writing - 12 item formats & sample test items for evaluation
  - Total of 36 item formats evaluated

- **Total participants**
  - 27 participants evaluated listening (0.98 reliability score in Likert scale)
  - 21 evaluated reading (0.91 reliability score in Likert scale)
  - 26 evaluated speaking (0.95 reliability score in Likert scale)
  - 44 evaluated writing (0.92 for set of 26 participants and 0.93 for set of 28 participants in Likert scale)

- **Data gathering**
  - Quantitative (mean & reliability scores) on a Likert scale
  - Qualitative (data displays and coding format) on an open ended questionnaire
  - Both Likert Scale and Open Ended Questionnaire with 16-20 variables including construct of test & practicality for classroom use and ESL appropriateness

- **Reverse-engineering process**
  - Re-definition of Reverse-engineering process (missing gap)
  - Process to RE item formats into item test specifications
  - Creation of Specplates (specification of specifications) & RE item test specification

- **Teachers**
  - Teachers' participation in current study
  - Teacher input to inform suitability of item formats to RE specs
  - Teachers as decision makers
  - Teachers' participation in assessment panels
Glossary

**Connected Contributions Approach:** This refers to the use of the strengths of a research design method to contribute to the different strengths of another. For instance, using a core method as quantitative (QUANT) and then redefining key areas or addressing gaps through qualitative methodologies (qual) (Morgan, 2005).

**Constructed-Response Items:** Constructed-response items require students to create their own answers rather than select from prewritten options. These items generally allow for various ways in which they can be answered correctly. Sometimes, however, they may have only one or two specific answers. Constructed-response items allow the awarding of partial credit for partial knowledge or partial completion of the task (Pearson ELL Item Writing & Reviewing Guidelines, 2009, p. 3).

**Cut Score:** Selection of test items that belong to specific standards (in the case of ELL assessments, the notion of language proficiency levels). In standard setting committees for ELL assessments, selected items may constitute a ‘cut score’ between one proficiency level to the next one.

**Item:** An assessment question is often referred to as an item (Pearson ELL Item Writing & Reviewing Guidelines, 2009, p. 1).

**Item Format:** Format refers to both the type of item and the item’s layout and appearance on the page of a test form. For example, item format encompasses stimulus art and how it is presented and labeled, the use of tables, charts, and graphs,
the orientation of option art or formatting of text options in a multiple-choice question, etc. In addition, item format contains an intrinsic relationship to a set of skills or criterions tested depending on their design (Adapted from Pearson ELL Item Writing & Reviewing Guidelines, 2009, p. 1).

**Item Type:** At the broadest level, items are considered to be either selected response, in which the student selects a response from provided possible answers, or constructed response, in which the student constructs his or her own response. Multiple-choice items are the most common type of selected response item, though True/False and Matching are also selected response assessment item types. Short answer (or short response) and extended response items are the two most common types of constructed response item. A labeling task and a performance-based task are other types of constructed response assessment items. Typically, the notion of item types contains implications for psychometrics in terms of scoring (e.g. MCh items contain one total score point, SA [short answer] items can range from zero to one or two score points, and ER [Extended Response] items can range from zero up to four points or more, depending on the test design) (Adapted from Pearson ELL Item Writing & Reviewing Guidelines, 2009 p. 1).

**Prompt:** The prompt is the part of a constructed-response item that asks the student to do something (describe, discuss, explain, compare, contrast, persuade) (Pearson ELL Item Writing & Reviewing Guidelines, 2009, p. 3).

**Reverse-Engineering:** The creation of test spec from representative test items/tasks (Davidson & Lynch, 2002, p. 41).
**Rubric:** A rubric (also known as scoring guide and rating scale) is a set of guidelines for scoring constructed-response items. The rubric sets forth the expectations for performance and delineates what a response must be in order to earn each of the possible score points (Adapted from Pearson ELL Item Writing & Reviewing Guidelines, 2009, p. 3).

**Specplate:** A combination of the words “specification” and “template.” It is a model for a specification, a generative blueprint which itself produces a blueprint. It can be also referred to as a specification of specifications (Davidson & Lynch, 2002, p. 50).

**Stimulus:** Material associated with one or several test questions that sets the context for and/or provides information related to the task(s) presented by the test question(s). The stimulus material precedes the associated question(s), and the student must process the material in order to respond to the question(s). Examples of stimuli include a reading passage, listening passage, drawing, photograph, graph, chart, or other representation of information. (Pearson ELL Item Writing & Reviewing Guidelines, 2009, p. 1)

**Washback:** also known as “backwash”, merely referring to the consequences of the test, and the impact on the people who use it (Davidson & Lynch, 2002, p. 1).
CHAPTER 2
LITERATURE REVIEW

The following section covers the most important literature concerning the main topics driving the purposes on this thesis: language assessment, types of language testing, policies affecting state testing, testing for language minority students, test construction (including test specifications), and reverse-engineering. The areas presented in this literature review are the backbone of the study where the final objective is to understand how teachers’ voices can be of relevancy in the development of large scale assessment. The core aspect of the study lies onto determining what aspects of the test in question can be better defined in terms of producing item test specifications based on teacher input (spec reverse-engineering).

Evaluation, Assessment & Testing

In order to define properly the overarching notions of testing and assessment, it is critical to differentiate the functions of each as sometimes they tend to be confused with evaluation as well. Evaluation is defined as the systematic gathering of information to make decisions about language teaching programs. Similarly, Takala (1998) postulates that evaluation is an activity whose purpose is to determine the quality, activities or performances of educational programs. Takala claims that curricula and syllabi normally function as the criteria and there should be a link between objectives and evaluation. Hence, evaluation can make use of assessment
and test instruments beyond evaluation purposes (Lynch, 1996; Bachman, 1990; Turner, 1991). To the contrary, assessment is the ‘systematic use of information to make decisions about individuals and their language ability’. Assessment addresses other skills and purposes such as proficiency, diagnosis of needs, achievement of syllabus objectives, and the performance of specific tasks (Lynch, 1996). Lynch also declares that assessment allows for the opportunity for reflection on such results and who to take appropriate action on the outcomes. Testing is devised to determine measurement procedures to inquire on aspects of individual knowledge and ability.

McNamara (2000) points out that not all language tests are of the same type as they vary according to their design and purpose. In other words, there is a relationship between test method and test purpose. Additionally, tests are classified due to the decisions they reveal from the test results (Bachman, 1990; Davidson & Lynch, 2002). Additionally, Takala (1998) presents a series of procedures and questions that need to be devised while using tests. Therefore, the assessment procedures should be open and comprehensible (transparent), internally consistent, and could be related to other assessment systems. Takala (1998) describes that all assessment needs to be informed by and based on answers to questions such as “What information (Why test? When? What for?); how to organize and report information (What for? Who for?); what to test? (What model of L2 competence? Content? Sampling of content/students?); and how to evaluate performance? (Count or judge?)”. (p. 4).

This view of language assessment leads to the notion of the language test construct. McNamara (2000) defines this term as those aspects of knowledge or skills
possessed by the test-taker that are measured. This notion of construct has critical implications in the design of the test since it needs to be operationalized so test items can be created. Defining the construct means making clear what knowledge of language we are measuring, as well as how such knowledge is translated into actual performance or language use.

**Measurement & Types of Language Testing/Assessment**

McNamara (2000) and, Davidson and Lynch (2002) present two main approaches to testing according to the measurement assumptions they reflect: Norm-reference measurement (NRM) and criterion-reference measurement (CRM). NRM represents the comparison between individuals against each other as represented in test scores (norm). In other words, how a specific test-taker is ranked-ordered to a single score or other test takers. Lynch declares that this approach is relative, as it needs to be in relation to some sort of ability or knowledge. This mastery is also known as standard setting. Given this, NRM approaches assume that test scores will be normally distributed across the population of interest. In contrast to this notion, CRM mirrors individual performances at any given level. Thus, individual test-takers reach their own performance goals at their own rate. Lynch (2002) asserts that CRM allows for verifying whether or not an individual has mastered the ‘criterion’ as opposed to comparing him/her to other test-takers, the criterion is a skill, ability or knowledge.

As previously stated, language tests are classified according to their purposes
or decisions made on test scores. Takala (1998), presents several types of testing/assessment that have been developed over the years. Table 1 depicts this taxonomy.

Table 1. Takala’s Types of Testing/Assessment (1998)

<table>
<thead>
<tr>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norm-reference testing vs. criterion-reference testing</td>
</tr>
<tr>
<td>Achievement testing vs. proficiency testing</td>
</tr>
<tr>
<td>Diagnostic testing vs. formative testing vs. summative testing</td>
</tr>
<tr>
<td>Standardized tests vs. teacher-made tests</td>
</tr>
<tr>
<td>External vs. internal testing/assessment</td>
</tr>
<tr>
<td>Self assessment, peer assessment, teacher-assessment, external assessment</td>
</tr>
<tr>
<td>High-stakes vs. low-stakes assessment</td>
</tr>
<tr>
<td>Tests, examinations vs. national assessments (representative samples)</td>
</tr>
</tbody>
</table>

Using Takala’s (1998) classification as a reference, I will start describing only those assessments/tests significant to the scope of this study. NR and CR tests are defined under the two main approaches postulated above. McNamara (2000, p. 5) also makes an additional differentiation between paper-and-pencil language tests from performance tests. Paper and pencil tests reflect the familiar examination question paper format as they are used to assess separate/discrete components of language; let it be grammar, vocabulary, syntax, etc. or repetitive understanding, for instance, listening and reading compression. They are usually created on a fixed response format, in which test items are scored automatically. On the other hand, performance language tests reflect language skills since they are assessed in an act of communication. The typical domains assessed are speaking and writing and judged by raters on a rating scale system.

The best well-known distinction in terms of test purpose is achievement and proficiency tests. Thus, achievement tests are interrelated with the process of the
instruction or how much knowledge and progress students have achieved over a
completion of a course or program. Proficiency tests investigate the future language
situations in language use by not looking at any prior instruction (Bachman, 1990;
Bailey, 1998; McNamara, 2000; Davidson & Lynch, 2002; Murray, 2001).
McCutchen (2003) points out the advantages of proficiency tests since they allow
educators to gather data on pupils’ performance, which leads to make decisions about
specific mastery of language form, weakness in language knowledge, competence or
ability. Moreover, Cohen (1994) and Lynch (2003) point out that proficiency testing
correlates with administrative purposes and achievement testing matches instructional
purposes.

**NCLB, AMAO, Title I & Title III Mandates in Relation to Oregon Education for ELLs**

The overarching federal mandate stated in 2001, is the reason why the Oregon
ELP test is being crafted. The USDE (U.S. Department of Education) describes the
legislation as follows: The No Child Left Behind Act (NCLB) is a landmark in
education reform designed to improve student achievement and close achievement
gaps (No Child Left Behind at Glance, 2004 p. 1)

According to No Child Left Behind at a Glance (2004), such a mandate,
reflects accountability for results, an emphasis on executing what works based on
scientific research, expanded parental options, and expanded local control and
flexibility. Such accountability mirrors the projected goal of having every child make
the grade on state-defined education standards by the end of the 2013-14 school year. To reach the goal, states have developed benchmarks to measure progress and corroborate for students’ learning. States are required to separate or disaggregate data from students’ achievement (in this case test scores from an English Language Proficiency Assessment), holding schools accountable for subgroups of students. The USDE reports that a school group or school district that does not meet the state’s definition of ‘adequate yearly progress’ (AYP) for two consecutive years (schoolwide or any subgroup) is considered to be ‘in need of improvement’ (p. 1). Further, the umbrellas under which NCLB is implemented are Title I & III.

Title I deals with subject matter instruction by mandating inclusion of ELLs in math, language arts and science. Whereas Title III provides accountability for the effective teaching of English. In addition, it requires states to develop standards and assessments of English language proficiency and to report test results for two indicators: the percent of students transitioning from English proficiency programs, and the percent of students making progress in English.

Under NCLB, states are required to establish three kinds of standards and two types of Annual Measurable Achievement Objectives (AMAO). According to the US Department of Education’s non-regulatory guidance on the title III state formula grant program report, these standards and objectives are different under Title III and I, but are interrelated in establishing State achievement targets and accountability systems. Therefore, the three standards and the two sets of annual measurable achievement objectives are: a) English Language Proficiency Standards, Title III (for LEP students
only); b) Student Academic Achievement Standards, Title I (for all students, including LEP); c) Title III English Language Proficiency AMAO (for LEP students only); and d) Title I Annual Measurable Achievement Objectives (for all students, including LEP students).

Since in the present test design we are dealing with English language learners, Title III has critical implications to this population in terms of assessment. According to West and Kenworthy (2004), the purpose of Title III is to ensure that limited English proficient students master English and meet the same rigorous standards for academic achievement as all children are expected to meet, and to support to the extent possible, the native language skills of such children. They claim that the main purposes of Title III are for ELL students to attain English proficiency, to achieve at high levels in core academic subjects, to develop high-quality language instruction educational programs, and to promote parental and community participation.

Under the umbrella of Title III English language proficiency, the Annual Measurable Achievement Objectives are: state defined English language proficiency achievement targets; they are also based on State English language proficiency standards and baseline data; and are used to evaluate the effectiveness of language instruction educational programs funded under Title III. According to the USDE the fulfilling requirements for the AMAO, such annual objectives must include: (i) at minimum, annual increases in the number of percentage of children making progress in learning English; and (ii) at a minimum, annual increases in the number of percentage of children attaining English proficiency by the end of each year and (iii)
making adequate yearly progress for limited English proficient children as described in section 1111 (b)(2)(B)

In addition, NCLB says to the field that a goal of 100% of LEP students will be proficient in English within 3-5 years as a result of instruction in an English language acquisition program. However, many scholars disagree with this assumption (e.g., Collier, 1987; Hakuta, Butler & Witt, 2000; Klesmer, 1994; Cummings, 1981; Baker, 2001). These advocates suggested that unequal and discriminatory assessment is being claimed where a mandate of law process follows linguistic research ‘critical theory’.

Additionally, Cummings postulates the time line of BICS (Basic Interpersonal Communication Skills) & CALP (Cognitive Academic Language Proficiency) acquisition depends on the instructional practices whether or not it is context embedded or context reduced. Thus, according to him it takes approximately two years for the second language learner to reach the same level of proficiency as a monolingual in context embedded language proficiency and five to eight years to reach the same level as the monolingual in context reduced language proficiency. Likewise, Shohamy (1999) reported that seven to nine years are needed for heterogeneous immigrant students in Israel to catch up with native speakers in Hebrew literacy. She says that the immigrants are chasing a moving target. Finally, Cummings (1984) claims that the failure to take account of the BICS/CALP (conversation/academic) distinction has resulted in discriminatory psychological assessment of bilingual students and premature exit from language support programs.
(e.g. bilingual education) into mainstream classes.

A common definition of language proficiency in English adopted by the Oregon Department of Education is that of Hakuta (2000, p.2):

English language proficiency is demonstrated through the mastery of reading, speaking, listening, and the writing system (Hakuta, 1999) & Achieving full proficiency in English includes far more than mere fluency in conversation. It means that students know English well enough to be fully competitive in academic uses of English and their age equivalent speaking peers. (Hakuta, 2000, p.2).

The Oregon Department of Education (ODE) agrees with this definition since they state that linguistic proficiency is an indication that proficient students are ready to understand or comprehend (listen to) oral instruction in English. It also indicates that the student can comprehend and obtain information (reading) from content area textbooks, as well as write and speak about what the student is supposed to know and be able to demonstrate.

As mentioned before, the AMAOs are critical objectives for English language learners’ attainment of English proficiency. Therefore, NCLB states that if students are not meeting these objectives four years in a row, different sanctions will be implemented. For instance, if a district fails to meet AMAOs, ODE will require the district to modify the district’s curriculum, program, and method of instruction; or it will make a determination whether the district shall continue to receive funds related to the district’s failure to meet the AMAOs. ODE will also require the district to
replace educational personnel relevant to the district’s failure to meet the AMAOs. As a consequence of the previous sanctions, the Oregon K-12 ELP Test plays a very important role in measuring English proficiency for ELLs, but also to report scores in order to compensate for the AMAO requirements. In the section below, I will present the implications of measurement as a result of NCLB and AMAO mandates.

**NCLB and its Repercussion on Large Scale Assessment**

Given the outcomes from the No Child Left Behind Act of 2001 and the enormous push for assessment and accountability of the law, large scale assessments throughout the US are both changing their dimensions and affecting school districts in search to attain improvement targets and raise student achievement levels. As a consequence, Sweeney and Kahl (2004) declare that the testing industry is required to produce ‘faster, better, and cheaper’ assessment measurements. They claim that testing companies, states departments of education, and local school systems are able to produce technically reliable tests with all the psychometric properties required. Nevertheless, they claim that to evaluate if a test is better, the question becomes “better for what purpose?” (p. 2). They postulate that a specific test must fulfill the requirements of testing the right requirements leading to appropriate information. If such a test does not meet these two criteria, then the faster, cheaper, and even technically better may be useless on the long run. Unfortunately, due to the NCLB pressures Sweeney and Kahl (2004) affirm that many educators fall into little understanding of assessment issues and the complexity around them. So cheaper and
faster tests, which do not cover the right requirements or assess the expected criterion are utilized. Therefore, in order to mitigate the misunderstanding, they point out several options convenient for schools, as well as their pros and cons.

They mention the significance of understanding standards-based tests since they convey ‘standards’ in two ways. First, “items are addressed to a particular set of content standards, which are statements of objectives defining the domain of knowledge and skills to be learned and assessed at a particular grade in a particular subject” (Sweeney & Kahl, 2004, p. 2). Secondly, the results must be reported in the context of performance standards, which reflect various threshold scores leading to test score ranges that correlate to different categories or levels of performance. Additionally, the notion of content standards and alignment is critical in order to fulfill NCLB stipulations. Thus, the coverage of the appropriate content is a critical rationale to develop standards-based tests. Some temporary options are the so-called ‘Off-the-Shelf’ commercial tests marketed to many school districts, since they address content. However, NCLB and Title I require state level accountability for tests to be aligned closely with the state’s own content standards. They even require what is known as a two-way alignment. Such alignment makes sure that all test questions address the standards and that all measurable standards are addressed by questions in the test. Thus, the test questions must reflect the depth of knowledge (cognitive level and complexity) and the amount of knowledge communicated by the content standards. In addition, there must be a balance between items across the standards.
Another denominator for NCLB accountability is performance standards and achievement levels, which reflect student performance relative to established cut scores as the reporting tool of results in a standards-based program. Therefore, a student’s performance level is reported with the added numerical test score.

NCLB mandate requires the notion of accountability of proficiency, which calls for states to use at least three performance categories or achievement levels. However, the level of concern in the field is the ‘proficient level’. Thus, by 2014, NCLB expects 100 percent of the students in every school and in each required subgroup in a school to attain ‘proficient level’. Sweeney and Kahl (2004) note that because each state independently determines the quality of work earning the proficiency designation, the percentage of students performing at proficient, as those vary from state to state. So they say this ‘proficiency expectation’ is quite questionable for schools in states with very high standards. They also remark that many states have already implemented standards based assessment programs prior to NCLB but they face the challenge of identifying cut scores across grades. Thus, with the added-grade testing mandated by NCLB, states need to comply for this as in the past they just had scattered cut scores on just individual grades.

Ethical Considerations for Large Scale Assessment Including Language Assessment

Given the critical requirements (as devised by NCLB) to make decisions about individual students, many states use the results of standards-based tests for
decisions about graduation, advancement, summer school, etc. This of course affects English language learners’ accountability for proficiency vis-à-vis funding and sanctions to school districts since the use of testing results draws attention to the use of cut scores and tests. So a student’s score to either above or below a cut score is based on the proficiency based on the evidence from the test. Due to this, Sweeney and Kahl (2004) state that there is a fifty-fifty chance that their true proficiency levels are reversed. They state that even for a test of the highest quality, abnormalities occur as the result of a psychometric concept called measurement error. In other words, they claim that measurement error does not refer to a mistake that has been made but it indicates that any test is subject to some degree of imprecision. Therefore, Sweeney and Kahl (2004) recommend that when using test scores to inform high-stakes decisions, students must be offered multiple testing opportunities. Such rationale is reflected below:

A truly proficient student is not likely to fail due simply to measurement error When given several retest opportunities. The fact that different students can Demonstrate their proficiency better by different methods is a reason it is also important that districts and states offer alternative measures and take into account additional, relevant information in making decisions about individual students. (p. 3)

Additionally, Shohamy (1998, 2001), a critical theorist, suggests that language testers must realize that much of the strength of tests is not their technical quality but the way they are used in social and political contexts. This has implications in the use of tests because their power reflects important decisions about a test-taker’s life that
can result from them. Thus, depending on what the political agendas are, tests scores can have detrimental consequences on test-takers. Similarly, Norton (1998, p. 313) mentions the significance of accountability in language testing. She refers to the responsibility of language testers to the ‘stakeholders’ in the testing process. These include the test takers, teachers, school administrators, community agencies, public officials, etc. Therefore, the idea behind these actions would be to make language testers accountable to these stakeholders and to ensure positive impact of testing practices on them. Norton (1998), and Fulcher and Bramford (1996) also point out the notion of fairness in assessment because they say that in the U.S., legislation exists to prevent the discrimination against any subgroup of the test taking population, and this legislation has had a direct impact upon the adoption of testing standards.

Similarly, Lynch (2001) refers to fairness in two ways: from the alternative assessment perspective in which reaching a fair consensus on the meaning of a student’s portfolio is critical. Second, from the traditional validity/reliability frameworks that deals to statistically significant inter-rater correlation criterion of fairness to be prominent.

Additionally, Wall and Alderson (1993), Alderson and Hamp-Lyons (1996), and Hamp-Lyons (1997) indicate the notion of ‘washback’ is being investigated more carefully that even before due to ethical consequences of the test use on teaching practice. They claim that washback is more complex that thought before, since it seeks to investigate the relationship between test use and the society, which it is used. As a consequence, Lynch (2001) drawing on Foucault (1982) goes beyond this notion
of washback while talking about the importance of free and ethical power relations, as they postulate:

Do participants change the way in which they relate to each other and to themselves (Do students assume and obtain more responsibility in the curriculum; do teachers gain control over assessment policies previously established by others)? – Do these changes become fixed, or are they established as reversible, mobile relations of power? (p. 366)

His view on the notion of washback is relevant for the present study as it currently seeks to see teachers’ views on the assessment in question but also it contains implications to see any effect on their communities.

Finally, Sweeney and Kahl (2004) warn us to carefully make use of Norm Reference Tests since they are not designed to comprehensively cover a set of content standards that may be used by a state or local district. Thus, NRTs do not measure how well schools are teaching or students are learning the material defined by the relevant content standards. They also claim that NRT items do not cover the full range of many state content standards. Due to its inappropriateness in some items, the USDE only allows the use of off-the shelf NRTs for NCLB assessment and accountability purposes if such tests have enough additional items to ensure adequate coverage of standards. Consequently, this process results into another issue. Thus, test security is compromised as many states use commercial off-the shelf NRTs since 1980 as many districts have been using the same tests at the local level due to their easy availability.
Assessment/Testing for Language Minority Students

According to Geisinger and Carlson (1992), 15% to 20% of school children speak a foreign language at home and they do not speak English as their primary language and their culture frequently differs from that of most Americans. As a result, they claim that these differences in language and culture have critical influences in how ELP (Limited English Proficiency) students do in school and on the various tests educators use to assess students. For instance, they warn about the need to understand how cultural background influences LEP students’ approaches to test taking and educators need to consider the students’ individual level of acculturation.

An additional consideration in this matter is the need for test publishers to document their efforts to ensure the fairness of their tests by having sensitivity panels review tests questions in early development, by documenting evidence showing the effective use of the test with LEP students, and by citing differential validation research showing that the test is equally valid and appropriate for use with language majority and minority students. A careful analysis of the scores of certain tests must be considered, especially while used to predict success in future academic experiences.

Duran (1989) has advocated using a test-teach-test- paradigm. This means that after testing to ensure that an LEP student does not know a particular concept, a test administrator teaches the student the concept and then tests the student again. However, this technique is hard to quantify, standardize, and validate, as well as being time-consuming. Yet Geisinger and Carlson (1992) suggest it may informally
report a student’s learning ability.

**Test Construction/Development & Test Specifications**

Hughes (1989) defines that “test development is best thought of as a task to be carried out by a team” (p. 58). He declares that it would be very difficult and unwise for an individual to develop a test because there is a need to look objectively at what is being processed at each stage of development. Hughes (1989) recommends the following procedures for test construction as depicted on table 2.

**Table 2. Hughes’ Procedures for Test Construction (1989)**

<table>
<thead>
<tr>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make a full and clear statement of the testing ‘problem’</td>
</tr>
<tr>
<td>Write complete specifications for the test</td>
</tr>
<tr>
<td>Write and moderate items</td>
</tr>
<tr>
<td>Trial the items informally on native speakers and reject or modify problematic ones as necessary</td>
</tr>
<tr>
<td>Trial the test on a group of non-native speakers similar to those for whom the test is intended</td>
</tr>
<tr>
<td>Analyze the results of the trial and make any necessary changes</td>
</tr>
<tr>
<td>Calibrate scales</td>
</tr>
<tr>
<td>Validate</td>
</tr>
<tr>
<td>Write handbooks for test takers, test users and staff</td>
</tr>
<tr>
<td>Train any necessary staff (interviewers, raters, etc.)</td>
</tr>
</tbody>
</table>

The series of steps for test construction as presented in the previous table, can be taken into consideration after the results of this study are show in order to see if a given step was missing within this process. In addition Hughes (1989) provides a framework for crafting the content as depicted in table 3.
Table 3. Hughes (1989) Framework for Crafting the Content of a Test.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Operations or tasks that candidates should be able to carry out</td>
</tr>
<tr>
<td>2</td>
<td>Types of text such as letters, academic essays, etc.</td>
</tr>
<tr>
<td>3</td>
<td>Addressees of texts or the types of people that the candidate is expected to be able to write or speak to</td>
</tr>
<tr>
<td>4</td>
<td>Length of texts or how expanded a passage or spoken text it should be</td>
</tr>
<tr>
<td>5</td>
<td>Topics as connected to the suitability for the candidate and the type of test</td>
</tr>
<tr>
<td>6</td>
<td>Readability or the level of readiness</td>
</tr>
<tr>
<td>7</td>
<td>Structural range or a list of structures that should be included, rejected or a range indication of structures</td>
</tr>
<tr>
<td>8</td>
<td>Vocabulary range</td>
</tr>
<tr>
<td>9</td>
<td>Dialect, accent, style that test takers are meant to understand or expected to write or speak</td>
</tr>
<tr>
<td>10</td>
<td>Speed of processing or the number of words to be read per minute or the rate of speech, or the speed at which texts are spoken</td>
</tr>
</tbody>
</table>

As presented above, this process is a traditional view in preparing and dealing with test development. In addition another view is becoming more predominant as the Oregon Department of Education has adopted this model. Davidson and Lynch (2002) have believed that the CRM approach to test development is superior to that employed by NRM for many testing purposes. Davidson and Lynch (2002) claim that:

- good tests involve clear thinking, and regardless of the use of the test score, certain fundamental practices always seem to apply . . . Iterative, consensus-based, specification-driven testing . . . we advocate tests that are developed in an iterative manner: there are cycles of feedback-laden improvement over time as the test grows and evolves . . . we advocate tests that are consensus-based: the test should result from dialogue and debate among a group of educators, and it should not result from a top-down dictate, at least not without negotiation from the bottom-up as well. And finally, we advocate tests that are specification-driven: a specification is an efficient generative
recipe for a test that fosters dialogue and discovery at a higher, more abstract level than achieved by analysts of a simple item or task. (p. 7)

Popham (1994), and Davidson & Lynch (2002) share the notion that specifications must be as clear as possible as to both teachers and students can understand their functionality. They suggest a “boiled-down general description of what’s going on in the successful examinee’s head be accompanied by a set of varied, but exhaustive, illustrative items” (p. 17-18). Therefore this approach to specifications should provide enough clarity for both item writers and teachers by not falling into the notion of instructing to a particular skill from the test. This, avoiding the ‘teaching for the test’ effect. Figure 3 depicts the role of test specifications in stages of test development.

Figure 3. The role of test specifications in stages of test development, from Davidson and Lynch (2002)

<table>
<thead>
<tr>
<th>Mandate</th>
</tr>
</thead>
<tbody>
<tr>
<td>↓</td>
</tr>
<tr>
<td>(1) ----&gt; (2) ----&gt; (3) ----&gt; (4) ----&gt; (5)</td>
</tr>
<tr>
<td>select skill</td>
</tr>
<tr>
<td>↑ from spec</td>
</tr>
<tr>
<td>↓ ↓ ↓ ↓ ↓</td>
</tr>
<tr>
<td>+------------+-----------------+</td>
</tr>
<tr>
<td>for test revision</td>
</tr>
</tbody>
</table>

As observed in the figure 3 above, Davidson and Lynch (2002) outline the importance of selecting the skills or criterions that need to be tested, following into writing the spec, creating the items, assembling the test trials and operationalizing the measure just as outlined in the Hughes’ model. However, notice the fact that this
model allows for ‘iterative’ feedback for the test revision since it involves teachers, item writers, administrators, and item developers. In addition, Davidson and Lynch (2002) remark that specs are the result of development of CRM leading to a test result that portrays the test-taker’s ability on some criterion or set of criteria. They also declare that specs can help both NRM and CRM test purposes. Since specs mirror clarity, they allow feedback, consensus among the test development team and the spec is stabilized and can reach a stable agreement. Consensus among the team members can be documented overtime and it becomes evidence to support content validity (p. 16). Finally, in order to describe the characteristics of this model in terms of writing test specifications, the following components of the specs need to happen: a) Mandate – the policy, mandate, or standard from which the spec is devised or motivated, b) General Description – indicates behavior or skill to be tested, c) Prompt Attributes – detail what will be given to test taker, including directions, d) Response Attributes – describes in detail what test taker will do, e) Sample Item – explicit format and content patterns for item or tasks that will be produced from the specs, and f) Specification supplement – additional information including rating/scoring procedure, time allotment, etc.

Reverse-Engineering

Davidson and Lynch (2002) describe the notion of spec-reverse engineering as the creation of a test spec from representative test items or tasks. They claim that not all testing is spec-driven since ‘spec-driven’ tests are those that have been created
from a specification and for which much of the test development, evolution, and maintenance of specs has been devised. Davidson (2003) provides the definition of spec-reverse engineering as “the production of test specifications when none exist: all we have are test items or tasks, and from those we induce what the blueprint might be” (p. 1). In addition he postulates there seem to be four types of reverse engineering:

1. Straight RE – the spec seeks to replicate the item or task exactly as it is.
2. Historical RE – specs are straight-reverse-engineered across many versions of a test, in order to better understand how the test has changed over time.
3. Critical RE – the spec replicates and improves the item or task.
4. Test Deconstruction – the spec is used to better explain the contexts (social, political, curricular, theoretical) in which the test operates. The spec is, itself, a waystage tool toward a larger analysis.

Davidson and Lynch (2002) also state that cases of existing specs could be possible in which are usually ignored in favor or some other common consensus among the test developer team. Thus a difference among the ‘real spec’ in use and the ‘supposed spec’ can be part of the testing manual. Therefore, they claim that the best solution for this issue is to re-create the specs that might exist and to review them so reverse engineering can help clarify a particular test setting which can actually reflect spec-driven characteristics. Further, RE allows analyzing a set of existing test tasks in order to decide how similar they are and then create the spec that could bring them together.
Another application from reverse-engineering is in case of a need to move from a testing context in which there was not any formal instructions or guidance for test development to another reflecting spec-driven. They mention that even in a spec-driven situation, it could be hard to start creating a test specification effectively for a new test criterion. Davidson and Lynch (2003) declare that reverse engineering is a powerful notion since it permits the test development team to clarify what they are trying to measure as an integral part of producing a viable spec. Another application from this procedure is the fact that it can be used to critique existing specs and tests.

Since there are several applications to reverse engineering, they also call the attention to the notion of ‘Item/Task Fit-to-Spec’. This consists of assuring that items or tasks generated by a test specification matches what is described in the spec. Hambleton (1980) declares that this process reflects a form of evidence to establish test validity so practitioners in the testing team can argue whether or not a test is measuring what is supposed to measure. He claims that if test items or tasks from a test procedure are indeed the result of valid inferences about the examinee’s abilities this should be identifiable in the characteristics from the test specifications. Further, an interesting notion behind spec reverse engineering is Davidson’s (2003) concern on the policies of NCLB as he states:

In the realpolitik of the “No Child Left Behind” era, I find myself worried about kneejerk reactions to complex social decision problems. If we put stress the testing corporations and agencies, then they will produce those tools with which they are most accustomed: measures that maintain stasis. Stasis
provides comfort and predictability, particularly in times of political pressures. You can maintain stasis through straight reverse engineering. You can produce tests that are consistent and equivalent to past measures. You can do even better if you couple reverse engineering (which is essentially a content-focused procedure) with careful statistically oriented test development. This leads us to a double-whammy: content determinism plus statistical determinism (p. 2)

As observed in his previous citation, he maintains that NCLB pressures to the field may enhance the use of traditional tests as produced by testing companies and educators may rush into these alternatives. However, he supports the idea that straight reverse engineering allows mitigating pressured decisions and will revive a better development of tests consistent and equivalent to past measures. He also talks about adopting what he calls ‘critical reverse engineering’ to change testing practices for the good of the field since he claims this process may be slow but diplomatic and carefully processed.

As a result of these statements, Davidson (2003) argues that spec reverse engineering should be Iterative, Flexible, and Consensus-Based as a predictive descriptor of a successful product of test specification crafting. Thus, for him, ‘Iterative’ means that the specifications are responsive to feedback from peers’ comments, from trialing and from operational use. The second principle ‘Flexible’ implies specifications that do not get set in stone, but that they change and adapt to shifting contexts. And third principle, ‘Consensus-based’ refers to the unique benefit
of specifications as a cooperative aspect of all interested parties in the test development; this last characteristic is particularly related to the promise of test reform.

Conclusion

Language testing plays a critical role in the collection of systematic information about English Language Learners' ability and progress in K-12 public schools in Oregon. After having made emphasis from the previous literature review on the main aspects surrounding large-scale assessment, the process of testing development was the motivation for this thesis.

In the ideal world and situation, test developers and teachers should be able to interact in the process of test development as outlined in the previous literature review. The Oregon ELPA instrument contains various of the components mentioned previously and it goes through some of the processes as posed in the literature. Such elements are the notion of a CRM, implications and design for tests addressing minority (ELL) students, proficiency measurement types of testing, and of course the need to have item test specifications and if not, make use of reverse-engineering.

Additionally, my intention is to being able to have access for teachers to a more intimate participation in the design of this particular test, to have their voices heard by those who make decisions in terms of standardized large scale-assessment. Popham (2001) declares that testing is relatively unknown area for teachers and the public in general, saying, “Unfortunately, the public knows less about tests than even
educators do. And that's really unfortunate, because tests are becoming such a significant criterion by which to judge the quality of schooling” (p.1).

As the previous statement from Popham outlines the need to inform teachers in the development of tests, teacher evaluation in this project is essential; the last section of the chapter outlines the research question leading this thesis project.

**Research Questions**

1. What is the relationship between teachers’ evaluation of item formats & test items and the original test construct and ELP standards?
2. What is the relationship between teachers’ evaluation of writing item formats & test items and the original test construct and ELP standards?
3. What are the most suitable writing formats to reverse-engineer item specifications based on teacher input from the variables reflected in the instruments of data collection?
CHAPTER 3

METHODOLOGY

Introduction

The purpose of this study was to identify what item formats were more suitable to develop sample draft item/task test specifications using Davidson and Lynch’s (2002) ‘reverse-engineering’ process for test specifications. In this case, the specifications to be reverse engineered are for the ODE’s ESL test, which was not originally developed using item test specifications. The only tools used were the theoretical framework for the test construct (language functions & forms), some linguistic background on second language acquisition (levels of proficiency), the ELP standards, and the projected draft blue print devised by the project psychometrician.

Further, during the first pilot testing for the development of this assessment, the main objective was to detect what item formats would be problematic for test-takers. A taxonomy of item formats taken from the language testing literature was developed by Perea (2004) and used to categorize the initial item bank developed by item writers. About 6000 students took the pilot test and the results indicated that all of the item formats performed very well across all students so that there was no way to eliminate bad item formats, but just some individual test items.

This evaluation and judgments from teachers on item formats will establish in the future a better framework and suitability as to what items/formats can be reverse-engineered. Additionally, item writers will have a better picture of item formats and
illustrative items underlying the principles behind the crafting of the test specifications and test item writing. The teacher evaluation pursued in this study allowed determining what item formats were better for ESL students as well as their suitability to the construct/standards.

Eventually, the results from this study will shed light future work to create detailed item test specifications. Teachers may know what students will encounter during future facets of the test as well as the expected response, and all the elements needed to create these test items.

If using Davidson and Lynch’s (2002) specification guidelines, these will allow novice to experienced item writers (who are teachers), to get a more comprehensible sense of how to generate language test items in a more systematic, organized and fair way. Given this, future teachers writing items may have explicit examples of each target item as well as congruent instructions on how to create the items.

**Research Design & Paradigm**

After having the opportunity to constantly interact with the ELP Assessment Team and fully participate in the process of creating and developing this test, initially the design thought to be applied in this project was ‘Case Study.’ For instance, Alderman et al. (1976: 140) as cited in Nunan (1992); defines case study as the study of an ‘instance’ in action. So to speak, one selects an instance from the class of objects and phenomena one is investigating and determines the way this instance
functions in context.

Yin (1993) declares that case studies can be used for theory building, they favor theory testing, it considers context as essential part of phenomenon of being evaluated, assumes a single objective of reality that can be investigated through the traditional rules of scientific inquiry.

For the original and initial scope of this research, a neo-ethnographic case study with evaluative features was suggested. For instance, neo-ethnographic case study deals with the in-depth investigation of a single case by a participant observer. Concurrently, evaluative case study present an investigation carried out in order to evaluate policy or practice (in this case the language testing development). Nevertheless, given the results and the characteristics of the phenomenon under study, a real life ‘exploration of test spec design and teacher reaction’ within the positivist paradigm is fully recommended as more fitting. In other words an exploration to observe the implications in a given large scale assessment development process when processes shift and these are reflected on policy.

Additionally, Davidson and Lynch (2002) remark the importance of a cooperative approach to developing language tests as in a ‘flexible’, ‘iterative’, and ‘consensus-based’ fashion. Thus, the test development process should be carried out within a group of educators interested in improving the quality of the test. Hence, all the parties involved in the designing of the instrument are participating, their voices are heard (and this includes those individuals outside of the test development environment, e.g. teachers applying the test), and the resulting output (e.g. item
formats, specifications) is always evolving as it is adapted to the needs of the context. For the present thesis study, the bounded system from which case study methodology is applied is the group of teachers who bring their knowledge, education, and experience while interacting with K-12 English language learners.

Along with this, the phenomenon under study is also bounded by the notion of assessment items & item formats and test specifications that interrelate to each other. In other words, the evaluation of items and item formats from teachers has critical implications for test specifications as this teacher judgment and expertise within this study established a more detailed and accurate framework / vision for what item formats are more suitable for English Language Learners given their level, characteristics, teacher / instructional practices, etc.

Similarly, Nunan (1992) relies on Smith, cited in Stake (1998) to find the systematic connections among the speculations, judgments (in this case from teachers), causes, etc. to their judgment on language testing.

Finally, it is important to address the paradigm in which the present study is weighed. Given the fact of using a ‘true exploration of test spec design and teacher reaction’ with the possibilities of survey (e.g. questionnaire) elements, the ‘Post-positivist’ (a modified version of positivism) paradigm was used in order to pursue the perspectives of objectivity as a regulatory ideal. Lynch (1996), Bechtel (1998), and Cook and Campbell (1979) describe Positivist perspectives as those reflecting the notions that a complete objectivity can be achieved and that reality can be perfectly perceived, objectivity as a ‘regulatory ideal’. Cook and Campbell (1979 as cited in
Lynch, 1996, pg. 42) define the (post) positivist paradigm as:

a perspective that enables us to recognize causal perceptions as ‘subjective’ or
‘constructed by the mind’; but at the same time it stresses that many causal
perceptions constitute assertions about the nature of the world which go
beyond the immediate experience of perceivers and so have objective contents
which can be right or wrong. (pg. 29)

Therefore, the (post) positivistic paradigm view permits to study and interpret
the phenomenon we are observing in an objective and systematic way as it has
traditionally been the case in the language-testing environment. Despite the nature of
the proposed exploration design, the desire to use objectivity as a regulatory ideal,
and the use of quantitative instruments and analysis called for the positivistic
paradigm to guide this research.

**Setting**

In order to describe the setting in the completion of the present thesis project,
it is important to mention that teachers from two settings, from school districts in
Oregon and some from graduate courses in the field of education, are the key
locations for the data collection.

The process was divided into two sources of data gathering. The first part was
to pick some of the most representative and diverse school districts (e.g. Woodburn,
Salem/Keizer, Portland Public Schools, Canby or Hillsboro) and ask ESL teachers to
evaluate item formats based on the existing items. The other source of data collection
comes from already practicing teachers in the Portland Metro Area who are taking their ESL endorsement classes or other educational ESL related courses at the graduate level in Portland State University, Washington State University, and William Patterson University.

School districts that reflect a representative range of teachers serving from ESL up to sheltered and mainstream classes were chosen. Districts such as Woodburn, Canby, or Portland among others were targeted since they have transitional or bilingual English language programs depicting a distribution of language proficiency levels in their student population as required in the ELP assessment scale & ELP standards. In other words, those districts possessing a wide range of teachers addressing different range levels and English levels were key in this study. Similarly, teachers in graduate classes represented various school districts in Oregon as this allowed more diversity in the sample.

Participants

In order to describe the participants in the completion of the present thesis study, I will do so by differentiating their function in the current thesis project.

Teacher Participants (for Item Format Evaluation)

These participants played a very critical role in the development and evaluation of ELL assessment since their voices needed to be heard and considered in the decision making of assessments. They are the primary and direct contact to the students. They are the ones who can perceive whether a positive washback effect is
reflected on the test takers because of these large-scale assessments. They can observe the examinees reactions to the test and technology used to deliver it. This is the reason why knowing their perceptions and allowing them to make decisions on the analysis of items and item formats is rather critical for validity in future development of specifications (Davidson & Lynch, 2002). Therefore, a group of these teachers were eventually approached through a questionnaire instrument in order to obtain their perceptions and feedback on the evaluation and improvement of item formats and their relationship with test items and the construct.

I selected teacher participants from Education and ESL Graduate classes and from School districts from Oregon. Two criteria were used: 1) They were already practicing educators in public schools and had contact to English Language Learners and/or 2) they had a strong background in ESL training with some experience teaching ELLs. The participants from the graduate classes were from Portland State University (PSU), Washington State University (WSU), and William Patterson University (WPU). Participants’ gathered data from WSU and WPU were used for piloting the data collection instruments in the present study. This allowed re-defining any discrepancies or reactions reflected in the surveys. They were already practicing teachers or near to be practicing, but with strong educational-theoretical background.

Despite the fact that they were graduate students, they were already involved in school districts. These teachers were current teaching practitioners taking graduate courses to get their ESL endorsement or any other related ESL educational course at these universities. The practicing teachers (non-graduate students) were mainly from
Beaverton, Woodburn, Portland, Hood River, Salem/Keizer, Canby, Hillsboro, Gresham, Tualatin, among other school districts. At least 35 different districts among Oregon and Texas data were captured from. As a result, data captured from the teacher background checklist (see appendix 1 for the specific variables) is illustrated in the following table 4.
Table 4. Demographics of All the Participants in All Modalities Based On the Teachers’ Background Checklist

<table>
<thead>
<tr>
<th>Total participants in all four modalities</th>
<th>Listening</th>
<th>Reading</th>
<th>Speaking</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total by modality</td>
<td>27</td>
<td>21</td>
<td>26</td>
<td>54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean distribution of teachers by teaching level/setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total by grade level</td>
</tr>
<tr>
<td>Elementary</td>
</tr>
<tr>
<td>Middle</td>
</tr>
<tr>
<td>High School</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total average years of ESL teaching experience</th>
<th>Listening</th>
<th>Reading</th>
<th>Speaking</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>= 8.46</td>
<td>7.7 years</td>
<td>13 years</td>
<td>7.14 years</td>
<td>6 years</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean distribution of grade group teacher level experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-1</td>
</tr>
<tr>
<td>19%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean distribution of English proficiency level taught from participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entering</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>19%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean distribution of student ethnicity teacher experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>34%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean distribution of ELL/ESL method teaching training/ experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistics courses</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>16%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean Distribution of ESL program teaching experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total mean distribution of teachers’ test experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
</tr>
<tr>
<td>10%</td>
</tr>
</tbody>
</table>
This table summarizes all the demographic and background data of all participants (in the total scope), and also makes reference to the modality reported in this study (writing). The purpose of demonstrating this data is to allow the reader to see the scope and characteristics desired in the participants of this study. Most importantly, I will discuss the implications and value of this data in the discussion chapter in order to establish an argument to the importance of having teacher input in large scale assessment and to support the notion of face validity for this study.

**Sampling**

The instruments for item format evaluation procedures were applied to teachers across different school districts in Oregon (including teachers from graduate classes) in order to gain a better representative sampling, which reflects these teachers having exposure to students at all the grade groups and proficiency levels as reflected in the Oregon ELP Standards and the expected test construct scale. These participants performed the actual item format evaluation for practicality in the classroom (item formats vs. test items). Thus, those school districts, which reflected these types of teachers teaching this particular ELL student population were selected.

In order to find the representative teacher participants that reflected a range of teaching & proficiency levels, grade groups, and diversity, a purposive selection of data sources was approached (Sage, 2002; Morgan, 2005). This involves collecting detailed information from a relatively small number of carefully chosen sources. Since it was virtually impossible to sample every teacher in Oregon at the K-12
levels, those school districts, which may have a good distribution of ESL and Sheltered teachers and student language proficiency levels (given the nature of their program), were selected. In this case, Canby, Woodburn, Salem/Keizer, Hood River, Tualatin, Hillsboro, and Portland Public Schools were good candidates to approach teachers given the scope of their programs. For instance, they have early and late exit transitional programs as well as dual language programs. Simultaneously, the teachers in ESL endorsement education graduate classes held at Portland State University were sampled since educators from various districts were represented. Patton (2002) and Morgan (2005) postulate four basic strategies for the purposive selection of data sources.

These are: selecting sources by defining characteristics, for systematic comparisons, with unusual information and for theory development. Given the nature of the present study, selecting sources by defining characteristics is the most appropriate. This uses “carefully-stated eligibility criteria to select specific types of cases (or people, experiences, etc.) for in-depth investigation” (Patton, 2002; as cited in Morgan, 2005, p. 1). To mention some practical variations, typical cases (selects sources that typify the subgroup of interest) and information intensity (selects cases based on amount of data) can be used. In our case, since we focused on the school districts with the most diverse and major distribution of ESL students and focused on the particular teachers teaching these students, our sample reflected a wide range of teachers with different experiences, background and points of view (as long as they teach ESL students).
The aim for the number of teachers to give the questionnaire was at the minimum 30 per domain (Listening, Speaking, Reading, and Writing) with a total of 120 in order to gain a normal distribution in the data. The more teachers that were thought potentially to be interviewed and the wider the range of their teaching level, the better for the purposes of this study.

The process to select these teacher participants was by first contacting the ESL coordinators of the proposed school districts. Coordinators were able to refer to specific schools that had the characteristics of the teachers needed for this study. Then an invitation was extended to the teachers (with their voluntary consent) to participate in the evaluation of item formats. ESL Teachers were selected based on a variety of experience teaching all grade levels in order to obtain a more consistent perspective. For instance, the coordinator pinpointed teachers with the desired characteristics and I invited them to participate in the study. All of these teachers were volunteers for this study as they were not paid for their contribution. A teacher background checklist provided better information regarding their qualifications and experience. At least six teachers per grade group (a total of 30 per domain) were to be approached to answer the instrument of data collection. After this step, a 1-hour information training was presented to the participants so they had a better understanding of the construct of the test, and the item formats used. Consequently, they were believed to be able to respond appropriately to the questionnaire instrument.

The criteria used to provide the questionnaire to teachers taking ESL endorsement classes or any related class to the teaching of ESL consisted of inviting
the entire class. First of all, the professor teaching a related ESL endorsement course at PSU (WSU or WPU) was contacted to request his/her permission to approach the graduate students and explain the purpose of conducting this project.

All necessary materials regarding how to answer the questionnaire were forwarded to the professor first. Then the professor determined if the information was or was not appropriate for his/her students and the curriculum of the class. With the professor’s permission, an invitation was forwarded to the entire class of graduate students (current practitioner educators in the K-12 field in Oregon) to request their participation in the current study as evaluators of item formats. Classes ranged from 10 to 20 students. The minimal expected number of participants to gather data from was 6 per grade group (K-1, 2-3, 4-5, 6-8, 9-12) with a total of 30. They only evaluated item formats and items in one single domain (e.g. speaking). Another PSU class performed the same analysis but with different item formats in another domain (e.g. writing). They were provided a 1-hour training / presentation for their understanding of the construct of the test and the components of the item formats so they could answer the questionnaire appropriately. Teachers who did not wish to participate in the study, were given the materials as well, but they were free to not answer the instrument.

Teacher participants in this study (while considered to evaluate item formats) from school districts and graduate classes, needed to cover any of the following characteristics in order to obtain appropriate valid judgments and recommendations in the validation of item formats. In other words, to consider participation in the study,
teacher participants needed to meet one or more of the following characteristics:

1. Have past experience with or currently be teaching English Language Learners in at least any of the following grade bands (K-1, 2-3, 4-5, 6-8, 9-12).

2. Have limited teaching experience but a moderate to substantial training in any of the following areas: teaching ESL, having exposure to linguistics courses, second language acquisition courses, assessment courses (or proctoring experience), training in SIOP, GLAD, CALLA, Systematic ELD or any other sheltered techniques.

3. Have experience teaching sheltered classes at any K-12 level.

4. Have experience proctoring and testing students with some known ‘off-the-shelf’ testing instruments used in public school districts and some state instruments (Woodcock-Muñoz Language Survey (English), SOLOM (Student Oral Language Observation Matrix), SELP (Stanford English Language Proficiency) test, SLEP (Secondary Level English Proficiency), IDEA (IPT) English Proficiency Test, LAS (Language Assessment Scales), LPTS (Language Proficiency Test Series), DRA (Developmental Reading Assessment), IRI (Individual Reading Inventory), Oregon State Writing, Speaking, and Reading work samples (rating scales). *(Refer to appendix 1 for the teacher background checklist).*

This checklist initially allowed determining how many participants in the ideal situation actually would cover the desired qualifications and it informed a cross-check.
to target the minimal sample of 30 participants per domain as well as their representativeness among different school districts and exposure to different students from distinct backgrounds.

**Data Gathering Instruments and Procedures**

*Procedures*

In order to gather data that allowed performing the ‘item format evaluation’ process and including teachers in the actual decision-making process of redefining these item formats based on test items and their ESL/classroom expertise, the use of a questionnaire instrument was pursued.

Given the fact that the questionnaire instrument with different formats (quantitative and qualitative data gathering) was used, but yet keeping consistency with the post-positivistic paradigm a ‘connected contributions’ approach was suggested since it allowed to use the strengths of one method to contribute to the different strengths of the other (Morgan, 2005). This also reflected elements of an ‘expanded coverage’ approach to research design since multiple pieces of data were taken into account to gather the appropriate information that could improve the item formats. Another characteristic of this mixed method was its flexibility to provide an iterative process to involve teachers in the data collection. The following structure in figure 4 for mixed research methods captures the stages of this process:
My first step was using a quantitative gathering approach and then chose to supplement with qualitative data and then repeat the pattern until finalizing the data gathering and analysis with the support of both research methods (data displays and statistical procedures such as mean and reliability scores. The first step was to pilot a survey with a subgroup of graduate student teachers who had some experience in the teaching field, especially in ESL. These teachers were from WSU (Washington State University) and WPU (William Patterson University). Following this, I made a revision in the results of the instruments and data were interpreted in order to modify the existing instruments and apply them to the remaining participants. Ultimately the target participants were teachers in the field at K-12 settings and teachers at graduate classes at PSU (Portland State University) & Oregon schools. The participants then performed the item format evaluation in order to RE specifications based on their input for future studies & developmental stages of the test.

After having selected the teachers, it was expected at least to gather 30 teachers per domain in the original larger scope (reading, speaking, writing, and listening) for the evaluation of item formats and their correlation to test items so we could have obtained a normal distribution in the data. Potentially more than 30 teachers could have been used to answer the instruments of data collection per each domain, but consistently at least 30 teachers were thought to be approached.
It was critical that in each 30-teacher sample per domain, we could obtain teachers who had experience teaching ELLs or training to address ELLs’ needs from diverse grade groups. Thus, ideally it was expected to have at least six teachers evaluating item formats with their items per one grade group (as depicted in the Oregon ELP standards and the projected ELP assessment scale).

The following table 5 represents the initial minimal desired distribution of teachers in the evaluation of item formats and their correlation to test items per domain. This table represented a guiding selection / cross check of participants after having gathered data from them. As the data gathering and recruiting of participants took place, if gaps appeared in the cells, more teachers were asked to answer the questionnaire to reach ideally 30 participants per modality.

<table>
<thead>
<tr>
<th>Reading test items &amp; Item formats</th>
<th>Writing test items &amp; Item formats</th>
<th>Speaking test items &amp; Item formats</th>
<th>Listening test items &amp; item formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade Group</td>
<td># of Teachers</td>
<td>Grade Group</td>
<td># of Teachers</td>
</tr>
<tr>
<td>K-1</td>
<td>6</td>
<td>K-1</td>
<td>6</td>
</tr>
<tr>
<td>2-3</td>
<td>6</td>
<td>2-3</td>
<td>6</td>
</tr>
<tr>
<td>4-5</td>
<td>6</td>
<td>4-5</td>
<td>6</td>
</tr>
<tr>
<td>6-8</td>
<td>6</td>
<td>6-8</td>
<td>6</td>
</tr>
<tr>
<td>9-12</td>
<td>6</td>
<td>9-12</td>
<td>6</td>
</tr>
<tr>
<td>Independent group of 30 teachers</td>
<td>Independent group of 30 teachers</td>
<td>Independent group of 30 teachers</td>
<td>Independent group of 30 teachers</td>
</tr>
</tbody>
</table>

As observed in the previous table depicting the expected numbers of teachers to be sampled for the present study, it is observed that the 30 teachers would have
evaluated item formats with their items for the same domain (e.g. Writing only).

In other words, at least 30 teachers would have focused on the same set of item formats and test items but at least 6 teachers would have answered the questionnaire from the K-1 perspective given their own training and teaching experience. Subsequently, another 6 teachers would have evaluated the same item formats for the 2-3 grade group, 4-5 grade group, and so on. Notice that each group evaluating each domain was totally independent from the other groups. This process allowed practicality in giving the questionnaire to the participants and reducing the load of information.

Ultimately this is the teacher sample I draw from about the role of teacher input for future studies. Despite the fact that in the ideal situation a minimum amount of 30 participants was desired to account for a normal distribution, G-studies (reliability) were ran for the Likert scale instrument in order to see consistency within scores as observed in the following table 6 below.

Table 6. Final Distribution of Participants per Modality

<table>
<thead>
<tr>
<th></th>
<th>Total participants in all four modalities = 128</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total by modality</td>
<td>Listening 27 participants</td>
</tr>
<tr>
<td></td>
<td>Reading 21 participants</td>
</tr>
<tr>
<td></td>
<td>Speaking 26 participants</td>
</tr>
<tr>
<td></td>
<td>Writing 54 participants</td>
</tr>
<tr>
<td>Reliability scores</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>0.92 (26 scores)</td>
</tr>
<tr>
<td></td>
<td>&amp; 0.93 (28 scorers)</td>
</tr>
</tbody>
</table>

**Instrument**

The instrument of data gathering was the use of survey instruments (e.g. questionnaire) as the main format. Cohen and Manion (1985) state that surveys are useful to obtain a ‘snapshot’ of conditions, attitudes, and / or events at a single point
in time. Similarly, Jaeger, 1988 (p. 307, as cited in Nunan, 1992) points out that a survey allows researchers to know about teachers’ honest attitudes and their perception on specific educational issues. Similarly, Morgan (2005) declares that surveys reflect deductive, hypothesis testing since they can measure a relatively large number of variables and investigate the relationships between them. In addition, some conditions exist in order to use surveys. For instance, what population is considered?, what reflects a representative sample of the target population?, what type of sampling can be done?

Nunan (1992) presents some attributes of a survey: a) defining the objectives, b) identifying the target population, c) covering the existing literature review on the topic, d) determining a sample, e) identifying the survey instruments, d) designing the survey procedures, e) identifying the analytical procedures, and f) determining the procedure to report.

In order to avoid the confusion to represent a survey design model in this thesis study, for purposes of consistency, we refer to the data-gathering instrument as questionnaire.

Consequently, from the use of questionnaires (as a type of survey) as means of data gathering, they permit the researcher to collect information in field settings as they allow to process quantification procedures. Some advantages of these instruments are the closed or open-ended question formats. In addition, a Likert-scale format can be applied and quantified.

Alternatively, a set of open questions can allow more freedom to the recipient
to answer (Nunan, 1992). If we want to go along a more objective approach, for the purposes of this study, a questionnaire (Likert-scale) and (open-ended) format was quantified in order to establish statistically reliable results.

Whenever possible, it was indispensable to pilot these two questionnaire variations in order to make the appropriate arrangements prior to using them for actual data gathering.

As mentioned in the participants section, a sub-group of teachers who were graduate students were asked to answer a preliminary questionnaire (Constructed response and Likert scale format) in order to gather qualitative data from one and quantitative data from the other, with both aiming to establish appropriate item format evaluation and correlation to their items to allow RE of specifications. The subsections below contain the instruments of data collection.

1st instrument: Constructed response method questionnaire in the evaluation of item formats vis-à-vis individual test items.

In the following instrument (see appendix 4), teachers looked at the selected item formats from one domain only (e.g. writing). Then, sample items were shown to them. After having been explained the characteristics of the ELP test construct and the item formats, participants provided their input/judgment to decide whether or not there was a proper suitability between the items and the item formats in which they are embedded and provide comments for enhancement.

The constructed response questionnaire inquires on variables such as item formats, item format difficulty (proficiency & grade group level), item format
appropriateness to ESL students (e.g. matching typical instructional tasks), relationship between items/tasks and item format to the construct of the test, and necessity to pre-teach students to take these test items. This format permitted to gather more in-depth data, suggestions, editing, and to capture teachers’ experiences with their students in regards the test items and item formats.

2nd instrument: Likert scale method questionnaire in the evaluation of item formats vis-à-vis individual test items.

As for the Likert scale format questionnaire (see appendix 5), it allowed me to gather similar data, but to the degree of ‘quantifying it’ and obtaining accurate numbers to visualize the overall opinions of teachers. Some of the variables considered in here were the compressibility of the item formats, ambiguity, appropriateness, difficulty and relationship with the construct (grammatical, illocutionary and textual competence). In order to avoid neutral responses, a 4-point scale was used depicting the following values:

*Strongly disagree - Disagree – Agree – Strongly agree.*

(1) (2) (3) (4)

Purposely a 4-point scale was used in order to avoid the middle value of neutrality that would allow weighing the value responses from the participants to one side or the other in the continuum of the ordinal scale. Consequently, this format allowed the researcher to make more clear-cut and accurate decisions in the selection of appropriate item formats to establish RE of specs.

In addition, having this characteristic in the Likert scale assured validity in the
results as the researcher felt confident on the robustness of the projected data and results.

Notice that in question ‘n’ of the Likert scale instrument participants were asked to provide their judgment on what levels of language proficiency were more suitable for each item format.

These levels of English language proficiency as those as depicted in the Oregon ELP standards and the ELP scale. For purposes of practicality for participants, the version developed and adopted by Scott (2004) was used (see appendix 2 as a reference). Similarly, a taxonomy of language functions as used in the ELP Test construct was used (see appendix 3), Bachman (1990).

Overall, when teachers used the previous two instruments of data collection, they were able to look at each item format from one domain (listening, speaking, reading and writing). Concurrently, they were given a description, function and type of task that each item format was supposed to assess (as found in the language testing literature by Perea, 2004). Perea (2004) conducted an in-depth study at NWREL (Northwest Regional Educational Laboratory) in order to identify the item formats available in the testing literature. These were taken as reference to establish a preliminary taxonomy on item formats that could be used to place items newly created into these specific categories. Further, participants were given some sample items taken from the ELP item bank in order to present a more concrete visual of what these item formats look like.

Finally, participants presented their judgment of these categories considering
their teaching experience, their exposure to ELLs from different ages, proficiency levels and how they believed test-takers would react to specific item formats. As presented previously, this analysis of item formats was carried out vis-à-vis specific grade groups (K-1, 2-3, 4-5, 6-8, and 9-12). These results then allowed re-defining item test specifications in future studies (e.g. reverse engineering of test specs).

After the preliminary instruments had been piloted, the remaining group of participants actually evaluated the item formats with the ‘revised’ questionnaire as given the results from piloting the instrument of data collection.

Moreover, participants needed to understand the definition, function, rationale and reason for using the determinate item formats. An important component in the development of items and item test specifications was to observe what item formats and items best fitted the potential production of RE specifications. Participants needed to identify the most appropriate and suitable item formats and items. These allowed redefining what item formats could be used to future RE specs (e.g. characteristics of the prompt attributes). For instance, certain items were more suitable for older and higher proficiency students than others. Some items would need pictures while others would no or some items would be better on a cloze or sentence completion approach than others.

Thus, participants’ experience with ELLs informed whether or not items/item formats were appropriate for students and to the spec itself, but also their evaluation on appropriateness level of items and item formats informed a careful and more valid use of these formats to grade grouping and proficiency levels from a theoretical
As stated earlier, the 2004 item bank was organized, grouped and coded previously by a group of item writing facilitators and the researcher. Hence, the hypothesis was also to find out (based on teacher evaluations) how these items as grouped in item formats were more suitable onto the creation of RE specifications from both individual items and bigger groupings of them (item formats). What item formats best fit certain grade groups (e.g. K-1, 2-3, 4-5, etc.) and student characteristics to make better decisions to RE specs? These item formats allowed a better approach to organize items and observe their relation to the skills they were intended to test.

Further, as teacher evaluations emerged, the identification and in-depth description of item formats and items (contained within the prompt attribute) will be critical in the evaluation of future studies on ‘item/task-fit-to-spec’ (out of the scope of this project). A finishing approach to these item formats is the fact that they allow to group great quantities of items (entire item banks) according to specific characteristics, skills, levels, etc. Therefore, this is when teachers’ perceptions and experience in dealing with ELLs comes handy to evaluate how specific items and item formats can be used to RE future specs.

Therefore, the list below depicts the item formats used to categorize the ELP item bank in the modality of writing.
Writing

a) Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response)
b) Essay (Extended Response w/Picture, Diagram Stimulus)
c) Multiple Choice (Picture/diagram stimulus w/Sentence Completion)
d) Multiple Choice (Picture/diagram stimulus w/Answer Questions)
e) Multiple Choice (Sentence Completion-Speech Parts Construction)
f) Multiple Choice (Simple sentence stimulus w/Combining Sentences)
g) Multiple Choice (Simple sentence stimulus – correct sentence identification)
h) Short Answer 1 (Sentence Completion-Dichotomous-No Picture)
i) Short Answer 1 or 2 (Sentence Structure/Single Sentence)
j) Short Answer 1 or 2 (Picture Stimulus – Single Word/Phrase/Letter Response)
k) Short Answer 1 or 2 (Picture Stimulus - Sentence Structure/Single Sentence)
l) Short Answer 1 or 2 (Picture Stimulus-Sentence Completion)

Ultimately, participant teachers answering these instruments were able to evaluate the effectiveness and validity of the item formats by looking at some test items taken from the 2004 ELP Test item bank. Qualitative data was gathered from the constructed response questionnaire and quantitative data was taken from the Likert-scale format questionnaire; both aiming to establish valid item format selection and guidance to RE specs from their test items. The concurrent use of both instruments by each participant permitted a more valid notion of triangulation to double check responses.
In regard of proper consent requested to participants, permission was asked beforehand to the professor teaching a selected graduate class. Similar process was pursued with the principal of those schools representing the previously mentioned school districts; and/or the ESL coordinator of that representative school district. Thus, teachers (participants from graduate classes and school districts) were first presented what the purposes of this study were and were voluntarily invited to participate in the evaluation of item formats.

A consent form was presented to the participants in order to ask for their voluntary permission to participate in this project as well as all the expected outcomes and purposes of this research. While they provided their evaluative comments through the questionnaire instrument, their names and the school where they teach was kept anonymous. However, some important data for the purpose of this research was asked to them. Such information refers to the current level of ESL or Sheltered English they teach, average age and known proficiency level of the students they teach. Further, as an important aspect in their validity as participants evaluating item formats/correlation of items, it was recorded what experience they have had while teaching English Language Learners. This was important information to know since those teachers who had plenty of exposure to teaching techniques and to teaching ELLs definitely provided a lot of expertise and advice in how students are assessed as well as to what item formats and individual items best fit these populations. They were presented a general training in the construction of this test in regards of the construct, the item formats and types of items projected for the test. This allowed a better
understanding about how they made recommendations for test items and item formats by using the questionnaire. They were asked to evaluate the suitability of item formats, construct and items by relying in their experience and teaching intuitions in order to document their judgment properly. After they had answered the questionnaire, any other comments and suggestions they had were recorded on paper. After the data had been processed, they received a report outlining the resulting outcomes from their comments.

Finally, teachers were trained in the characteristics of the test construct, proficiency levels, grade groups and language functions in order to address the linguistic framework. In regards to the assessment framework, teachers were informed about the characteristics of these item formats and the definitions of how they were intended to be used.

**Data Analysis Procedures**

Due to the open-ended / constructed response questionnaire format, responses were qualitatively recorded, coded and the resulting information was analyzed through a data display (Lynch, 2002). The final coding categories of data to consider (in table 7 below) were:
Basically for the qualitative data gathering a three-step process was pursued. This included a data coding procedure, a reduction procedure and a final display of results.

The first step consisted of transferring data into data displays by item format, grade group (K-1, 2-3, 4-5, 6-8 & 9-12), and into a positive or negative outcome.

The second step consisted of transferring positive outcomes into a second set of data displays and reducing a lot of the comments into core comments and deleting unnecessary information.

The third step consisted of creating a data set or data map in order to guide the researcher (or a test developer) create sample item test specifications.

Regarding the second format of the questionnaire (Likert scale), it was analyzed in GENOVA software in order to establish a better validity of results and to make generalizations beyond descriptive statistics. Thus, principles of analysis based on the G-theory (Generalizability theory) for reliability purposes were used. For
instance, Brennan (1992) and Lynch (2003) declare that a ‘generalisability study’ (G-study) can provide estimates for percentage of total variability in the test scores across all test takers or raters that each facet (and interaction) accounts for. They called these estimates ‘variance components’, which in turn are used to estimate reliability, or dependability, or inferences drawn from the scores. In addition, they postulate the use of a ‘decision study’ (D-study) to see how different numbers of conditions for the facets, in different combinations, will affect the reliability or dependability estimate.

Despite the fact that the number of participants suggested from the sampling did not result as desired and because what was achieved in terms of sampling numbers, the proposed statistical procedure to analyze the responses to the questionnaire is a non-parametric version of it.

For the purposes of the present study I used descriptive statistics as the main instrument of quantitative data analysis using mean scores in a Likert scale to indicate opinion, and I did not expect nor need the sort of discrimination required from tests where the objective is to make relative decisions concerning individuals (Lynch, 2003). Then I simply examined reliability to see how consistent the participants approached the questions asked in the Likert scale instrument.

The variables reflected in the questionnaire were suitability of items and item formats to: a) construct of test, b) difficulty to student, c) test item/item format relationship, d) compressibility, e) appropriateness to ESL students, f) proficiency level, g) instructional tasks, h) pre-teaching to test-taker, and i) test-taking procedure
to test-taker. The mean scores from each variable identified their suitability to each item format based upon the hierarchical judgments from teachers (1-strongly disagree to 4-strongly agree). The statistical results determined the relationship between test items and item formats with the construct/ELP standards and their relative importance and applicability as applied to test-takers.

Finally, mean results from each variable and the participants from different grade groups (K-1, 2-3, 4-5, 6-8, 9-12) established a total score to make a definite decision on item format applicability to students per domain (listening, speaking, reading, writing) based on teacher judgment, but of course I am reporting on the writing modality only.

Therefore, the results from these descriptive statistics (mean scores) determined the most important item formats upon teacher judgment (perceptions) and how they actually have a meaningful weight to select item formats to RE specifications.

In general, mean scores allowed determining better decisions to establish suitability of item formats with all of the variables mentioned above and reliability studies confirmed the degree of consistency in which raters evaluated the variables in question using the Likert scale instrument.

**Establishing Validity for Research Findings**

In order to establish validity for the present thesis, it was desired for the instruments of data collection to include a minimal sample of 30 teachers per
evaluation of each item format as part of one domain (in this case writing). However, since 54 participants were able to evaluate 26 full questionnaires, the researcher believed these were enough numbers.

Technically, to avoid statistical assumptions to internal validity, the appropriate tests or procedures applicable were using a non-parametric version since we ended up with non-distributed data or small sampling. The statistical procedure used to calculate reliability was Genova Version 3.1 (Brennan, 2001) for the Likert scale resulting in high levels of reliability.

For external validity, I used a purposive sampling of participants (teachers) having taught students with diverse language proficiency levels, which was important as a sign of ‘representativeness and diversity’. This compensated for the possible threat to validity for 'interaction of selection and treatment or setting'. Therefore, the choosing of sites based on representativeness or diversity accounted for this purposive sampling (Lynch, 1996). In other words, the sample teachers who completed the format evaluation possessed enough teaching experience, training and understanding of assessment to provide accurate judgments. So to compensate for the treat to validity of ‘selection’, careful stratification of participants through a purposive sampling was significantly used.

Further, as Morgan (2005) describes the power of using a core research design (e.g. Quantitative), but also to allow qualitative work to help and inform the core method in the research design used, the same approach of mixed research method was implemented. So to speak, the strengths of one method compensated for the
weaknesses of the other (a connected contributions approach).

Finally, I would argue that the resulting outcomes or processes from this project can be generalized to other research contexts in the sense that other school districts or states would confine themselves in the need to create item test specifications while just possessing an item bank (see chapter five for additional discussion on this).
CHAPTER 4
RESULTS

Quantitative Results (Yes/No Questionnaire Open-Ended Results)

Item Format Relationship with the Construct

Writing item formats.

The following chapter illustrates the results of the item formats analysis by teachers when using the quantitative and qualitative data collection instruments.

The table in appendix T also contains the results from participants in their evaluation of item formats in writing regarding their connection to item formats used for this test. Twelve item formats were identified in this modality as listed below:

1) Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended
Response)

2) Essay (Extended Response w/Picture, Diagram Stimulus)

3) Multiple Choice (Picture/Diagram Stimulus w/Sentence Completion)

4) Multiple Choice (Picture/Diagram Stimulus w/Answer Questions)

5) Multiple Choice (Sentence Completion-Speech Parts Construction)

6) Multiple Choice (Simple Sentence Stimulus w/Combining Sentences)

7) Multiple Choice (Simple Sentence Stimulus – Correct Sentence Identification)

8) Short Answer 1 (Sentence Completion – Dichotomous – No Picture)

9) Short Answer 1 or 2 (Sentence Structure/Single Sentence)

10) Short Answer 1 or 2 (Picture Stimulus – Single Word/Phrase/Letter Response)
The following research question is answered for the sections below:

*What is the relationship between teachers’ evaluation of writing item formats & test items and the original test construct and ELP standards? (Test constructs include: grammar, vocabulary, morphology, representational, manipulative, heuristic, imaginative language functions, and cohesion & coherence.)*

**Grammar**

The results observed in the table indicate that all of the item formats were considered by participants as being suitable for assessing grammar. The observed means ranged from 73% to 100%.

**Vocabulary**

Subsequently, writing item formats also correlated well with the teachers’ judgments that they had a good connection to testing vocabulary. Means varied from 73% to 100%, similar to grammar, and in fact having the highest item format in SA 1/2 sentence completion with picture as stimulus.

**Morphology**

Another skill that accounted for the teachers’ positive relationship toward the item formats was morphology, which ranged between 69% and 90%. The highest mean for a format identified was MCh items, with sentence completion and pictures as stimuli.
Representational language functions

In regard of language functions to convey information, all the item formats were shown to suitably test these functions. The lowest mean identified was 72% (MC with sentence completion) and the highest mean was 97% (ER with writing discourse from all genres).

Manipulative language functions

Some discrepancies were found in these functions. Only ER with different genres, MCh with sentence stimulus/correct sentence identification, SA1/2 for sentence structure in single sentences, SA1/2 for sentence structure in single sentences with pictures as stimuli, and SA1/2 with sentence completion with picture stimulus were identified as having a suitable relationship to test these language functions. Percentages appeared from 50% to 59%. The remaining item formats presented a weak relationship to this construct.

Heuristic language functions

The item formats as related to heuristic functions were considered overall as suitable to the skill. The only exception was SA1 (dichotomous item) sentence completion with no picture, which accounted for barely 50%. The remaining item formats indicated a mean range of 55% to 65%.

Imaginative language functions

Language functions for enjoying language were not so suitable according to participants of the item formats in question. The exceptions were ER with different genres (62%), ER with pictures as stimuli (58%), and SA1/2 sentence structure in
single sentences (59%). The rest of the item formats did not account for a true relationship to imaginative functions. Their percentages ranged from 19% to 45%.

**Cohesion and coherence**

Finally, in regards to the writing construct of cohesion and coherence nearly all item formats were shown to be highly suitable for testing these skills in writing discourse. The items with the lowest mean identified were 62% (MC sentence diagram stimulus with sentence completion, MC picture diagram stimulus with answering questions, and SA1 sentence completion with no picture). The highest mean was 96% for ER with picture diagram stimulus.

The next section includes the results on the (yes/no) open-ended questionnaire in regards to the second set of variables. The following research question taps specifically on these remaining variables:

*What are the most suitable writing formats to reverse-engineer item specifications based on teacher input from the variables reflected in the instruments of data collection?*

**Item format difficulty for grade level (IF).**

The table in appendix U shows the results pertaining to the notion of item difficulty for the selected grade groups. As observed, participants believed all the item formats were neither difficult nor ambiguous for each grade level evaluated. As evidence of this, the lowest mean supporting these results ranged from 55% in SA1/2 (picture stimulus – sentence structure/single sentence) to 85% of acceptance in SA1/2 (picture stimulus-sentence completion).
**Item format & test item relationship (IF/TI↔Relationship).**

In regards to the whether the test items fit the item format from which they were produced, the table in appendix U demonstrates that all formats were estimated by teachers to be well constructed, in the sense that the test items contained all the characteristics described by their recipe. As confirmation of this, the item format with the lowest percentile approving this notion was 73% (MCh (simple sentence stimulus – correct sentence identification), while the highest percentile was 96% as seen in item formats such as Essay (ER w/picture, diagram stimulus), and SA1/2 (picture stimulus – sentence completion).

**Item format comprehensibility / generative to similar test items (IF↔Comprensibility).**

The results in appendix U indicate that the participants believed that all item formats were comprehensible for test takers and to the teachers evaluating such formats. The format with the lowest mean was SA1 (sentence completion-dichotomous – no picture) with 77%, and the format with the highest mean supporting this result was MCh (simple sentence stimulus w/combining sentences), which accounted with 100% acceptance.

**Item format appropriateness to ESL students (IF↔ESL Sts).**

As it can be observed in the table from appendix U, the results for the modality of writing in the variable of item format appropriateness to ESL students, all formats contained an approving result for this notion. As verification of this, the item format with the lowest percentile was SA1 (sentence completion – dichotomous – no
picture) with 65%, while the formats with the highest percentiles were MCh (simple sentence stimulus – correct sentence identification) and MCh (sentence completion – speech parts construction) with 90% acceptance.

**Item format & test appropriateness to proficiency level (IF → Prof Lvl).**

Appendix U shows the results for the relationship between suitable proficiency level and item format and test items. Such results also indicate that all item formats were considered to be appropriately constructed, in regards to the proficiency level allocated for each sample test item. For example, the lowest mean in support of this factor was 69% in item formats such as: SA1 (sentence completion – dichotomous – no picture), SA1/2 (sentence structure/single sentence), and SA1/2 (picture stimulus – sentence structure/single sentence). Conversely, the mean with the highest acceptance was 90% in format MCh (picture/diagram stimulus w/sentence completion).

**Task from item format & test item vs. regular instructional tasks (IF/TI ↔ ClassTask).**

Concerning the factor of test tasks matching any instructional tasks used for teachers, the table in appendix U shows that most of the item format tasks were believed to fulfill this characteristic. The only exception came from item format MCh (simple sentence stimulus w/combining sentences), which accounted for a mean of 54% in disapproval of such a feature. Alternatively, the remaining format tasks were believed to be similar to those used in classroom-based settings. The resulting ranges were from 58% to 88% in the following item formats: MCh (picture/diagram stimulus
w/answer questions) and SA1 (picture stimulus – single word/phrase/letter response).

**Need to pre-teach/train students on selected task/item format (IF/TI↔PreTeach).**

As shown in the table from the same appendix U, mean results for the need to pre-teach the item format tasks presented in this study were almost in favor of the need to pre-teach. For example, the item format tasks that resulted in favor with the lowest percentile was SA1/2 (picture stimulus – sentence completion) with 54% acceptance, while the item with the highest percentile of acceptance was Essay (narrative, imaginative, informational, persuasive, descriptive – ER), which accounted for 90%. Opposite to this, the only format that showed a split 50/50 percentile agreement/disagreement was SA1/2 (picture stimulus – single word/phrase/letter response).

**Need to pre-teach/train students test taking procedure on selected task/item format (IF↔PreTeach Proc).**

The table in appendix U demonstrates the results showing the need to practice or pre-teach test takers in the process of test taking as a whole, based on the analysis of the item formats in question. Therefore, participants recommended the test takers be exposed to this process in relation to almost all item formats, with exception to SA1/2 (picture stimulus – single word/phrase/letter response), which contained a 50/50 percentile result in favor (and in disfavor) of applying such activity. Therefore, the item format with the lowest mean approving this notion was SA1/2 (picture stimulus – sentence completion), with a mean of 58% approval, while the format with
the highest percentile was Essay (ER w/picture, diagram stimulus), containing 81%.

**Grade groups most appropriate per item format.**

The table located in appendix V depicts the grade groups that participants believed to be the most fitting for each item format they evaluated.

Concerning the writing modality, the item format with the highest, or most complete, distribution are Short Answer 1 or 2 (picture stimulus – single word/phrase/letter response), in which participants selected all grade levels to be applicable to each of the grade level. Interestingly, the remaining 11 item formats were selected by the teachers to also have a level of acceptability to all the grade levels in question, with the exception of K-1. The lowest mean of these eleven items ranged from 16% up to 31%. Overall, almost all the item formats discriminated appropriately with each grade level (except K-1).

**Quantitative Results (Inferential Statistics)**

**Reliability of the Likert-Scale Questionnaire**

In order to account for the appropriate inferential statistical procedures for the present study, the researcher met and consulted with a psychometrician at the Pearson Assessment & Information Psychometrics Department. The expert research scientist (psychometrician performed G-theory studies and the main researcher was provided with counsel during several meetings, both prior to and after running the statistical procedures.

As previously stated in the methodology chapter, in order to calculate the
reliability of the Likert scale instrument provided to the participants who examined the diverse item formats, generalizability theory (G-theory) principles were used. Reliability is important in this study, as it allowed the researcher to know the consistency to which participants evaluated the item formats through the Likert scale. Hence, the following section presents the results for this modality.

**Reliability of the Writing Likert Scale.**

For the modality of writing, the G-studies were separately conducted for two sets of raters. The first G-study employed the output for the set of 26 raters and the second G-study processed the output for the set of 28 raters. The reliability for the 26 set is 0.9285 and for the 28 set is 0.93127. At this point, it is necessary to clarify that two G-studies were conducted due to the fact that two independent groups of raters evaluated different formats. For instance, the first group (26 participants) of raters evaluated formats 2, 4, 6, 8, 9, and 10 while the second group (28 participants) evaluated formats 1, 3, 5, 7, 11, and 12. The reasoning behind this, as stated in the methodology chapter, was to ease the load of work by having each rater evaluate six formats as opposed to twelve. Finally, based on these two data sets, it is clearly shown that the reliability for both groups was very high, providing the desired consistency of grading amongst all raters.

**Likert Scale Results for the Relationship between the Test and the Construct**

Appendix W illustrates a table with the condensed results between the test items and item formats in relationship to each construct of the test. The variables or constructs included: grammar, vocabulary, morphology, representational,
manipulative, heuristic, and imaginative language functions, as well as cohesion and coherence in writing.

In short, the mean scores represent a combination of the 2 subscales in the Likert scale (agree & strongly agree), as a positive or expected attribute in the results. The results regarding the other subscales (strongly disagree & disagree) were not tallied in the table, since these numbers do not provide illustrative and meaningful data for the interpretation of the argument. In other words, these results represent what the teachers for each construct did not agree had a relationship with the item formats in question. Further, those mean scores shaded in yellow merely represent those formats that did not fulfill the minimum expected requirements from teachers in their analysis.

Therefore, appendix W exemplifies that for the construct of grammar all item formats in writing were deemed to be addressed by all writing formats. The lowest mean (73%) appeared for MC (picture/diagram stimulus w/answer questions) and the highest mean (100%) appeared for the format SA1/2 (picture stimulus – sentence completion). This demonstrates a good correlation between the item formats, this construct, and the teachers’ opinions about such relationships. Other strong formats showing good tasks for assessing grammar included: essay with picture, MC with picture and sentence completion, MC with sentence completion addressing speech parts, MC with correct sentence identification, and SA with picture addressing sentence completion and sentence structure.

The second construct, vocabulary, also resulted in good relationship to the
item formats. The lowest mean was 77% for the format SA1 (sentence completion – dichotomous – no picture) while the format with the highest mean (100%) was SA (picture stimulus – sentence completion). This demonstrated that all formats for writing are suitable for assessing vocabulary, since they are inherently part of grammar.

The third construct consisted of morphology. Results indicated all formats were appropriate to test this skill. For example, the item format with the lowest mean (61%) was MC (picture/diagram stimulus w/answer questions) and the format with the highest mean (93%) was MC (picture/diagram stimulus w/sentence completion). This could be taken as a logical outcome since this specific format calls for completing sentences and teachers believe this activity is a good task for addressing morphology.

The next construct referring to the representational language functions also indicated positive acceptance from teachers. As evidence of this the formats with the lowest mean (69%) were MC (picture/diagram stimulus w/sentence completion) and MC (sentence completion – speech parts construction). On the other hand, the format with the highest mean (96%) consisted of Essay (ER with picture/diagram stimulus). Interestingly, most of the item formats besides the ones mentioned ranged in the upper measures 80% to 90%. This indicates most of the formats have good properties for assessing functions to convey information.

The variable consisting of manipulative functions shows only 5 out of 12 formats were (barely) deemed acceptable to address this construct. Means ranged
between 52% and 65%. The formats included in this collection were: Essay (all genres) [55%], MC (picture/diagram stimulus w/sentence completion) [52%], SA1/2 (sentence structure/single sentence) [56%], SA1/2 (picture stimulus – sentence structure/single sentence) [52%], and SA1/2 (picture stimulus – sentence completion) [65%]. The remaining formats did not fulfill the 50% cut score deemed as acceptable to address this construct (see appendix W).

Similar results were found with heuristic language functions, where 9 out of 12 item formats were considered by teachers to actually address this construct. The lowest acceptable mean was 51% for SA1/2 (sentence structure/single sentence) and the highest mean consisted of 65% for the Essay format (ER with picture diagram stimulus). Basically, the accepted item formats, as noted in the table, were those tasks where students needed to complete sentences, combine sentences, use contextual clues based on the sentences such as pictures, etc.

In terms of imaginative language functions, only 3 formats out of 12 were regarded as acceptable to assess such functions. These consisted of Essay (all genres) [64%], Essay (ER w/picture/diagram stimulus) [54%], and SA1/2 (sentence structure/single sentence) [63%]. Teachers may be choosing these formats for the ER tasks as they are perceived to allow more room for writing discourse and elaboration, leading to the production of imaginative functions.

The final construct was cohesion and coherence in writing. Interestingly, as show in appendix W almost all formats technically were acceptable to address these two skills. The only exception was format MC (sentence completion/speech parts
construction), which arguably (based on teachers’ perceptions) only taps into the speech part identification, as opposed to manipulating writing discourse to complete or produce a sentence all the way to a set of paragraph as it is the case of the ER tasks. The minimal acceptable mean was 62% for the format SA1 (sentence completion-dichotomous/no picture) and the highest mean was 92% for the format Essay (ER with picture/diagram stimulus).

The following variable, called “task/item format degree of difficulty or ambiguity for the present level of student,” also received positive feedback in all formats. Of course the expectation in the results was that teachers would either strongly disagree or disagree with the item. Therefore, all formats were deemed non-ambiguous; neither difficult for test takers. The mean scores in this variable were less powerful, as noted in the table. The lowest mean was 63% for MC (simple sentence stimulus – correct sentence identification) and the highest mean score was 90% for MC (picture/diagram stimulus w/sentence completion).

**Likert Scale Results for the Remaining Variables in the Questionnaire**

As observed in appendix X, the table exemplifies the results for the remaining variables addressed in the Liker scale instrument. Notice that the table reports the summary of those answers in the subscales ‘agree’ and ‘strongly agree’ that merely represent the positive or expected outcomes in the scale. On the other hand, variable B of the two subscales is considered (as expected) to be ‘strongly disagree’ and ‘disagree,’ since the item was tapping at the degree of difficulty or ambiguity to students.
The shaded cells representing those mean scores for the formats do not fulfill the requirements of the teachers’ evaluation. Therefore, regarding the variable “items in the item format assessing the purpose of the test,” all writing formats were considered by teachers to address the purpose of the test purpose. The lowest mean (77%) accounted for the formats MC (picture/diagram stimulus w/answer questions) and SA 1 (sentence completion – dichotomous – no picture) while the highest mean score (93%) accounted for Essay (all genres).

The next variable regarding “test items fitting the item format” was also given positive teacher feedback in all item formats. In fact the overall mean scores were higher when compared to other variables. For example, the format with the lowest mean (77%) was SA1 (sentence completion – dichotomous – no picture) and the formats with the highest mean scores (100%) were Essay (ER w/picture, diagram stimulus) and MC (simple sentence stimulus w/combining sentences).

Concerning the variable “item format comprehensibility,” all formats were deemed understandable for both students and teachers as item writers producing new test items. Most of the item formats had high mean scores. Thus, the lowest mean score was 79% for the formats MC (simple sentence stimulus – correct sentence identification) and SA1/2 (sentence structure/single sentence), while the highest mean score was 97% for the format MC (picture/diagram stimulus w/sentence completion).

In terms of the variable “appropriateness of items and item format for ESL students in the given grade group,” all formats in writing were considered acceptable. As evidence of this, the lowest mean was 73% for MC (sentence completion – speech
parts construction) and the highest mean score was 96% for SA1/2 (picture stimulus – single word/phrase/letter response).

Concerning the variable “items and item format fitting the selected proficiency level,” all item formats received positive criticism from the participants. However, the mean results were lower, as compared to other variables. For example, the highest mean was 86% for MC (picture/diagram stimulus w/sentence completion) and the lowest mean score was 59% for MC (simple sentence stimulus – correct sentence identification).

Regarding the variable of “items and format to be applicable and practical to use at this particular grade level,” all formats fulfilled teachers’ positive acceptance with most of the formats having high mean scores. For example, the format with the lowest mean accounted for 69% in the format MC (sentence completion-speech parts construction) and the highest mean accounted for 89% in the format Essay (all genres). Nevertheless, most of the item formats oscillated in the upper 80 mean percentiles.

For the variable “the task represented in the format would match teachers’ typical instructional tasks on a daily basis,” most of the formats were deemed acceptable in matching typical instructional activities, with the exception of formats MC (picture/diagram stimulus w/answer questions) and MC (simple sentence stimulus w/combining sentences), with means of 50% and 46% respectively. This suggests that such activities for these two tasks either need to be practiced during regular instruction or removed from the test altogether. On the other hand, for the
remaining item formats that correlated with instructional activities the lowest mean was 58% for Essay (ER with picture, diagram stimulus) and SA1 (sentence completion – dichotomous – no picture). The highest mean was 81% for the format SA1/2 (picture stimulus – single word/phrase/letter response).

For the variable “need to pre-teach items/tasks from the item format to a given group of test takers,” as noted in the table in appendix X, all formats were strongly considered by teachers to be explicitly taught to test takers due to the test takers unfamiliarity with these tasks. Perhaps the only format that could have been considered acceptable based on mean scores (50%) was SA1/2 (picture stimulus – single word/phrase/letter response), yet the scores represent the cut off point. In addition, most of the formats oscillated between 61% and 86% of agreement for teachers to pre-teach such tasks.

Similar results were perceived in the variable “need to pre-teach the procedure of test taking to a given group of test-takers.” Again the only format that perhaps would have been acceptable not to pre-teach was SA 1/2 (picture stimulus – single word/phrase/letter response). Again the remaining formats ranged in acceptance to be pre-taught from 61% to 81%, as noted in appendix X.

The final variable, called “the appropriateness of items and item format in age and cognitive level of students at a particular grade band,” received positive expected judgment from teachers. Thus, all formats received acceptable mean scores ranging from 73%, the lowest at MC (simple sentence stimulus – correct sentence identification) and 92% the highest for the format SA1/2 (picture stimulus – single
Qualitative Results

In order to interpret the qualitative results from the analysis provided by the educators, table 8 makes reference to the variables that were used for the current study and the appropriate codes used.

Table 8. Coding Format for each of the Variables found in the Open Ended Questionnaire.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gram</td>
<td>Grammar</td>
</tr>
<tr>
<td>Voc</td>
<td>Vocabulary</td>
</tr>
<tr>
<td>Morp</td>
<td>Morphology</td>
</tr>
<tr>
<td>RLF</td>
<td>Representational Language Functions</td>
</tr>
<tr>
<td>MLF</td>
<td>Manipulative Language Functions</td>
</tr>
<tr>
<td>HLF</td>
<td>Heuristic Language Functions</td>
</tr>
<tr>
<td>ILF</td>
<td>Imaginative Language Functions</td>
</tr>
<tr>
<td>C&amp;C</td>
<td>Cohesion and Coherence</td>
</tr>
<tr>
<td>IF diff</td>
<td>Item format difficulty for grade level</td>
</tr>
<tr>
<td>IF/TI↔Relationship</td>
<td>Item Format &amp; Test Item Relationship</td>
</tr>
<tr>
<td>IF↔Comprehensibility</td>
<td>Item Format Comprehensibility / Generative to Similar Test Items</td>
</tr>
<tr>
<td>IF↔ESL Sts</td>
<td>Item Format Appropriateness to ESL students</td>
</tr>
<tr>
<td>IF→Prof Lvl</td>
<td>Item Format &amp; test appropriateness to proficiency level</td>
</tr>
<tr>
<td>IF/TI↔ClassTask</td>
<td>Task from Item Format &amp; Test Item vs. Regular Instructional Tasks</td>
</tr>
<tr>
<td>IF/TI↔PreTeach</td>
<td>Need to Pre-Teach / Train Students on Selected Task / Item Format</td>
</tr>
<tr>
<td>IF↔PreTeach Proc</td>
<td>Need to Pre-Teach / Train Students Test Taking Procedure on Selected Task / Item Format</td>
</tr>
</tbody>
</table>

Based on the open-ended questions provided in the questionnaire, responses were transferred into data displays. Due to the fact that the amount of quantitative data was enormous because of the scope of the present study, the researcher applied a three-step process to data coding, reduction, and display of results. The first step consisted of transferring the data into data displays by item format, grade group (k-1, 2-3, 4-5, 6-8 & 9-12), and the data were classified into having a positive or negative outcome.
The second step, or cut-data process, consisted of transferring only the positive outcomes into a second set of data displays and reducing many of the teachers’ comments to reflect more core comments by deleting unnecessary information. Such data displays were not included in the present study since they provided an enormous amount of data too large for the scope of this thesis. However, this information provides valuable information to a test developer in terms of constructing item test specifications for the item formats presented in this study. Due to the scope of this study, it would be overwhelming trying to discuss all the comments presented by participants across 36 different item formats evaluated or even 12 item formats for the writing modality. For this reason, further discussion that includes both quantitative and qualitative data into one is presented in chapter 5.

The third step in the process and last cut-data process consisted of creating a data set or data map in order to guide the researcher (or in another context, a test developer) in creating sample item test specifications based on order to re-define reverse engineering process to item test specifications.

**Final Writing Data Display**

The table in appendix R of this thesis illustrates the item formats that received commentaries from participants in the modality of writing. In the relationship between the item formats and the construct of Grammar, almost all formats were considered to be suitable to test such a construct. The exception was the format SA1 (sentence completion with no picture), which did not receive any comments at all. As observed in the outline below, participants believe that when students produce written
discourse through an essay grammatical features would be addressed. The same conclusion was reached for the short answer items producing single sentences, completing sentences, producing words, and even letters, as all of these tasks were seen as testing several aspects of grammar. Finally, the multiple choice items that the comments showed were suitable tasks for addressing grammar, consisted of combination of sentences (addressing syntax), identifying correct sentences (mechanics), completing sentences with speech parts, and addressing questions,

Similar to Grammar, the construct of Vocabulary also tapped in on the same item formats and left out the format SA1 (sentence completion with no picture). As observed in the table below, the recommended grade levels varied amongst the formats, but most of them started at grade 2, as was the case with Grammar. Teachers tended to choose to comment on these formats due to the fact that vocabulary goes hand-in-hand with the grammatical aspects of language.

Interestingly, the selection of item formats for the construct of Morphology was rather similar to the selection of formats for both Grammar and Vocabulary (discussed above). The teachers recommended starting at grade 2, and in some cases at grades 6-12. Perhaps the most suitable items in the minds of the teachers for addressing morphology were those requiring the completion of a sentence or a word, and of course those where larger pieces of discourse needed to happen, as in the case of producing an essay (a more holistic approach to assess grammar, vocabulary, morphology, etc.).

On the subject of Representational language functions, only 10 formats out of
a total of 12 received comments from the teachers as the most prominent for
addressing language functions to convey information. Such formats covered the
production of essays (with picture and with the different genres), which logically
seem to be the most salient for conveying information. The other formats consisted of
multiple – choice addressing sentence completion, answering questions, identifying
speech parts, and identifying correct sentences and combining sentences. The short
answer formats included producing single sentences, single words and letters, and
sentence completion. These tasks in turn were considered by participants as an act to
convey information in one way or another. The target grade levels in most of the
formats were clustered mainly in grades 2-3 and grades 6-12.

For the construct of Manipulative language functions only 3 formats received
comments from the participants. These were the essay with multiple genres and the
essays with picture or diagram as stimuli. The other format was the MCh with correct
sentence identification. Thus, teachers believed these formats were suitable to address
manipulative functions, which in turn can be a logical notion while reacting to a
prompt and writing an essay, depending on the nature of the prompt.

The next construct, Heuristic language functions, was addressed in half of the
12 formats in both formats addressing essays, MCh addressing sentence completion
and answering questions, and the SA addressing the construction of single sentences,
and producing single words, phrase, or letters as a response. Based on teacher input
on these formats, they can be expected to tap in on heuristic functions (or functions to
learn about language itself) when students are creating a writing piece, and even when
completing a sentence, a word, a phrase or a letter to address the task, as these activities seem to pinpoint meta-linguistic features.

In regards to Imaginative language functions, only 5 formats were selected by participants to address these functions. The selected formats were the ones to create an essay (picture based and several genre based) and short answer addressing sentence completion, creating single sentences and creating sentences with pictures as aid. As observed, teachers believed these item formats would allow the use of imagination to produce written discourse in writing in the case of the essays, and also to come up with appropriate sentences to complete the task in the case of the short answer item formats.

The final construct in this set of variables, and perhaps the most prominent for the writing modality, is Cohesion and Coherence. As a result of this prominence, participants commented that 11 out of 12 item formats as good tasks to address cohesion and coherence in writing. The only exception was the format addressing MCh items with sentence completion through the selection of speech parts. This is perhaps because teachers believed this particular task involved testing speech parts as a discrete skill as opposed to the rest of the formats where examinees have to actually produce discourse, sentences, and phrases, or even complete them. This in turn suggests the use of appropriate cohesion and coherence in writing.

In the next set of variables, item format difficulty was regarded to be appropriate in 9 out of 12 item formats. The selected formats were: essay (various genres and picture stimulus based), MCh with picture and sentence completion,
speech parts construction, correct sentence identification, SA2 with sentence structure for single sentences, picture stimulus with single word/phrase/letter responses, picture stimulus with single sentence and picture stimulus with sentence completion. The selected grades ranged from K-1 and 4-12 in some cases.

The notion of item format and test item relationship was considered mostly as positive in all item formats, suggesting that the sample items were appropriately constructed from the instructions provided for analysis to teachers. The recommended grades for these item formats were deemed acceptable for most all levels (K-12).

In regards to item format comprehensibility, 10 formats were deemed as clearly comprehensible for test takers. The only exceptions were MCh picture stimulus with answer questions and SA2 picture stimulus single word/phrase letter response. Grades identified were clustered from K-5, K-8, and 6-12, respectively.

For the next variable pertaining to item format relationship to ESL students, 11 out of 12 formats were regarded as suitable for ELL test takers. The exception in was SA1 sentence completion with no picture. Grade levels recommendations mainly oscillated between K-3, 2-3, and K-12.

In terms of the appropriateness of language proficiency levels to the sample items analyzed, teachers deemed 11 out of the 12 formats as having their items in accordance to the proficiency level selected. The item format that received no comments for this variable was SA1 sentence completion with no picture. As observed in the data display below, the grade levels indicated for this variable were mostly from 2-12, but there were some instances also covering K-1.
The final three variables consisting of the relationship between the item format tasks versus instructional tasks, the need to pre-teach the current tasks from the item formats, and the need to pre-teach the general test taking procedure to examinees received a high level of comments for all item formats. As for the first variable comparing current tasks to instructional tasks, teachers agreed at all levels (K-12) to use tasks in class analogous to the ones presented in this writing modality, including some practice tasks in preparation of testing. The second variable mentioned here also illustrates the need to pre-teach such tasks from K to 12, and in some cases starting from grade 2. Similar results were found in recommending the need to pre-expose test takers to general principles and guidelines for taking standardized assessments.

Conclusion

As observed in the results from the yes/no open ended questionnaire, teachers gave preference to all the item formats in writing in relationship to those constructs that truly convey information through the language as a whole. For instance, formats tapping into grammar, vocabulary, and morphology were highly preferred, since a strong relationship amongst these language features exists. In other words, item formats assessing grammar were intrinsically embedded with vocabulary and morphology. In the same way items tapping into representational, imaginative and heuristic functions were given preference. The reason for this is that these constructs seem to be more manageable for testing, but also they contained a strong relationship
between specific formats. For instance, ER formats are more likely to be testing language functions (imaginative, representational and manipulative). SA and MC formats are likely to tap into heuristic functions, grammar, vocabulary, and even morphology; especially if designed in a sentence completion or fill in the gap. In the same manner, an ER and even a SA2 item format was more suitable for tap into cohesion and coherence, depending on the amount of discourse required by the prompt.

Similar results were observed in the qualitative data analysis. For example, for the construct of grammar, all formats were selected, with the exception of SA1 sentence completion with no picture. Teachers commented on a strong correlation of this construct with the production of tasks such as constructing single sentences, completing sentences, producing words, letters, sentence combination, identifying correct sentences, and speech parts. The same results were observed for vocabulary with the exception of SA1 sentence completion no picture, mainly because participants declared that vocabulary goes hand-in-hand with grammatical aspects of language.

Regarding the construct of morphology, all formats were deemed acceptable to assess this skill, especially if students had to complete sentences, produce essays, which was viewed as a more holistic approach to address grammar, vocabulary, and morphology.

In terms of representational language functions, all formats were acceptable, especially in the production of essays, sentence completion, answering questions,
identifying speech parts, correct sentences, combining sentences, single sentences,
single words, and sentence completion.

For manipulative language functions, the formats suitable to produce these functions were both of the essay formats and MC correct sentence identification (students reacting to a prompt). Similarly, heuristic language functions were highly correlated to sentence completion, answering questions, construction of single sentences, producing single words, phrases or letters. In other words, the production of metalinguistic skills was the main point of assessment.

In regards to imaginative language functions the selected formats were the essay formats, SA completion, creating single sentences, and sentences with picture aid. Teachers believed these were the best tasks to trigger students’ imaginative use of language.

Finally, the construct of cohesion and coherence was, not surprisingly, selected in all formats in this modality, given the properties of writing as a discourse. The only exception was MC with sentence completion of speech parts, because teachers believed testing speech parts was a more discrete skill vs. producing discourse, sentences, or phrases.
CHAPTER 5
DISCUSSION

Putting it All Together (Quantitative and Qualitative Data Results)

After having disclosed both the quantitative and qualitative data results in the previous chapter, my purpose in this section is to bring it all together, and present a framework that can be applied for future development, mainly establishing a process to be used potentially in future large scale assessment initiatives. In doing so, I will draw a more detailed discussion on the purposes of the item formats themselves and on some of the qualitative data coming from the data displays (see appendix R for data display sample and appendices F-Q for detailed descriptions on each item formats used including sample items). Such discussions will allow the discovery of the possible reasons why the results showed up to be into this manner.

Addressing the Research Questions

The next section will address in detail the following research questions as initially proposed in the literature review chapter.

- What is the relationship between teachers’ evaluation of item formats & test items and the original test construct and ELP standards?
- What is the relationship between teachers’ evaluation of writing item formats & test items and the original test construct and ELP standards? (Test constructs covering: grammar, vocabulary, morphology, representational,
manipulative, heuristic, imaginative language functions, and cohesion & coherence).

Writing Modality

*The construct of grammar in writing.*

Both the quantitative and qualitative data results were very promising in regards to considering the writing item formats that tap into Grammar as a construct. All the formats from the quantitative data results were accepted, since the mean scores ranged from 69% for the lowest to 100% for the highest. Similar result was attained in the data displays, with the exception of the format SA2 (sentence completion-dichotomous-no picture).

Now if we analyze the formats from the initial theoretical perspective in which they were created, the ER formats required students to produce an essay at different genres, and to pay careful attention to the writing process; in some cases also pictures are used to aid in this process. Thus, since these two item formats (ER – essay with several genres and ER – essay with picture/diagram stimulus) require extensive discourse production, it is evident that grammatical features need to be considered. The next set of formats are Multiple-Choice based where students are required to identify parts of speech in a blank segment of a sentence, put together two or three sentences into one, or to identify the correct grammatical sentence. Thus, we can certainly agree that these formats are tapping into grammatical features and syntax while combining sentences. The final set of item formats are Short Answer
with a maximum awarded score of two points. Basically the emphasized criterions in these formats tap into completing a sentence with the correct grammatical fragment, identifying the correct speech part, producing an accurate phrase or sentence, or even using the appropriate time frames given the structure of the prompt or sentence to be completed. So as previously discussed, these formats really focus on grammatical features of language, such as vocabulary, time frames, and syntax.

Now, to make the argument even stronger, some of the commentaries identified in the data displays postulated that grammar was an essential construct as part of the writing process. Further, it was stated that grammar was inherently perceived as part of the expected response from the prompt, as it allows students to generate a number of sentences. Others indicated the presence of grammatical attributes to most syntactic intelligibility in the written discourse provided.

Additional skills that were identified consisted of the uses of tenses, construction of paragraphs (as used in the state test), construction of sentences, proper structuring, word and sentence order, ideal testing of grammar for the upper grades, conjugations, inflections in verbs, parts of speech, prepositions, tenses, recognition of syntactic structures, morphology, syntax, negative sentences, conjunctions, relations between subject and predicate, sentence fluency, comprehension of all areas in a writing mode, comprehension of sentences and questions, uses of personal pronouns versus object pronouns versus pronominals, etc.

In conclusion, I maintain that all of these item formats presented are suitable for testing grammatical competence, including syntax and speech parts. The wide
range of item formats allows for this, since for the advanced language proficiency students the ER formats are ideal to get a considerable amount of discourse. Then for the intermediate proficiency levels and middle grades, the SA2 item formats allow for examinees to produce from one to two sentences (maybe more), and yet permit the ability to check their understanding of grammar. Finally, the MCh formats permit us to tap into grammatical features in a more controlled and discrete way, since no need of a rubric is necessary; such a format also provides the choices and allows low proficiency students and the low grades to comfortably address these items at their own level.

*The construct of vocabulary in writing.*

As it can be observed in the quantitative data results, all formats were deemed to be appropriate to test vocabulary. The mean ranges were between 73% and 100%. Similarly, qualitative data results indicated that all item formats address this construct with exception of SA2 (sentence completion – dichotomous – no picture).

Thus, if we look at the commentaries from the data display, these make a stronger argument for such formats testing vocabulary. For example, some teachers postulate that vocabulary is essential, as it is part of the current state assessment. Moreover, test takers can demonstrate an understanding of daily activities by naming them, and test takers can express themselves accurately since they can use the knowledge of vocabulary to provide a base from which to write. Other comments indicated that the use of vocabulary as observed in the item formats shows better command of language, because one can accurately comprehend more of what one
individual wants to say. In some of the formats, test takers need to know specific vocabulary to answer and write the expected essay from the item format, and they need a fairly good grasp of vocabulary to answer in-depth. In addition, others indicated that vocabulary allows a range and use for the context used (as proposed in the prompt). It was also recognized that vocabulary is necessary to convey ideas essential for essays, it allows one to find abilities to use language, it allows one to construct sentences, and it is key factor for comprehension. Some other comments indicated the actual ramifications that the use of vocabulary allows. For example, the knowledge of verbs, spatial prepositions, some connecting words such as subordinators/conjunctions, verb inflections and their meaning, vocabulary pertaining to the pictures used, knowledge of vocabulary and grammar altogether to answer correctly, abstract words, vocabulary needed to understand a sentence as the missing lexical item is seen in the blanks, vocabulary of the school, subjects, food, conjunctives, key words, general vocabulary, vocabulary related to relative seizes, comparatives, descriptive vocabulary to explain pictures, personal pronouns, pronominals, letter response recognition, and possessive pronouns, among others.

Finally, in closing the argument regarding the use of this construct, it is evident that the use of lexical items is necessary to construct small sentences that become more sophisticated complex sentences, leading to paragraphs which form essays. In addition the multiple choice items are designed to identify parts of speech, which in turn require the knowledge and understanding of these terms to identify such lexical items. In short, we can argue that in several cases all item formats are
necessary and suitable to address vocabulary along with grammar.

*The construct of morphology in writing.*

The criterion of Morphology was deemed acceptable to be tested in all item formats (as observed from the quantitative data results). The mean range was 62% and 90%. Similar results were attained from the qualitative data results where all item formats were considered appropriate to address this construct, with the exception of the format SA2 (sentence completion – dichotomous – no picture).

In analyzing the qualitative commentaries from the educators some of the findings follow. Teachers declared that through the item formats and the use of morphology, students can analyze the mastery of syntax, grammatical structures at this level, while others indicated this was an essential skill to be tested, as it is part of the state writing assessment. It was also declared that the presence of morphology contributes to the intelligibility in writing. Morphology was also considered as part of the construct in the item format, necessary as part of the format process, since students need morphology to learn words.

Included in the specific skills detected in the data displays are: past tense inflection, past tense morphology less effective than in the essay formats, present, singular and plural of nouns (e.g. party vs. parties), choosing the correct structural differentiation from words, word endings, morphemes, suffixes, morphological particles embedded in sentences, morph-syntactic features, morphological particles denoting explicit meanings, inflections for comparatives, verb tense inflection, inflections affected by syntax, knowing the appropriate words and plugging them into
the blank appropriately, and inflected possessive forms as morphological tested features.

Other commentaries indicated that the construct could be embedded in the MCh choices provided as they recognized the effectiveness, validity, and speed in testing and scoring this construct through the MCh formats. Now in a more practical way, if we look at the original design of the item formats, those which require an extended response format in fact require the use of morphology as it relates to the proper use of words (inflected) as they are related to the overall sentences, then paragraphs, and lastly, an essay. Despite the fact that this construct may seem rather minimal as compared to the entire discourse being composed in an essay, it is very important to the cohesion and coherence of the entire essay, and will be noted by the scorers in the rubrics.

The second set of formats (SA2) is more suitable in targeting the construct of morphology in a more discrete manner, since a given sentence or prompt can specifically target these inflections. Finally, this is the most suitable and fitting format to address morphology is in the MCh formats since these address grammar and vocabulary. Criterions tested are inflections of verbs (tenses), nouns (singular/plural), adjectives (comparative and superlative forms), and modifiers (adverbs). Thus, this construct can be more discretely targeted in the formats if the test developer designs items tapping into the criterions listed above.

*The construct of representational language functions in writing.*

The results from the quantitative results indicate that all item formats for
writing were considered to be suitable to assess language functions to convey information; means ranged from 72%-97%. Interestingly, similar results were observed from the qualitative data, where 10 out of 12 formats were considered to address Representational functions, with exception of the formats Essay (ER w/ picture, diagram stimulus), SA1/2 (picture stimulus – sentence completion) and SA1 (sentence completion – dichotomous – no picture).

Returning to the commentaries from the data displays to support such results, several functions were detected by participants. For example, the identified functions were: creating information thru writing, conveying questions and information, generalizing activities & sequencing, answering questions, conveying a story, describing people, places & things, comparing and contrasting, conveying information to construct arguments, explaining, drawing conclusions, generalizing, describing a process, describing spatial & temporal relations, deciding what sentence better describes the stimulus or picture, choosing the appropriate/different way of expressing information, using junctives but not for essential meaning, learning conjunctives that convey information, choosing the correct construction of sentences to produce information, conveying likes, dislikes & actions, description of sizes and comparing them, using comparatives, relationships of things & people with descriptions, addressing narrative and expository discourse, expressing & supporting opinions, producing information phrased in students’ own words (circumlocution), describing pictures, synthesizing information, talking about a picture, indicating possession, and establishing a discussion about a picture.
Other comments indicated that these types of functions are at the heart of an essay, which in turn makes writing extremely necessary to convey any kind of information. It was pointed out that the presence of these functions for these formats is the same as used in the writing state test (e.g. using information, evaluating, predicting, etc.).

Finally, from the perspective from which the item formats were designed, the functions mentioned above make sense in that the ER item formats require test takers to produce an essay or a longer piece of discourse that evolved around a topic and genre. Such genres can be from simple descriptions or narrations to more sophisticated tasks, such as comparing and contrasting, defending an argument, creating a persuasion, etc. The rest of the item formats, SA2 and MCh, directly tap into more discrete skills, such as grammar, vocabulary, morphology, in which yet another task of conveying information is taking place (naming, identifying, choosing the right speech part or vocabulary piece or even combining a sentence). In short, all these writing formats have the properties to produce information whether if it is simply identifying a speech part all the way to producing a cohesive and coherent writing piece of discourse.

The construct of manipulative language functions in writing.

Qualitative data results indicate that only three item formats were acceptable in assessing Manipulative language functions [Essay (ER – several genres), Essay (ER – w/picture – diagram stimulus), and MCh (simple sentence stimulus-correct sentence identification)]. However, the results from the quantitative data indicated...
more formats to address this construct (overlapping with the ones from the data displays), these were SA1/2 (sentence structure/single sentence), SA1/2 (picture stimulus – sentence structure/single sentence), and SA1/2 (picture stimulus – sentence completion).

Therefore, if we rely on the qualitative data results, these are the commentaries in support of the formats addressing Manipulative language functions: suggesting & persuading, suggesting a course of action, interpretation involved as conversation and socialization inferred (essay formats), and interactional functions (interpersonal: socializing, greeting, introducing, meeting, leave taking (MC correct sentence identification)).

Secondly, looking at the formats chosen in the quantitative data results from the perspective in which they were created, the first two selected formats refer to producing an essay where it is clearly depicted to produce a persuasive piece (if the prompt requires). The next format called MC (simple sentence stimulus/correct sentence identification) was initially created to identify the correct meaning of a sentence, sentence construction, meaning, or inference. Thus, from the production of Manipulative language functions per se, I do not see a true connection as existed in the essay format. However, teachers relied on these in sense that if some of the stimulus of the format (2-3 sentences) were focused onto a communicative aspect (e.g. if someone asks “When doe she leave?”), then students needed to choose the appropriate choice that goes with the given prompt. What this means is that the test item somehow was tapping into the interpersonal/socializing subset of Manipulative
language functions. Consequently, my take on this is that it is possible to design explicit items that tap into the identification of these functions (socialization), as a subset of the item format structure. However, from a more purist perspective of writing test items, the item format that exists seems to be more suitable for reading comprehension as opposed to the writing construct. The item format SA1/2 (sentence structure/single sentence) was primarily designed to describe something about a specific word, item, or situation through the use of a sentence or set of sentences (depending on the prompt).

Test takers were supposed to compare and contrast a particular aspect of something, given the background information as depicted in the stem (prompt). Therefore, this specific attribute was what made teachers think that this format tapped into Manipulative language functions, in the sense that if comparing and contrasting was needed then test takers needed to make use of some of these functions.

For the last two item formats, SA1/2 (picture stimulus – sentence structure/single sentence) and SA1/2 (picture stimulus – sentence completion), the first format requires students to explain and describe past experiences or future events based on the set of pictures presented, and the second format involves completing a sentence with a missing blank located at the beginning, middle or end of such sentence, and a picture may be used to provide background information. Given the nature of the first item format, I would argue it is more plausible to produce Manipulative language functions if test takers are required to explain or describe past experiences or future events. However, the projected responses seem to be purely
descriptive, and I would argue that there is not enough room to use manipulative functions for even the subset dealing with social skills. The same argument is posed for the second item format where students merely need to complete a missing section of a sentence, since the only logical criterions to elicit are grammatical features of language, but not functional language per se. In these two cases, it is possible that participants needed more training and more understanding of what manipulative language functions look like and under what circumstances these could be operationalized. Despite the fact that these two item formats were selected quantitatively by participants, my argument derives from that fact that there exists limited to no evidence to support the notion that such formats address manipulative language functions.

*The construct of heuristic language functions in writing.*

Quantitative data results indicated that all item formats in the writing modality are suitable for tapping into language functions to learn about language. Mean scores range from 55% for the lowest up to 65% for the highest. Now only 6 item formats overlapped with the results found in the data displays: Essay (all genres), Essay (w/ picture/diagram stimulus), MC (picture/diagram stimulus w/sentence completion), MC (picture/diagram stimulus w/answer questions), SA1/2 (sentence structure/single sentence), and SA1/2 (picture stimulus – single word/phrase/letter response).

Therefore, if we consider the commentaries proposed in the qualitative data results, educators postulated the writing process itself to be a heuristic function since test takers are manipulating the language in their writing to make sense out of it.
Educators also considered organization of thoughts in writing, planning, and organizing/composing an essay as the same sort of functions. Others indicated the use of practice with correction rewriting to be very critical, since students can look up at variables needed to complete their writing. Another commentary indicated this was a more inherent use of Heuristic functions if students were allowed reference to access material, which would make it a more untainted definition of this construct.

Amongst other functions identified by participants were: finding the way about how to say something, finding the meaning of words in dictionaries (if allowed), planning to write, organize & rewrite the writing piece, using meta-linguistic activities explicitly, use of phonics & prepositions, demonstrating the meaning of sentences through demonstration of comprehension, using picture cues, figuring out syntax and vocabulary, demonstrating the understanding and meaning of sentences, comparison and contrasting, addressing hypothetical prompts (e.g. what would you do…), learning and applying how to make a sentence flow, conveying information using students’ own words, which is a more concrete use of the tasks if students were allowed to go back in source reading, tasks allowing for proper spelling, word construction (no matter if letters are to be learned first), etc.

Finally, considering the initial characteristics from which the formats were created, I argue that all the tasks have the room to allow addressing Heuristic functions. As evidence of this, the ER formats (essay all genres and with pictures as aid) simply require students to produce an essay or an extended piece of writing in addition to other parts of the writing process. We all know that composing a writing
piece requires thought, organization, drafting, editing, and checking for cohesion and coherence to achieve the desired final result. These activities all involve the use of language functions to learn about language and going through the process. The second set of formats are structured as MCh tasks, whether the construct is completing a sentence, phrase, or a word in the correct form, or whether the construct is selecting the right answer based on the meaning of the question, plugging in the appropriate speech part into the sentence gap, combining two or more sentences into one (manipulating syntax), or identifying the correct sentence, students are employing Heuristic language functions in a more structured and discrete manner through their understanding of language and the surrounding context.

The final set of formats structured in a SA – awarding one or two points to the right answer – possesses a construct is similar or the same as the ones used for MC (sentence completion, combining sentences, producing a sentence), but differing in the sense that students have to actually write the discourse to perform the task. Yet they are employing the use of Heuristic language functions because they are using the context around the incomplete sentence or the prompt to produce the answer in a more micro-level use of language.

*The construct of imaginative language functions in writing.*

Only three formats were considered as acceptable to address Imaginative language functions from the quantitative data results: Essay (all genres), Essay (w/picture/diagram stimulus), and SA1/2 (sentence structure/single sentence). Concomitantly, the same formats were deemed suitable to test Imaginative functions
in the qualitative data, but also for SA1 (sentence completion/dichotomous – no picture), and SA1/2 (picture stimulus – sentence structure/single sentence).

Having mentioned this, let us look at the commentaries from the teachers in the qualitative data results. In the analysis of the format Essay (all genres), participants supported the use of communication in general, since the test takers have to imagine a context and come up with a description of it. Similarly, in such use of communication the use of writing itself was seen as a vehicle of imaginative language functions, but also in the use of personal experiences in writing. Others indicated that given the structure of the prompt, the use of imaginative responses can be generated if test takers use vivid vocabulary. A similar comment indicated the item format to generate imaginative functions in the responses if the topics are compelling to the students, but also when these sorts of topics are geared toward higher levels of proficiency.

The comments on the format Essay (w/picture, diagram stimulus) pointed toward the fact that students were telling a story using the elements necessary to use Imaginative functions as they were creating such a story. Further comments mentioned that such writing about a picture can be an enjoyable activity to test takers. Another comment declared the presence of Imaginative functions dependent upon how the story is constructed and what the prompt is asking. A similar view was seen in the format SA1 (sentence completion – dichotomous – no picture), where it was stated that given the structure of the task, creative students would find this entertaining and enjoyable. Likewise, in the format SA1/2 (sentence structure/single
sentence) it was referenced that for these functions, as with the present item type, examinees could be creative in their production of sentences. The final commentaries for the format SA1/2 (picture stimulus – sentence structure/single sentence) pointed out a possible point of assessment of these functions in terms of sounds and patterns of words; another comment declared that explaining a picture could be an entertaining way for students to use their language skills.

**The construct of cohesion & coherence in writing.**

As observed in both the quantitative and qualitative data results, all the item formats for writing were deemed suitable to assess cohesion and coherence in this modality. Only one format (MC – sentence completion – speech parts construction) did not receive comments in the data displays, yet it was selected in the quantitative results with a mean of 66% of acceptance. The lowest mean score for a given format was 62% and the highest score was 96%.

In order to show strong evidence from these formats addressing cohesion and coherence, I will do so by briefly reflecting upon the initial purpose of selecting them for the assessment. For the ER formats, students are expected to produce an essay or an extended piece of writing discourse which will be scored through a rating scale containing several attributes that tap into the writing process. In addition, test takers are supposed to produce a writing piece in any of the given genres (depending on the prompt), or in another case describe a process or make up a story based on the picture(s) provided. This of course implies having an organized, sequential, grammatical, cohesive, and coherent piece of writing. In other words, these two item
format are the best vehicles to actually elicit and assess students’ ability to use Textual Competence in writing.

The next series of item formats are constructed in a MCh configuration. Some include looking at a picture as a stimulus and others require completing a missing blank in a sentence along with the insertion of the appropriate speech part. Others consist of choosing the proper sentence structure to answer a question or describe a process given a picture, ranging from simple to compound-complex sentences.

Another format requires the ability to combine two or three sentences into one or two in order to properly manipulate syntax. Thus, given the different characteristics and criterions to be tested within these MCh item formats, we can really see all of them tap into cohesion and coherence; perhaps some in a smaller scale than others as compared with the ER formats, yet these formats break down these two skills and test them in a more discrete way.

The last set of item formats is those constructed in Short Answer 1 or 2 points, depending on the complexity of the response. Some are designed to fill in the blanks to complete a sentence or phrase where examinees have to actually write down what the missing word or set of words are. This allows for flexibility in the language since at least two different answers can be correct, depending on the context of the sentence. Of course these tap into writing down the correct speech part. Others require producing a complete sentence or sentences in order to describe something about a specific word, item, or situation. Then, the expected criterions can be the correct use of time frames and aspect (present progressive, simple present, simple
past, simple future, etc.). Other tasks would include responding to a question or writing a statement describing or explaining the features of a picture(s), etc. Therefore, in practical sense these set of formats allow more room and freedom for test takers to produce a phrase, a sentence, or sets of sentences in their own individual way based on their background, understanding, and imagination to respond to the prompts. This way of responding allows more ways to produce writing discourse in a smaller scale than the ER but with more options than the MCh formats, as well as ample opportunities for scoring leading to the assessing of cohesion and coherence attributes of these sentences.

Finally, in order to bring out some of the evidentiary comments from teachers in support of these formats addressing this part of the construct, let us highlight some of the most prominent. For example, observations in the ER formats indicated that the skills assessed in this item format as organization pattern are critical even at K-1 levels. Others indicated this construct is extremely important as it is used in the state writing assessment, since cohesion and coherence it the basis for producing essays. Another comment pointed out that extended writing leads itself to demonstrating competency in academic and basic language skills, but also the necessity to organize an essay with implied language required versus literal language, which is an important element of cohesion and coherence skills.

The comments in the MCh item formats indicated the presence of this criterion in the construction of completed sentences and relationships between parts of them, allowing the identification and proper use and comparison of vocabulary
with a picture as stimulus. In addition, it was postulated that the responses had to be scored with actual writing discourse, as it creates a strong relationship to the knowledge of cohesion and coherence. Moreover, these item formats allow students to be able to recognize the order of events which must be coherent, or it will not be transparent. Similar comments indicated that the MC format allows cohesion and coherence since it permits sentences to flow well without mentioning explicit details; and to the use of comparison and contrasting as an efficacy that supports what is written as a test unity.

The final set of item formats which are constituted by SA1/2 points awarded indicated that the presence of this construct was indeed evident, as the task was designed to confront text unity, especially in those items which would require additional writing discourse (small paragraphs) and work with the flow of the writing. Of course the expected responses would be dependent on the question and how much writing was to be expected in the rubric. Another comment recognized the presence of this construct (if taken as an example from a short answer format in an off-the-shelf-test in the LAS [Language Assessment Scales]), which recognized the current task to be targeting cohesion and coherence. An additional commentary stated that indeed evidence for this construct in the SA item formats was seen even at the level of constructing a single sentence. A final set of observations acknowledged that for these item formats there were varying levels of difficulty and expression. It was also pointed out that if test takers could produce long enough sentences, the evidence of this construct would be stronger, but it was also recognized that the output could have
several levels of complexity and expression.

So to sum up the discussion of this construct, it is clear and evident from the initial design of the item formats and from the comments and quantitative results provided by teachers that all of the item formats in the modality of writing indeed tap into different levels of cohesion and coherence, from a simple single sentence all the way to producing an entire piece of discourse. Finally, reflecting upon all the results in which cohesion and coherence fluctuated across modalities, these formats were indeed the best indicators (based on teacher input) for tapping into such skills, even more so than in the speaking modality.

The Remaining Variables (Discussion)

As the results for the second set of variables were extensively discussed in chapter four, in this section, I will merely discuss the wealth of information and recommendations provided by the educators participating in this study. Hence, my purpose in this section is to discuss how such information can be applied in light of improving test development practices in similar contexts such as what happened to the Oregon ELPA. In order to remember what these variables were they are noted below (with the codes used to identify them), but also the research question they tap into.

Research Question Addressed

The following research question specifically taps into the remaining variables from both the Likert scale and the open-ended questionnaire.
What are the most suitable writing formats to reverse-engineer item specifications based on teacher input from the variables reflected in the instruments of data collection?

1) Grade Groups Most Appropriate to All Item Formats
2) Item Format Difficulty for Grade Level (IF)
3) Item Format & Test Item Relationship (IF/TI↔Relationship)
4) Item Format Comprehensibility/Generative to Similar Test Items
   (IF↔Comprehensibility)
5) Item Format Appropriateness to ESL students (IF↔ESL Sts)
6) Item Format & Test Appropriateness to Proficiency Level (IF→Prof Lvl)
7) Task from Item Format & Test Item vs. Regular Instructional Tasks
   (IF/TI↔ClassTask)
8) Need to Pre-teach/Train Students on Selected Task/Item Format
   (IF/TI↔PreTeach)
9) Need to pre-teach/train students test taking procedure on selected task/item format (IF↔PreTeach Proc)

Grade groups most appropriate for all item formats.

As it was discussed in the results chapter, participants were asked in this study to actually make an educated judgment (based on their experience and analysis of the item formats) as to which grade groups were the most appropriate throughout the different modalities. Thus, this section reflects upon these rationales.
Writing

Interesting results were found in this modality. For example, the only format deemed to be acceptable to all grade levels was SA1/2 (picture stimulus – sentence structure/single sentence), perhaps due to the fact that it included a picture as a visual aid and the expected response was simply to produce a simple sentence (describing such picture). In any case, participants believed such format was applicable.

Correspondingly, the remaining item formats in writing were deemed acceptable for use starting at grade 2. This is rather surprising since the formats dealing with multiple choice structures and the production of one or two sentences (Short Answer) can logically be thought to be suitable for lower grades such as 2-3 and 4-5. However, the Extended Response (essay type) formats yet were deemed by teachers as applicable at grade 2. My assumption from the test developer perspective is that using an extended response format like this in grade 2, and even at grade 3 is a stretch for the cognitive demands of an ELL at such grades. Nevertheless, let us remember that this analysis was conducted by teachers with considerable experience, expertise, and exposure to the grade groups in question. Of course this provides valuable information since, based on their judgment test developers are making final decisions on a standardized assessment. Such input is applicable as teachers are the ones who have first-hand contact with students in the classrooms, and they are the ones driving the curriculum. Now in order to compensate for possible invalid applications of ER formats in younger students, a fair scoring can be applied in the rating scales and appropriate training can be delivered to scorers addressing these
extended pieces of writing discourse (and moving away from the typical essay format).

**Item format difficulty for grade level (IF).**

As noted in the results section, in all four modalities (listening, reading, writing, and speaking) all of the item formats were successfully deemed non-difficult for the grade level in question (with the exception of listening – visual short conversation). So at this point I pose the following question — what can the test developer and the contributing teacher to the assessment development do with such information?

In practical terms, in order to ensure validity of the test items along with the item formats, and eventually in the design of item test specifications, it is critical to have test items be examined by a committee or panel of expert teachers, in order to deduce whether or not test items and item formats are suitable for various grade levels. If this step can be completed then formats can be pinpointed to appropriate grade levels. This practice is usually performed in state committees reviewing the content of items.

**Item format & test item relationship (IF/TI→Relationship).**

The results supported the argument for how the sample test items presented in this study indeed related to the format from which they were produced. All formats in all modalities were considered by teachers as good fits. Therefore, this is a very important characteristic to maintain since it relates to the notion of ‘item/task-fit-to-spec’ (Davidson & Lynch, 2002), where they postulate that it is critical to make sure
test items fit correctly the test specs and the intended constructs.

Thus, this is perhaps one of the most important principles in testing: to make sure the items are testing what they are intended to measure. Based on my own experience dealing with committees of teachers reviewing test items, this important principle is performed, but at a different level. Most of states that are currently developing their own English language proficiency assessments (and all other subject areas) merely have their teacher panels see the relationship between test items and their state standards. This I agree is a crucial factor in validating tests. However, there is not validation from test items to the item writing guidelines (item test specifications) that examines whether standards, constructs, item, and response characteristics are being validated. In fact, not all states even produce writing guidelines (recipes) which are “set in stone,” which would allow their teachers, item writers, or even subcontractors to write suitable test items (as in theory item writing should be test specification-driven).

*Item format comprehensibility / generative to similar test items*

*(IF ↔ Comprehensibility).*

As noted in the results chapter, teachers deemed all item formats and their test items in all modalities as comprehensible to both test takers and item writers, tentatively producing items from these item format descriptions (which in turn, would become item test specifications). This is an important principle to verify since item formats contain specific characteristics that test specific skills. This in turn leads to descriptions of the characteristics and intended constructs for creating item test
specifications which are used by the item writers. Therefore, this has significant implications in the way test developers create item specifications.

If such guidelines are too technical, it can be very difficult for item writers, and even teachers who have not had a lot of experience writing test items, to figure out what information the specifications are seeking. Specifications need to be, at least to a certain degree, straightforward, so that item writers can understand how to create specific items. Concurrently, specifications need to be detailed enough so that no characteristics are overlooked. The results of a good, well-written specification and its degree of comprehensibility to item writers will be reflected in how comprehensible and unambiguous the items are for test takers. This of course goes hand-in-hand with the creation of items and the stem themselves that are straight to the point and avoid extraneous linguistic loads which can be understood by English language learners.

*Item format appropriateness to ESL students (IF↔ESL Sts).*

Similarly, all formats in all modalities were considered by teachers as suitable to be applied to English language learners. Given the fact that we are dealing with an English language proficiency assessment, it is critical to make sure test items are appropriate and sensitive to English language learners. For example, some important issues that both the test developers and item writers need to consider include making sure that test items are free of bias, with a cautious awareness of various sensitivity issues.

Another concept which is closely related to this discussion is the notion of
Universal Design Review (UDR) guidelines. A current trend in the development of large scale assessments concerns the application of Universal Design Reviews which include bias and sensitivity protocols. As evidence of this, test developers, item writers, and teachers can make sure test items and test specifications are carefully written to accommodate ELLs’ needs, in addition to the needs of the entire test taking population.

For example, Bias simply refers to a review of an item and prompt through the eyes of a test-taker for characteristics that would result in an unfair advantage or disadvantage for any group of students (this could be ELLs, learning disability students, other culture and origin students, etc.). Similarly, the notion of Sensitivity simply refers to a review of items and prompts for acceptability through the eyes of non-test takers, parents, and other adults. Effective and appropriate test items should be grade appropriate, free of bias, and engaging to the test takers. In principle, good items need to avoid any content that may be offensive to some people, e.g. contexts that include or imply stereotypes (including, social, racial, ethnic, gender, or religious). Some of the areas in which bias could be found are gender, culture, language, race and ethnicity, economic or social class, and religion.

Some other aspects to consider are in terms of language (applicable to ELL test takers) are more subtle. In principle, a correct response to a question should demonstrate that the student has learned the targeted content or skill, and an incorrect response should show that the student has not learned the targeted content or skill. Frequently, however, English Language Learners — even those who are quite
proficient in English — may not understand certain words and expressions that are common to other students at their grade level. These students may also be unfamiliar with products and cultural practices in the United States. It is important for item writers and reviewers to minimize obstacles that may unintentionally keep students from demonstrating whether they have learned the targeted content or skills. If this is done successfully, the result will not be a less challenging assessment, but rather an assessment that is more accurate in its representation of a curriculum that is rigorous, challenging, and grade-appropriate.

A final set of guidelines to consider is a taboo topics list for item and passage writing, which is a guide that assists writers in avoiding subject matter that would probably cause a passage to be deemed unacceptable. In general, a topic might be unacceptable for any of the following reasons:

1. The topic could evoke unpleasant emotions in the test-takers that might hamper their ability to take the remainder of the test in the optimal frame of mind.

2. The topic is controversial among the adult population and might not be acceptable in a state-mandated testing situation.

3. The topic has been “done to death” in standardized tests or textbooks and is thus overly familiar and/or boring to students.

4. The topic will appear biased against (or toward) some group of people.

Some of these topics may be perfectly acceptable in other contexts, but they do not belong in an assessment. A basal reader, for example, may contain a story
about a child dealing with death; but in such an instance, the teacher has a chance to prepare students before they read the story, and students have the opportunity to talk through their reactions. No such opportunities are available in a testing situation, so we must be much more circumspect in the selection of such topics.

**Item format & test appropriateness to proficiency level (IF→Prof Lvl).**

In regards to this variable, results indicate that all item formats (with the exception of reading – picture stimulus/sentence comprehension) were deemed acceptable in their assignation to proficiency levels based on the sample test items. Basically, this variable is important to consider when creating new test items since typically English language proficiency assessments are based on the state standards, break down a grade level (or a grade band level) into different proficiency levels. Instead of identifying proficiency levels, some states use the other models such as DOK (Depth of Knowledge), which merely assign a level of difficulty to a given item. Such level of difficulty relates to a category of higher-order thinking skills based on the state standards (e.g. summarizing, predicting, making inferences, etc.). Therefore, while evaluating test items, teachers need to be careful of how they empirically assign a given item DOK or a proficiency level. This also has implications in the way item test specifications are written. Test specifications can be designed to create a certain level of test items (tapping into certain language proficiency and depth of knowledge levels). Thus, item writers following such guidelines need to be careful in how they write items, so that they are neither too difficult nor too easy.
Task from item format & test item vs. regular instructional tasks

(IF/TI\rightleftharpoons\text{ClassTask}).

This variable was deemed as acceptable in almost all item formats and all modalities. The only exceptions were in Writing (MC simple sentence stimulus w/ combining sentences). This suggests that tasks such as combining sentences in writing are not typically used in classrooms. However, all other tasks used in the present item formats were found to be related to classroom tasks that educators use.

This information is a rather significant contribution, since in a testing situation, and especially in a large scale assessment such as this one, the use of familiar tasks is critical for the success of test takers. Since students do not have an opportunity to really prepare for a test, having similar classroom-based tasks makes it more meaningful and less stressful for them.

Therefore, I should argue that in order to start developing any assessment, especially of such a large magnitude, it is critical to create an inventory of item formats to select from that are as close as possible to classroom activities. In addition, there comes a point during test development, perhaps after the items are written but before they are ever presented to test takers, which allows teachers in content reviews to verify what those tasks that would be more meaningful to students based on their classroom activities. If the process does not allow for the actual selection of these tasks, then I recommend having teachers make adjustments to test items and test specifications so that the tasks/items are as realistic as possible.
Need to pre-teach/train students on selected task/item format

(IF/TI→PreTeach).

As noted in the results, almost all item formats across all modalities considered the need to provide some sort of pre-teaching prior to taking the assessment. This argument is an “eye-opener” for test developers since it is always assumed that test takers will be ready for the battery of tasks provided in any given standardized assessment. In addition, as shown in the data displays (qualitative data collection), most of the recommendations of this need were based on the fact that culturally, examinees are not familiar with some the tasks used in the U.S. Other global comments indicated the need to pre-teach these tasks so that students would be familiar with the overall test, including a familiarity with the specific needs in the item formats/tasks that students would not have known. It was advocated that by familiarizing students to such tasks succeeds in minimizes their stress during test taking. Other commentaries indicated the need for students have the opportunity to know certain concepts and constructs as part of the test, so that they would be ready.

Consequently, this factor has wider-ranging implications for large scale assessments regarding what to do when the items in a test are considered “secure.” I recommend, especially for the testing companies, the creation of a battery of test items (similar to those used in the actual test) for student practice purposes. Some states allow the “release” of previously-used items for the public to review, and sometimes these items are used for practice purposes. So apart from the security considerations, it is important for examinees to be provided with the opportunity to
engage in prior practice with similar tasks so that they at least have basic test-taking strategies for the test.

In addition, even in instructional tasks and skills as reflected in state standards, there is a big discrepancy and gap, since culturally the way certain concepts are taught in other countries are totally different or simply not used. Some examples of these are the notion of analyzing the parts of a literary piece (climax, plot, raising/falling action, etc.) or even the writing process in order to produce an essay or any other writing piece. So as these cultural discrepancies in both teaching and testing are present, we (test developers, curriculum developers, and teachers) need to be aware of these differences and be ready to address students’ needs.

Need to pre-teach/train students test taking procedure on selected task/item format (IF ↔ Pre Teach Proc).

The final variable consists of the need to pre-teach, in a general way, the process of test taking to ELLs. As observed in the results, all item formats in all modalities (the exception was Writing – ER – with picture/diagram stimulus). Similar to the notion of pre-teaching tasks, almost all educators agreed with the need for teachers to expose students in class to some sort of training regarding test-taking strategies. The main reasons for this assertion include a lack of exposure and familiarity from students, cultural mismatch, lower test anxiety, etc. Further, some teachers even argued for the need to have a special subject for tapping into test-taking strategies, so that students know what to do on both paper-and-pencil and computer-based assessments. As evidence of this, verbatim comments taken from teachers’
evaluations (from the qualitative data) are provided below.

Commentaries are tapping into the variable of pre-teaching the test-taking procedure to students in the 4-5 grade group, for the item format #1 Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response):

(Comment 1) “Familiarize students on diverse test taking procedures such as reference material use, focusing on form, and explicitly pre-teaching this process to lower anxiety and raise scores because not all students possess these skills and background knowledge.”

(Comment 2) “Young learners who aren't yet writing in complete sentences need this pre-teaching aid.”

(Comment 3) “All test takers need to be exposed to the process of testing because everything a student is expected to do, if it is demanding and adequate, needs to be taught.”

(Comment 4) “No need to for this test item to pre-teach process.”

(Comment 5) “Modeling and pre-teaching a skill, but the final goal is to ultimately know what students are capable of doing (avoid students mimicking everything so that they can move beyond formulaic writing).”

As a result of these commentaries and the recommendations posed by teachers, important implications are brought to light for better test development. For example, they provide important suggestions for what to seek when testing students, what item formats to use or avoid when designing the test, but most importantly, the
notion of consensus-based decisions. This permits the test development to be fluid and for the stakeholders at the bottom of the test-development process to have a say in the decisions of the assessment design.

**Limitations of the Study**

The flaws in this study include its reliance on having a small sample representative of what constitutes the entire population of K-12 Oregon ESL teachers. Despite the fact that 30 participants were thought to be able to evaluate item formats per each domain, different opinions and experience were reflected from teachers who specialize in specific grade groups (e.g. K-1, 2-3, 4-5, etc.). According to the Oregon Department of Education data provided in 2004, approximately 124 school districts in Oregon alone serve English language learners. Concurrently, another limitation in the sampling relies on having a set of teachers who are normally distributed in terms of their respective school districts. ESL populations characteristics differ across the state since school districts serve ESL students from different socioeconomic levels, degree of L1 education, cultures, ethnicities, mother tongues, type of ESL program applied to students (ESL Pull out, Transitional, Dual language, etc.), and so forth. These are critical factors to consider while choosing teachers who have this wide range of expertise and experience.
Implications for Future Research and Applications to Future Assessment

Perspectives

Several implications can be anticipated as possible contributions to the field of K-12 ESL and the language testing environment. For example, the results from teacher evaluation of item formats and test items will establish a better understanding and framework for creating better specifications for the Oregon ELP assessment. For example, during the first pilot test for ELPA in 2005, McCall (2004) declared that, overall, students reacted very well to the item formats piloted. Some particular items discriminated negatively but in fact there was not a transparent way to get rid of bad item formats (as it was the purpose of the pilot test). Therefore, this result has implications for conducting a more teacher-oriented judgment of item formats to obtain more detailed information on their structure and tasks presented.

Along with this, further studies in item/task-fit-to-spec can be applied with the future redefined specifications and even teachers can use some of this data to evaluate such specs themselves.

In terms of the processes used in these studies, they can be used as possible framework to design item specifications in similar testing situations, evaluate item formats/tasks, redefine the RE process to spec writing, better document the RE process of item banks, etc.

Further, the more intimate participation of teachers in the evaluation of items, specs, etc. can be more evident as they need to be active participants in the design of large-scale assessments such as this. Creation of panel teachers who become
specialists in their professional judgment of various stages of language testing development can allow more validity to a given test and apply this knowledge in their formative assessment classroom practices.

**Redefining the Process to Reverse-Engineering.**

**Introduction.**

The following section concerns a systematic procedure that can be carried out by test developers, teachers, item writers, and administrators for categorizing test items into appropriate item formats in a large-scale K-12 English language proficiency assessment and even in other contexts where the need to create stable item formats is imminent (e.g., a localized placement or summative assessment for a school district, a school, or a community college). This process opens the window for a careful analysis of test items to reverse-engineer draft item test specifications. Thus, my own learning from this project (Oregon ELPA, the making of this thesis paper), and my first-hand experience as a test developer, has brought out these ideas in an attempt to establish a framework that can potentially help in similar situations for developing language tests, where the only entities presented are item banks and no prior guidelines or specifications for devising these test items exist.

This notion of reverse-engineering affects the last question in the thesis project:

- *What are the most suitable writing formats to reverse-engineer item specifications based on teacher input from the variables reflected in the*
In order to fully understand the purpose of teacher input in this study and how it relates to and affects reverse-engineering, I am discussing its implications and applications based on the results from this project. In addition, I define these arguments as an opportunity to identify a big gap that has not been formally documented in the literature.

**Reverse-engineering.**

As postulated in the Literature chapter of this thesis, Davidson and Lynch (2002) describe the notion of spec-reverse engineering as the creation of a test spec from representative test items or tasks. They claim that not all testing is spec-driven, since “spec-driven” tests are those that have been created from a specification and for which much of the test development, evolution, and maintenance of specs has been created. Davidson (2003) provides the following definition upon spec reverse engineering, “Reverse engineering refers to the production of test specifications when none exist: all we have are test items or tasks, and from those we induce what the blueprint might be” (p. 1). In addition, he postulates that there seem to be four types of reverse engineering:

1. **Straight RE** – the spec seeks to replicate the item or task exactly as it is.
2. **Historical RE** – specs are straight-reverse-engineered across many versions of a test, in order to better understand how the test has changed over time.
3. **Critical RE** – the spec replicates and improves the item or task.
4. **Test Deconstruction** – the spec is used to better explain the contexts (social,
political, curricular, theoretical) in which the test operates. The spec is itself, a “waystage” tool toward a larger analysis (Davidson 2003, p. 1).

Further, Davidson and Lynch (2002) state that Reverse Engineering (RE) allows one to analyze a set of existing test tasks in order to decide how similar they are and then create the spec that could bring them together. Another application from reverse-engineering is toward the need to move from a testing context in which there were not any formal instructions or guidance for test development to another reflecting spec-driven. They mention that even in a spec-driven situation, it could be hard to start creating a test specification effectively for a new test criterion. Davidson and Lynch (2003) declare that reverse engineering is a powerful notion, since it permits the test development team to clarify what they are trying to measure as an integral part of producing a viable spec. Another application from this procedure is the fact that it can be used to critique existing specs and tests. One of the most significant reasons why RE is needed in test development lies in the need to write test items that follow a principled procedure. Following a recipe (so to speak) that allows both test developers and teachers to create items that mirror the target standards and construct is essential. In order to represent this notion, I present the following figure 5 a couple cartoons as a metaphor for a more practical situation.
Therefore, one situation captures the notion of uncertainty and confusion when trying to put together a bicycle. If no set of instructions is available, and we try to make sense of the process based on our limited understanding of the pieces, we might get a rather unrealistic and impractical product. Even if we analyze each element, without the existence of instructions for properly putting together the bike difficulties will arise in the future.

On the other hand, if we have a set of instructions and follow them accordingly, we have a systematic way to construct the bicycle, as we know where every piece fits, resulting in a functional, quality product. Item writing in language testing is the same. If we have a group of teachers who will develop new test items and if we have a “recipe” that clearly outlines each property of each item, target standard and skill, passage length, target levels, grade groups, characteristic and formats of the items, and so forth, then we will have perfectly sound items for the
Consequently, reverse-engineering is a procedure which allows both teachers and test developers reach a deeper understanding of the test items and fosters a dialogue for improving future items. Further, RE allows a deconstructing of all the pieces of the test to provide a recipe for future development. Thus, in order to demonstrate how RE can be executed, the sections below illustrate the experiential outcomes, like the notion of item formats (as executed by teacher judgment in the present study), and present two contexts where RE can be executed based on teacher input.

**Establishing the context and process to spec reverse-engineer.**

Having pointed out some of the important literature and the established gap existing in the language testing process of development, I will further elaborate my argument by providing examples of situations and also using as context some of the existing needs as they were posed on the Oregon ELPA instrument (but also applicable to the expansion into other testing perspectives).

**Context 1.**

In the specific context of the Oregon instrument, an approach to perform both straight reverse-engineering (RE) and critical RE was to be the target. Despite the fact that test items were initially produced, the process was not specification driven and item writers did not possess a guiding document that would allow them to visualize the projected items for this assessment. A taxonomy of item formats (within the present study) compiled by Perea (2004) was used to categorize test items in items formats
according to their structure, commonalties, skills tested, and possible relationship to the item types and construct features.

Teacher facilitators within the Oregon ELPA project were trained by the researcher in order to develop a deeper understanding of the process of analyzing test items. They were provided with the big picture behind designing the blueprint for statistical purposes (item types), background information regarding the set-up of item test specifications (in order to understand the reasoning behind categorizing items into item formats). They were given sample items across skills, which was generated from those item formats, so that they could better understand the structure of an item and the function of an item format. In addition, new item formats were created or adopted and edits were performed regarding the stimuli, stems, foils (also known as choices), proficiency levels, etc. Finally, a system of categorization was implemented. Table 9 below illustrates the final taxonomy of item formats (used in the present study) as captured by item writing facilitators and myself during this project.

*Application of process in context 1 to other testing contexts.*

Thus, in a similar testing development situation, the need to have item writing guidelines for teachers or item writers to create test questions or simply for theoretical informative reasons (let us say for parents and the community to understand the test rationales and test items; e.g., something what typically is called “released items”) is rather essential. Other contexts would entail having to further develop or expand an existing item bank for an existing project, but the lack of specifications or writing guidelines makes it very difficult to visualize what the developmental needs would
be, in terms of refreshing test forms and producing additional field test items that could mimic perhaps an existing core or operational form. In this case, reverse-engineering as a process not only for replication, but also to understand the construct and structure of items and rationales will help in this task. Now having further set some of the stage for the needs in these various contexts, a group of teachers or test developers can start classifying the existing sample test items or an entire item bank into groups of items that share a commonality and characteristic in regards to constructs, standards, item types (e.g. MCh, SA, ER, etc.), lay out, level of proficiency or depth of knowledge, etc. This will be the initial step into creating an item format.

I shall make the distinction between item type and item format to avoid confusion. An item format refers to the main structure of a test item where its purpose is to provide information regarding how many score points can be awarded to the test taker for a particular question. It has been my experience that item type indeed allows the creation (backbone) of a test design or “table of specifications” (Lynch, 2002). So an item type can be framed into dichotomous (0 to 1 score point) [e.g. Multiple Choice] and polytomous (0 to several score points) [e.g. Open Ended or Extended Response, a Short Answer, etc.], from which these are graded on a rating scale or rubric.

Now the item format provides the subcategories of item types which denote specific skills or criterions tested, and even the lay out or delivery for them. For example, a MCh item type could have item formats tapping into literal, interpretative,
or applied levels of reading comprehension (so to speak). But also, a given item format can be tapping into literal comprehension, which could be designed in a sentence completion structure, or with a picture as an aid, and so forth.

Obviously having a strong understanding of what the classification into item formats might be while dealing with an entire item bank is rather critical. This process will not only help classify test items into commonalities, but also will permit an intrinsic understanding of the diverse points of assessment that may be contained (or merely available) in given test for a given state project. Of course, having this understanding will allow one to account for and analyze what sets of skills our item bank contains, vis-à-vis our target construct, standards, and benchmarks (the mandate). This is a relevant step because it will allow us to see if indeed we (test developers and teachers) are meeting our blueprint in terms of standard/benchmark allocation, which has in recent days been quite crucial in standardized large-scale assessment development at all levels beyond ELL tests.

Thus, in short this process of understanding and identifying our item banks is the first step in a more principled manner for reverse-engineering test specifications. Now in terms of who should be doing this task — including those teachers who are involved in the committee reviews in their own states, those teachers who administer their tests, those who write the test items, and those who have the assessment experience — can be an initial cadre of individuals who can train their colleagues at the school district level in better understanding the needs of their students, because these individuals are the ones who who have day-to-day interactions with the test
takers and can comprehend the pedagogical and assessment needs from them.


<table>
<thead>
<tr>
<th>Item formats for WRITING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response)</td>
</tr>
<tr>
<td>Essay (Extended Response w/Picture, Diagram Stimulus)</td>
</tr>
<tr>
<td>Multiple Choice (Picture/diagram stimulus w/Sentence Completion)</td>
</tr>
<tr>
<td>Multiple Choice (Picture/diagram stimulus w/Answer Questions)</td>
</tr>
<tr>
<td>Multiple Choice (Sentence Completion-Speech Parts Construction)</td>
</tr>
<tr>
<td>Multiple Choice (Simple sentence stimulus w/Combining Sentences)</td>
</tr>
<tr>
<td>Multiple Choice (Simple sentence stimulus – correct sentence identification)</td>
</tr>
<tr>
<td>Short Answer 1 (Sentence Completion-Dichotomous-No Picture)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Sentence Structure/Single Sentence)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Picture Stimulus – Single Word/Phrase/Letter Response)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Picture Stimulus - Sentence Structure/Single Sentence)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Picture Stimulus-Sentence Completion)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item formats for LISTENING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answering Questions (De-contextualized)</td>
</tr>
<tr>
<td>Answering Questions about a Passage/Monologue</td>
</tr>
<tr>
<td>Sentence-Completion</td>
</tr>
<tr>
<td>Literal Comprehension/Dialogue</td>
</tr>
<tr>
<td>Pragmatic Inference</td>
</tr>
<tr>
<td>Picture/Literal Linguistic Level</td>
</tr>
<tr>
<td>Visual/Short Conversation</td>
</tr>
<tr>
<td>Picture Identification – Literal Comprehension</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item formats for READING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture Identification</td>
</tr>
<tr>
<td>Picture Prompt, Text Identification</td>
</tr>
<tr>
<td>Passage Comprehension</td>
</tr>
<tr>
<td>Fill in the Missing Gaps/Paragraph</td>
</tr>
<tr>
<td>Context Clues/Expanded Comprehension</td>
</tr>
<tr>
<td>Picture Stimulus/Sentence Completion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item formats for SPEAKING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extended Response (sample of the specified content)–Thinking / Organizing / Producing</td>
</tr>
<tr>
<td>Extended Response (pictures)</td>
</tr>
<tr>
<td>Extended Response (role-play)</td>
</tr>
<tr>
<td>Extended Response (topic of personal interest)</td>
</tr>
<tr>
<td>Extended Response (pictures/picture story)</td>
</tr>
<tr>
<td>Short Answer 2 (responses to audio or video recordings/semi-direct)</td>
</tr>
<tr>
<td>Described Situations</td>
</tr>
<tr>
<td>Simulated Conversation/Response w/Pictures/Video</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Appropriate Responses to Specific Questions in Situations)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Question &amp; Answer/Different Question Types/Cues)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Question &amp; Answer/Different Question Types/Cues/Naming)</td>
</tr>
</tbody>
</table>
**RE items into specs.**

After having discussed the initial step in understanding the items, the item bank, and the potential item formats to be used, I propose the establishment of a systematic procedure to RE items into specs. The following criteria can be pursued so data from items could provide information for how to write specifications. Table 10 depicts these elements.

Table 10. Criteria Extracted from Test Items in order to RE Specifications.

<table>
<thead>
<tr>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics of the stimuli (Oral/written text and/or pictures)</td>
</tr>
<tr>
<td>Characteristics of the stems (Questions, statements or prompts)</td>
</tr>
<tr>
<td>Characteristics of the foils (Distractors or choices)</td>
</tr>
<tr>
<td>Characteristics of the prompt attributes from the items within their item format</td>
</tr>
<tr>
<td>Characteristics of the response attributes from the items within their item format</td>
</tr>
<tr>
<td>Possible assessable ELP standards pertaining this item under its item format</td>
</tr>
<tr>
<td>Possible delivery, collection and scoring criteria</td>
</tr>
<tr>
<td>Coding of the construct of the test from each item</td>
</tr>
<tr>
<td>Coding from each item into a specific item format</td>
</tr>
<tr>
<td>Coding from each item to a specific academic context &amp; language category</td>
</tr>
<tr>
<td>Empirical characteristics of the expected responses from each observed item</td>
</tr>
</tbody>
</table>

After all of the elements mentioned above had been implemented and carefully considered, the next step consists of combining item formats into item types and taking into account their commonalties (e.g., skills tested). In order to make my argument stronger, I draw from the Oregon ELPA context as an illustrative example of how the combination of item formats presented in this study can be achieved.

Thus, I present Tables 11, 12, 13 and 14 to demonstrate the commonalties and differences among these existing item formats, and how those can potentially be used as the foundation to create draft item specifications in similar contexts (reverse-engineering).
The actual percentages presented indicate the actual allocation of items in all four modalities in the Oregon ELPA item bank (approximately 2000 items deemed usable for field testing), which were classified into each one of the item formats presented in this study after a complete taxonomy was made considering the attributes discussed in table 9.
**Possible combination of item formats into actual test specifications.**

### Table 11. Item Format Combination in the Modality of Writing

<table>
<thead>
<tr>
<th>Item formats</th>
<th>Commonalities</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>(21%) Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response)</td>
<td>Essays, covers all genres</td>
<td>Pictures as stimuli</td>
</tr>
<tr>
<td>(5%) Essay (Extended Response w/Picture, Diagram Stimulus)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11%) Multiple Choice (Picture/diagram stimulus w/Sentence Completion)</td>
<td>Both MCh, pictures as stimulus, share skills tested</td>
<td>Sentence completion vs. answering questions</td>
</tr>
<tr>
<td>(2%) Multiple Choice (Picture/diagram stimulus w/Answer Questions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4%) Multiple Choice (Simple sentence stimulus w/Combining Sentences)</td>
<td>Both MCh, simple sentences as stimuli</td>
<td>Combination of sentences into more complex sentences vs. identification of correct sentences from foils</td>
</tr>
<tr>
<td>(1%) Multiple Choice (Simple sentence stimulus – correct sentence identification)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(17%) Multiple Choice (Sentence Completion-Speech Parts Construction)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0%) Short Answer 1 (Sentence Completion-Dichotomous-No Picture)</td>
<td>Both contain SA1, both are structured into sentence completion, share skills tested</td>
<td>One contains SA2, one has pictures as stimuli</td>
</tr>
<tr>
<td>(12%) Short Answer 1 or 2 (Picture Stimulus-Sentence Completion)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1%) Short Answer 1 or 2 (Sentence Structure/Single Sentence)</td>
<td>Both are SA1&amp;2, both focused on sentence structure</td>
<td>Pictures as stimuli</td>
</tr>
<tr>
<td>(12%) Short Answer 1 or 2 (Picture Stimulus - Sentence Structure/Single Sentence)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(14%) Short Answer 1 or 2 (Picture Stimulus – Single Word/Phrase/Letter Response)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Or, an alternative option</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11%) Multiple Choice (Picture/diagram stimulus w/Sentence Completion)</td>
<td>Both MCh, both sentence completion</td>
<td>One possesses picture stimulus, one focuses on speech parts</td>
</tr>
<tr>
<td>(17%) Multiple Choice (Sentence Completion-Speech Parts Construction)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 12. Item Format Combination in the Modality of Reading

<table>
<thead>
<tr>
<th>Item formats</th>
<th>Commonalties</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>(22%) Picture Prompt, Text Identification</td>
<td>Both have text/passages as stimulus in all genres, both are MCh</td>
<td>Pictures as stimulus</td>
</tr>
<tr>
<td>(38%) Passage Comprehension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(15%) Fill in the Missing Gaps/Paragraph</td>
<td>Both are structured in missing gaps/sentence completion format, both are MCh</td>
<td>Pictures as stimulus</td>
</tr>
<tr>
<td>(3%) Picture Stimulus/Sentence Completion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7%) Picture Identification</td>
<td></td>
<td>Contain pictures as foils, targeted for lower levels, could be generated as ‘drag and drop’ item</td>
</tr>
<tr>
<td>(15%) Context Clues / Expanded Comprehension</td>
<td></td>
<td>Has passages as stimulus but focuses on a minimal part of it (context clue)</td>
</tr>
</tbody>
</table>

Table 13. Item Format Combination in the Modality of Listening

<table>
<thead>
<tr>
<th>Item formats</th>
<th>Commonalties</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>(17%) Answer Questions about a Passage/Monologue</td>
<td>Both share similar skills</td>
<td>One possesses spoken passages or monologues as stimulus vs. spoken dialogues</td>
</tr>
<tr>
<td>(9%) Literal Comprehension/Dialogue</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4%) Picture/Literal Linguistic Level</td>
<td>Both have visuals, pictures as stimulus</td>
<td>One may have spoken monologues, another one has short conversations</td>
</tr>
<tr>
<td>(1%) Visual/Short Conversation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8%) Answer Questions (De-contextualized)</td>
<td></td>
<td>Stems or questions come without stimulus spoken narration (out of context), may have pictures as stimulus</td>
</tr>
<tr>
<td>(5%) Sentence-Completion</td>
<td></td>
<td>Formatted into sentence completion structure</td>
</tr>
<tr>
<td>(13%) Pragmatic Inference</td>
<td></td>
<td>Focuses on more inferential comprehension</td>
</tr>
<tr>
<td>(43%) Picture Identification – Literal Comprehension</td>
<td></td>
<td>Contains pictures as foils and address literal comprehension (could be used as ‘drag and drop’ item)</td>
</tr>
</tbody>
</table>
Table 14. Item Format Combination in the Modality of Speaking

<table>
<thead>
<tr>
<th>Speaking</th>
<th>Commonalties</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6%) Extended Response (pictures) (4%) Extended Response</td>
<td>Spoken output (ER), can cover all genres</td>
<td>One has one picture as stimulus, another has series of pictures (picture story)</td>
</tr>
<tr>
<td>(pictures/picture story)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1%) Extended Response (role-play) (8%) Extended Response (topic</td>
<td>Spoken output (ER)</td>
<td>Role-play situation vs. personal interest topic</td>
</tr>
<tr>
<td>of personal interest)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6%) Short Answer 1 or 2 (Question &amp; Answer/Different Question</td>
<td>Both are SA1-2, focuses on responses to questions</td>
<td>Targeted for opposite proficiency levels</td>
</tr>
<tr>
<td>Types/Cues)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(25%) Short Answer 1 or 2 (Question &amp; Answer/Different Question</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types/Cues/Naming)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1%) Described Situations</td>
<td>Three are SA1-2, they have a picture or video as stimulus</td>
<td>Describe Situations vs. using functional language</td>
</tr>
<tr>
<td>(39%) Simulated Conversation/Response w/Pictures/Video</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5%) Short Answer 1 or 2 (Appropriate Responses to Specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions in Situations)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Context 2.**

The point I seek to illustrate through the specific context of the Oregon ELPA is that both the similarities and differences in those item formats potentially combine to allow the convenience of reducing many item formats (presumably items) to expose test takers with in an assessment.

Now, just for the sake of driving some discussion from the table above, let us consider the following two item formats in Writing: Multiple Choice (simple sentence stimulus w/combining sentences) and Multiple Choice (simple sentence stimulus – correct sentence identification). Notice that both are framed in a MCh format and contain a simple sentence as a stimulus. Now notice that their differences is that one
format taps into combining simple sentences into more sophisticated ones while the other format asks test takers to identify the correct sentence from the choices provided. Given this situation, a single test specification can be created and both skills can be included in the point of assessment for such items to be produced.

Let us look at another example in the modality of Listening: picture/literal linguistic level and visual/short conversation. Notice that their similarities lay in the fact that both possess visuals as stimuli and their differences involve having monologues versus conversations in their stimuli. Yet, a single spec can capture these characteristics and can produce a battery of test items to have students analyze monologues or small lectures along with conversations. The range of skill and level can also be incorporated and emphasized from literal to inferential conversation, and make use of visuals for support.

In short, the process employed to combine similar item formats (preferably upon the common skills tested) is essential as a starting point to create specs.

Finally, after these combinations of possible item formats, a preliminary rough draft depicting test specs can be completed in the following domains: reading, listening, writing, and speaking.

While starting to design item test specifications (item writing guidelines) the following information can be included for each spec: a) The name of the item format; b) the characteristics of these formats (e.g., stimuli, stems, foils, pictures, types of text used, genre, topics used, correlation between formats and proficiency levels across the grade bands, etc.); c) prompt attributes of the spec (in this case, the redefined
description/definition of the item format or a detailed description of what the student will do in response to the prompt and what will constitute a failure or success; d) response attribute of the spec (characteristics of the foils, rating scale 0-1, 0-2, 0-3 (collapsing them as observed on the items), observed skills/criterion from each item.; e) general description (included the selected standard, the language function and the linguistic form; and f) the specification supplement (the actual collapsed and unique standards) used during item writing. Of course, some of these elements are a specific of a given assessment or project (such as a given construct [language forms and functions, assessable collapsed standards, etc.]). However, the process can be applied to other ESL assessments.

After having combined similar item formats that could be blended into one single spec, some additional information can be incorporated. 1) Alignment of items in respect to academic content areas and competency areas (as observed in other state test specifications). 2) Add appropriate sample items per proficiency level on each spec. 3) Add information regarding “use cases” in terms of delivery, collection, and scoring, as recommended by the assessment team or teachers involved. 4) Add a template that could potentially mirror/generate expected responses for ER 4-point items since a need for rating scales was used for writing and speaking in further development. 5) Add proficiency level descriptors and suggested time and length for a prompt/stimulus (under speaking, listening and writing).

Application of process in context 2 to other testing contexts.

After having explained the wide-ranging possibilities for creating item test
specifications by combining item formats that share the same characteristics and
criterions tested, and after having experienced the process of creating a large-scale
assessment for an entire state, the general process that follows is suggested as it
mirrors logical steps. Figure 6 below illustrates a descriptive process used to spec-
reverse engineer this particular English language proficiency test (given its
characteristics and needs). Nevertheless, I use it also as a point of reference to
construct a more general process, which also applies to situations where the
development of other ESL assessments (either at the state, school district, or even at a
single-school level) is needed to have item writing guidelines. The components
established in such a diagram reflect the characteristics of good test design and
procedures that apply to all test development. In other words, I make emphasis on
Davidson and Lynch’s (2002) role of test specifications in stages of test development,
where I fill in the gap with the iterative process of RE in order to reach the stage to
“finalize the operational measure.”

As a consequence, based on some process documentation observed from this
figure, and based on the results from the current thesis study and the other
experiences employing this procedure, I propose this as a fitting process for the
development other ESL assessments, which illustrates an approach to identify what
type of RE is necessary to pursue. This allows one to embrace the philosophies of the
test and the thinking of the test developers.

Concurrently, I support the procedure to empirically identify the skills that
diverse test items are supposed to test, in other words, getting in the item writer's
mind and organizing the intended criterion each item assesses. Essentially, item formats can be applied to different targeted skills, but some item formats may be more suitable for such applications. At some point, creating a description of each item format permits one to start defining the prompt attribute that are an integral part of a specification. I rely on Davidson and Lynch’s (2002) model to create item test specifications. These guidelines consent to defining the characteristics of the items as well as the expected behaviors from the examinees. Finally, after extensive conversations among the test developers, item writers, teachers involved in the test construction process, etc. — and having put together the main pieces of a spec — I shall recommend other procedures to ensure validity.

For example, given the experience of this ELP assessment, performing a small pilot test allows for the trying out of item formats and to see how students react to them, so that bad item formats can be eliminated. If most item formats do not discriminate appropriately, gathering anecdotal information from both teachers and students to obtain their impressions is a relevant step. Another strategy is to perform more investigations to allow item/task fit-to-spec (See Davidson & Lynch, 2002, 2003) and inquire as to whether or not the existing test items correlate with the newly-created specifications. At last, executing a qualitative and quantitative teacher judgment evaluation of item formats is critical so that educators can select the best item formats related to their suitability for ESL students, cognitive levels, proficiency levels, instructional tasks, etc. In other words, this last process represents the entire point of this thesis paper: to seek teacher professional judgment to reduce item
formats and gather information regarding what item formats are both theoretically and practically most suitable for different skills and levels of ELL students.

Consequently, this information can better inform the ongoing creation of item test specifications in an alternative way than combining item formats into one test specification by simply having the input of teachers who dictate what the best formats are on the basis of creating item test specifications (merely a process of elimination).

Other important benefits of executing reverse-engineering are to create a test with established item formats that allow testing to be consistent from year to year. This is very important since test takers will be exposed to and eventually familiarized with the item formats used in the test. From the psychometric perspective, reverse-engineering is in fact a content task, and it permits the development of linking items in the core or operational forms for future tests and the ability to have parallel forms from year to year. In addition, reverse engineering allows the benefit of having teachers’ voices in the development of tests. For example, they might affirm certain skills covered in the test are not used in instructional activities, enabling the identification of a mismatch between the test tasks and the curriculum. Such identification is important to capture during test development.

To end this section, I support the idea of promoting spec reverse engineering in an Iterative, Flexible, and Consensus-based fashion (Davidson, 2003), and as a predictive descriptor of a successful product of test specification crafting. Thus, “Iterative” means that the specifications are responsive to feedback from peers’ comments, both from trailing and from operational use. The second principle,
“Flexible” implies specifications that do not get set in stone, but are able to change and adapt to shifting contexts. The third principle, “Consensus-based” refers to the unique benefit of specifications as a cooperative aspect of all interested parties in test development.
Figure 6. Spec Reverse-Engineering Process for an English Language Proficiency Assessment

- Identify the type of Reverse Engineering that will mirror the philosophy and aims of the specifications (Straight RE, Historical RE, Critical RE, & Test Deconstruction)
- Categorize items by item formats (or groups of items that share a commonality in their structure & criterion tested)
- Identify skills represented on each item
- Identify the characteristics of the items (stimuli, stems, foils)
- Identify the description of the item format (potential prompt attribute)
- Identify the attributes for delivery, collection and scoring
- Incorporate the test construct characteristics (e.g. standards, language functions & forms, competency areas, mandate, general description of the skill(s) tested)
- Redefine the item format description
- Define the prompt attribute
- Incorporate and define the response attributes
- Collapse similar item formats into one specification
- Incorporate the proficiency level descriptors to the form of the specification
- Redefine response attributes
- Incorporate specification supplement (including rating scale templates)
- Create an "Specplate" or specification for specifications in order to visualize the test form
- Obtain a consensus in the polishing of the specs
- Pilot specs with item writer facilitators (or a cadre of teachers involved in the assessment development) and observe if new items reflect original items

Yes
- Perform a small pilot test to try out item formats
- Gather anecdotal data on how examinees react to item formats
- Perform survey to allow item/task fit-to-spec process on existing specs by including teachers

No
- Perform qualitative and quantitative teacher judgment evaluation of item formats to select best formats/items to RE (process of elimination)

- Decide and create the item templates for item bank as well as for the technology platform (if computer delivered)
- Design the form of the test vis-à-vis the specifications
- Create a table of specifications (commonly known as test design or blueprint)
- Redefine & correlate the specplates to the original test blueprint, now looking at item types

- Assemble test and pilot trial (commonly known as field test)
- Finalize operational measure
Assembling a sample RE detailed item test specification.

After having discussed the process of executing spec reverse-engineering from single test items (or even just having an entire item bank), in this section I illustrate (based on the context of the Oregon test development and applying some of the results from the data and feedback contained from the evaluative comments on the item formats analyzed in the present study) a sample Reverse-Engineered Item Test Specification. Since it would be extremely overwhelming to include and RE all the item formats based on the feedback from the teachers, only one specification is included containing all the elements necessary to produce such items. Appendix X illustrates this finalized spec.

Notice that in order to compose such a specification, important information was added in order to have an ongoing discussion between the test developers and teachers. This discussion is very important in order to arrive at a mutual understanding of what a given item writer needs to produce in terms of test items. For example, a segment called the “GD,” or General Description, serves as a good preamble to the item writer in identifying the general skills or criterions to be targeted in the spec. In order to have a better illustration of this, the standards (in this given the case the ones used for the Oregon ELP) are included in addition to a brief description of the characteristics of the skills to be assessed. It is important to note that in any English Language Proficiency Assessment a set of pre-determined “assessable standards” should be only used in terms of assessment development.
Those standards which tap into skills that can be observed and evaluated directly by teachers within the classroom settings should not be included in any test since there is not formal process to actually evaluate those skills during a seated-testing situation.

Notice that the PA “prompt attribute” (or a description of what the test taker will encounter) includes the type of item formats that can be used to produce parallel items, and even provide detailed information into what the prompts and items are that need to be included. Information regarding the types of pictures that may used should be included in order to aid the item writer (and the artists who are composing specific art for a picture).

Another critical part of the test specification relies on defining the RP (Response Attribute), which allows the defining of the expected response that the test taker should produce in response to the test item. This is a significant part of the specification, due to the fact that it allows the item writer to have a connection and an illustration about what the expected response might be. Further, it helps the item writer define the boundaries of the structure and the level of complexity that the item should be. In this case an illustrative example of the rating scale (rubric) for this battery of items will be useful.

The next component to include is a Sample Item (SI) because it brings together the entire purpose of the specification by allowing the item writer to see what the examinees will encounter and then use that sample as a starting point. It is also important to include a sample item for each item format if more than one item format
is to be used in a single item test specification. Notice the inclusion of a “context”
segment which merely indicates what the possible topics are to be used when
producing the items. Those could be in different situations, such as describing
animals, school situations, geography (talking about cities), comparing people’s
appearance and personality, health, etc. Finally, given the situation and subject area
topics used (science, social studies, etc.), the appropriate level of vocabulary can be
used to tap into more difficult to less difficult item formats.

The last component is the Specification Supplement (SI), which basically
includes any additional information that the item writers will use in order to create the
test items. In this particular case it included the unique and consolidated Oregon ELP
standard that allow those standards to tap into a single skill, and also those standards
which cover more than one skill (consolidated). Besides the wording of the standards,
note the inclusion of the alignment to the construct, in other words, what criterions
from the construct these standards relate to (grammar, vocabulary, etc.). At last, the
inclusion of which expected sorts of skills that each item format (based on the
standards) will incorporate can be produced. This segment permits the item writer to
have a variety of ideas and skills to choose from and creates a battery of test items
that will tap into different criterions, such as talking about/describing animals, school
situations, geography (talking about cities), comparing people’s appearance and
personality, health, teacher instructions and commands, family situations, favorite
school subjects, sports, types of food, talking about the alphabet, means of
transportation, at the library, remembering past events, talking about present and
future events, school schedules, museums, at the cafeteria, activities performed by people, at the lab, school equipment, at recess, national symbols, the five senses, professions, at the park, etc.

*Creating a specplate (or specification of specifications).*

The previous section has examined a sample reversed-engineered item test specification, which indeed provides details of all the necessary elements that both item writers and test developers need to possess in order to produce sound-fitting, valid, and reliable field test items. Now, considering future applications for item test development (as discussed in the process of RE based on the diagram illustrating such procedure), the next step in the development of items specs is the creation of a “Specplate” or a specification of specifications. This part of the test can be considered by some test developers as a sort of detailed test design for item types. Now my point in bringing up this type of document is that instead of mapping the item formats that should be included in a given test form (whether field test or operational), this document can be used to bring together the use and combination of the “set in stone” item formats that have been evaluated and recommended by teachers and test developers. The document can then serve as a map to guide and connect item test specifications and see how they might correlate to the test design, blueprint, or table of specifications (usually created by the psychometricians). Based on my experience in the development of this project (Oregon ELPA), and based on my exposure to the development of several assessments across various states, I shall maintain the position that an item format test design, or a “specplate,” should be a document designed and
used by test developers who are in charge of content issues (whether ELL, Language Arts, Social Studies, etc.). The rationale behind this lies to the fact that this document aids the content specialist (as it shall also guide teachers in charge of the test development, let’s say at a district level) to map out what item formats need to be used, as well what specifications need to be included in a given form. I shall provide here, a context in which such documentation is not only important, but is indispensable, at all levels of test design.

Let us assume that a new project is underway, and new development needs to begin. After both the test developers and the teachers have executed a careful analysis on what is available in terms or items, item formats, the target test design, blueprints, etc. then all of this information needs to come together in a coordinated manner. Without the presence of a specplate, and with no item test specifications in place, we are at risk of jeopardizing future development if the need of newer items resembling the standards has begun. In this case, without documentation available, I recommend doing an inventory of item formats and then making educated decisions regarding which formats are the most suitable to produce. Thus, at a large scale, or even at a more localized level, performing a RE process or executing any of the small quantitative or qualitative studies to analyze item formats can be undertaken to achieve more validity (as it comes from the main stakeholders). In addition, the specplate can help to map out what item formats will be part of a test spec, and then this document can become the basis for a more stable item format design for content purposes. In other words, if in any testing situation no item format design or specplate
is utilized, a serious risk of exists of not having consistency in the test forms from year to year and from grade to grade. As a rule of thumb, good testing procedures require the existence of parallel forms from year to year in order to have the same constructs as dictated by the standards (as well as for psychometric reasons, e.g. for equating purposes). In addition, it is rather important to have consistency of what item formats are to be used in a test (year to year) since test takers need to have consistency regarding what sort of tasks they are to which they are exposed. Having consistency on item format used also implies having consistency on what items are to be used as “linking” or “anchor,” which indeed is a critical process in order to have one part of the test from given year be represented on the following year’s test. An additional reason for having a specplate is to have an organized process for embedding different item formats in a form. As previously noted, certain item formats can be more suitable to assess different skills, but also certain item formats might be more suitable to tap into different grade and even proficiency levels. In fact, I might have seen in other contexts where the overall test design (from psychometrics) is totally independent from an item format test design (specplate), since both documents represent different purposes (one is for scoring and psychometric purposes, while the other is for content and criterions needs). This separation results in the specplate being less understood by the psychometricians, as it is typically not given the importance it needs. So having an item format test design allows for the identifying of the strongest item formats and item specifications that will be set in stone and allow organization and consistency in future test administration.
Therefore, in order to illustrate a sample specplate, two instances are provided below. In the first case the documents map out the possible specifications to be developed after having combined those item formats used in the context of the Oregon ELPA (figure 7). The second sample provides a specplate that illustrates the item formats that resulted from the process of elimination, based on the feedback provided by the participants in the present thesis study (figure 8).

As pointed out in the results chapter, an additional step in the completion of this study required that all participants answer the surveys to globally determine what the best item formats would be to test for each part of the construct of the test. Therefore, the results upon the design of the specplate illustrated in figure 8 below is the effect of 118 participants evaluating such formats, vis-à-vis the different constructs of the test. The item formats selected for addressing the different constructs were determined based upon the highest mean distribution of each item format versus each construct.
Figure 7. Specplate (Item Format Blueprint) – Combination of Item Formats

### Writing
- Spec 1 Extended Response
  Visual & Non-Visual
  Prompted Extended Written Discourse
- Spec 2 MCh Visual
  Prompted Answering questions
- Spec 3 MCh Combinations of Sentences
- Spec 4 MCh/SA1-2 Speech
  Parts Construction
  (Grammar) – Sentence Completion
- Spec 5 Short Answer 1-2
  Visual & Non-Visual
  Prompted Sentence Structure
- Spec 6 Short Answer 1-2
  Visual – Single Phrase/
  Word/ Letter Response

### Speaking
- Spec 1 Extended Response
  Visuals & Picture Story
- Spec 2 Extended Response
  Personal Interest topics
- Spec 3 Short Answer 1-2
  Questions & Answers / Different Question Types/
  Cues/ Naming
- Spec 4 Short Answer 1-2
  Responses to Video Recordings – Responses to
  Specific Questions in Situations

### Listening
- Spec 1 MCh Dialogue-
  Passage Literal Comprehension
- Spec 2 MCh Visual-Short
  Conversation Literal
  Comprehension
- Spec 3 MCh Answering
  Questions
- Spec 4 MCh Sentence
  Completion
- Spec 5 MCh Pragmatic
  Inference
- Spec 6 MCh Picture
  Identification (drag & drop)
  Literal Comprehension

### Reading
- Spec 1 MCh Visual & Non-
  Visual Prompted Passage
  Comprehension
- Spec 2 MCh Visual & Non-
  Visual Prompted Sentence
  Completion
- Spec 3 MCh Picture
  Identification (drag & drop)
  Literal Comprehension
- Spec 4 MCh Context Clues –
  Expanded Comprehension
Figure 8. Specplate (Item Format Blue Print) – Selection of Best Item Formats based on Teacher Input

**Writing**

**Grammar**
- ER Essay (all genres)
- MC (picture/diagram stimulus w/sentence completion)
- MC (sentence completion-speech parts construction)
- MC (simple sentence stimulus-correct sentence identification)
- SA1/2 (sentence structure/single sentence)

**Vocabulary**
- ER Essay (all genres)
- MC (picture/diagram stimulus w/sentence completion)
- MC (sentence completion-speech parts construction)
- MC (simple sentence stimulus-correct sentence identification)
- SA1/2 (picture stimulus-sentence structure/single sentence)

**Representational Functions**
- ER Essay (all genres)
- MC (picture/diagram stimulus w/sentence completion)
- MC (sentence completion-speech parts construction)
- MC (simple sentence stimulus-correct sentence identification)

**Morphology**
- ER Essay (all genres)
- MC (picture/diagram stimulus w/sentence completion)
- MC (sentence completion-speech parts construction)
- MC (simple sentence stimulus-correct sentence identification)

**Manipulative Functions**
- ER Essay (all genres)
- ER Essay (w/ picture, diagram stimulus)
- MC (picture/diagram stimulus w/sentence completion)
- MC (simple sentence stimulus-correct sentence identification)
- SA1/2 (sentence structure/single sentence)
- SA1/2 (picture stimulus-sentence structure/single sentence)

**Imaginative Functions**
- ER Essay (all genres)
- SA1/2 (sentence structure/single sentence)
- SA1/2 (picture stimulus-sentence structure/single sentence)

**Heuristic Functions**
- ER Essay (all genres)
- MC (picture/diagram stimulus w/sentence completion)
- MC (sentence completion-speech parts construction)
- MC (simple sentence stimulus-correct sentence identification)
- SA1/2 (sentence structure/single sentence)
- SA1/2 (picture stimulus-sentence structure/single sentence)

**Cohesion & Coherence**
- ER Essay (all genres)
- ER Essay (w/ picture, diagram stimulus)
- MC (picture/diagram stimulus w/sentence completion)
- MC (sentence completion-speech parts construction)
- MC (simple sentence stimulus w/combining sentences)
- MC (simple sentence stimulus-correct sentence identification)
- SA1/2 (sentence structure/single sentence)
- SA1/2 (picture stimulus-sentence structure/single sentence)
Why teachers are not only important but crucial in the process of large scale assessment.

The results and table in chapter 3 portrayed all sorts of descriptive information in regards to the characteristics and geographical information of the participants in this study. In order to support the participation of teachers in the development of large scale assessment, I will start by briefly reflecting upon some of the characteristics of the participants in this study based on the teacher background checklist provided along with the questionnaire.

First, it is important to note that the teacher participants came from different backgrounds and schools districts and universities (approximately 38 different institutions) across all Oregon, and other states, such as Texas. In addition, I want to point out the diversity and distribution of teaching level backgrounds from which they came (elementary, middle, and high school levels). Now, the total means presented in the results chapter not only identifies one teacher having experience with only one particular level (e.g. elementary only or middle school only), but the means reflect the background level in which teachers reported having experience teaching. So in a practical way some teachers reported having teaching experience in two or more levels. Further, one of the most important factors in having chosen such participants came from the total average years of experience in teaching ELL was 8.46 years. Another important piece of data to reflect on is the distribution of the total means where teachers reported having experience teaching ESL/ELL students at the six different levels of English proficiency levels (as indicated in the Oregon ELP
standards), and of course as part of the Oregon ELPA. Thus, the distribution was rather appropriate for the study.

Another strong factor to point out is the diverse teaching experience of the teachers toward students from different backgrounds (Asian, African, Native American Indian, European and Hispanic descent). Thus, just having pointed out some of the characteristics of the teacher participants in the present thesis study, this brings to light some of the wealth of their experiences, background, and knowledge that they bring to the development of a standardized large scale assessment. In other words, these are some of the reasons why teachers’ voices, input, and wisdom are rather critical in any assessment; simply because of the diverse experiences they bring to the table to make the test more valid, reliable, and fair in its construction. The sole decisions posed by a group of experts (call it test developers, psychometricians, curriculum developers, or even policy makers) is not enough to say that a test is valid, or even fair, to all populations that which will be impacted by it.

Hence, let me bring out additional characteristics of the participants who provided their input in the present study. For example, in the teacher background checklist participants were asked which sort of setting (so to speak teaching program) they were currently operating in at the time of the interview or response to the questionnaire. Then ELL programs such as the traditional ESL pull out, late and early exit transitional, bilingual-dual language, sheltered English, mainstream classroom, literacy center, and even those performed in typical community colleges (e.g. community college, intensive English language programs, immigrant, international
students, etc.). Therefore, several teachers reported having experience in the aforementioned settings, which in turn demonstrates the diversity and understanding of programs and methods that impact the way this particular population is assessed. This contains implications in the methodologies that teachers use to address students needs, such as the tasks used in the classroom and how these relate to the tasks presumably used in a large scale assessment.

In addition, I would like to bring out the sort of training methodology teachers were asked to report when they answered the questionnaire. They reported a variety of ELL/ESL methods widely used throughout Oregon school districts and community colleges. For example, they reported taking linguistics courses tapping into the theoretical needs of ESL/ELL populations, assessment college courses and training provided in their school districts, TESOL methods courses, amongst other commercialized sort of training (systematic ELD forms and functions of language, GLAD training, CALLA training, and SIOP training). Consequently, this information from participants brings an immense amount of knowledge to the assessment construction since teachers can call upon their knowledge of teaching methodologies to evaluate if test items are appropriate for students in their areas of teaching. They can also propose new ways to assess students by stating what does work and what does not work for a given sub-population of ELL students, based on the teaching methods and training that they possess.

A final factor (based on the teacher background checklist use in the present study) is the test experience teachers had at the time of addressing the questionnaire.
For example, they reported having experience and training in the following off-the-shelf assessments used in the Oregon ELL teaching contexts: Woodcock-Muñoz Language Survey, SOLOM (Student Oral Language Observation Matrix), Stanford SELP (English Language Proficiency) by Harcourt Assessment, SLEP (Secondary Language English Proficiency) test, IDEA (IPT), LAS (Language Assessment Scales), LPTS (Language Proficiency Test Scales), DRA (Developmental Reading Assessment), IRI (Individual Reading Inventory), ELPT (English Language Proficiency Test) by the College Board, TOEFL (Test of English As a Foreign Language), and the official state assessment instruments, such as the State Reading, Speaking and Writing Rubrics. Initially all of this information was sought (the commercial ELL assessments used in Oregon) in order to demonstrate validity for the present thesis study, in terms of having recruited the appropriate body of educators to evaluate the item formats in question. However, I now bring out this information to not only report on the validity of the study, but also to highlight the relative importance of teachers’ background and training and demonstrate their experience with other assessment instruments as a basis for validating other assessment contexts (such as the one posed in the present study).

Therefore, all of the factors previously mentioned contain strong arguments for why it is not only important, but critical, to have teachers in the field propose courses of action and recommendations in assessment development. Allowing them to make decisions and be part of the assessment development improves the process and enables an environment in which educated decisions can be used in the
development of assessments, so that the resulting instrument is not only fair and
reliable, but also Universally Designed and Culturally Fair for the diverse populations
who make up English Language Learners.

After having posed some significant arguments based solely on the
characteristics of the participants in this study, I will now attempt to make such
arguments stronger and more relevant, and perhaps even illuminate the entire purpose
of this thesis, by bringing to light some experiential cases in which the teacher input
has been used as the breaking point in the final decisions in the development of an
English language proficiency assessment (and other sorts of standardized
assessments).

*Teachers as decision makers (the breaking point) in other contexts.*

To emphasize contexts beyond the scope of this study, I will point out my
experiential interactions with teachers, who were the first group of in the body of test
developers who saw the Oregon ELPA being created from scratch. In addition,
teachers have played an integral role in other contexts (projects) that I have worked
with while becoming a test developer at a testing company.

*Identifying assessable standards (special cases).*

Assessments are generally created after the standards have been set. In the
case of the Oregon ELP test, due to the time limitations imposed by the federal
government, the ELP standards were still under development when the initial test
construct was conceived. This construct was based on a framework derived from
three language models — Scarcella’s conceptual framework of academic English
(2003), Halliday’s language functions (1973, 1976), and The Canadian Language Benchmarks (2002). This framework defined English language proficiency as comprised of three main competences—grammatical, functional, and textual—underlying the ability to communicate in English in an academic setting. This test construct also incorporated Dutro and Moran’s (2002) taxonomy of 23 language functions that the department of education was using for the development of the state’s ELP standards (as previously noted in the Literature Review Chapter of this thesis).

Consequently, when the first draft of the standards was completed, it contained over 1,100 items covering the four skills: listening, speaking, reading, and writing across all grade levels, kindergarten through grade 12. Given the large number of standards, the next step was to determine which standards were suitable and critical for testing. First, a panel of teachers (content panel) was selected in order to identify those standards that could be assessed in a testing situation. This process was important, since teachers from different grade levels working with L2 learners were providing their personal and experiential insights, in terms of knowing what standards fulfilled the requirements. The panel analyzed all 1,141 standards and selected those deemed “assessable” based on the three competences of the test construct (grammatical, functional, and textual).

The panel began by selecting assessable standards. They ranked each standard from 1 to 3 in degree of assessability with 3 as the highest and B (broad) as having too many variables to assess and/or reflecting too many language functions. In
addition, they coded each standard by competence (grammatical, functional, and textual) and classified them according to their degree of linkage in the ELP construct. Standards with a score of 1 and a code of B were discarded. Finally, they organized standards by domain (reading, writing, listening, and speaking) and by individual grade level. It is important to note that for the sake of consistency with the ELP standards, the panel coded the standards as falling under functional competence if they fulfilled any of 23 language functions, as devised by Dutro & Moran (2002).

Thus, the absolute importance of teachers’ insight to make these sorts of decisions was crucial since they were the only body of individuals who truly knew their students and truly knew the curriculum. In addition, these teachers understood the standards and had the knowledge to really decide which standards should belong to assessment in the classroom and which standards should belong to a typical standardized assessment. Without their initial input, the development of an assessment would have been placed at risk, and the test developers would have not achieved what needed to happen in the development of the assessment. I would like to point out another context for these processes and concepts, in that it has been my experience that even on other states that did not have the same time pressures, there has not been a formal procedure to select assessable standards for testing. Perhaps the initial process in the development of a standardized assessment such as this one is something that needs to be considered for such programs.

Identifying appropriate item formats (special cases).

The specific case that follows represents the entire purpose of this thesis. In
the context of this assessment (Oregon ELPA), and during the first pilot testing for its
development, the main objective was to detect what item formats would be
problematic for test takers. A taxonomy of item formats available from the language
testing literature was developed by Perea (2004) and used to categorize the initial
item bank developed by item writers. Therefore, about 6000 students took the pilot
test and the results indicated that all of the item formats performed very well across
all students, so there was no way to eliminate bad item formats, however some
individual test items were removed.

Therefore, the evaluative judgments from teachers on item formats allowed
the establishment of a better framework and suitability onto what items/formats could
be reverse-engineered. Of course, these results can be indicative that item developers
have a better picture of item formats and illustrative items underlying the principles
behind the crafting of test specifications and item writing. However, the teacher
evaluation pursued in this study allowed for the determination of what item formats
were better for ESL students, as well as their suitability to the construct/standards.

Having mentioned this, the results from this study can shed light for future
work, and in the process, help to create detailed item test specifications. In the ideal
world of testing, teachers should know what students will encounter during a test,
including the expected responses and all the elements that are needed to create these
test items. The teachers’ input can fill in the gap in capturing a learning experience
from this process into other contexts of the development of large scale assessment,
these being the places where statistics cannot provide the “break point” in the
decisions that marks the course of this development. This is a clear example where capturing teacher’s judgments and opinions, which are based on their experience and wisdom, can really make the difference. Thus, the execution of a quantitative and qualitative study such as the one presented here can serve as a model for other contexts in similar situations.

**Selecting test items in data review (typical cases).**

This is a typical case where content decisions (and even better those coming from educators) can offer a supplement of better judgments in the creation of a large scale assessment than those that are only derived from the statistics.

Numerous times while attending data review committees, I have encountered the never-ending dilemmas that educators discuss when they have to face a decision of whether or not to accept or kill a test item based on either statistical data or content. In the end all psychometricians will agree that what matters is that the statistics are not the sole and absolute decision point . . . no matter how bad they might seem, the breaking decision point is derived from the content surrounding the test item. Such content decision should come from the teachers who teach the content and interrelate with the test takers, as they know the test takers, the curriculum, and the standards. The teachers’ decision weighs in more than statistical results coming from a field test. In the end, my experience dealing with theses dilemmas has shown that these decisions are balanced out toward the quality of the item, the content, and how it is constructed, rather than the numbers in a field test form. An analogous way to think about this process occurs when constructing a final operational form.
Statistics provide a relative piece of information to take into consideration, but
ultimately we must rely on the quality of the content, the balance of standards used,
the balance of ethnicities represented in the test items with art, the types of reading
and listening passages (informational, narrative, etc.), how the items match the
blueprint and test design, etc.; because the content is what matters in the end. The best
source of content in data review committees is the teachers who collectively make
decisions based on their knowledge, experience, and understanding of students.

_Teachers in other processes – item-task-fit-to-spec & test specification_
写作 / 审查（特殊情况）。

Based on the experience I went through in the process of completing this
thesis, as well as my teaching and test development experience, I contend that the
teachers’ voices and input are incredibly critical in future processes of test
development, both on a small and large scale.

This test construction is shown on a small scale, perhaps at a school district
level when there is a need to create a placement or summative ELL assessment for
elementary, middle or high school; or even when ESL instructors get together to
create a summative assessment for a given level of ESL to avoid relying on a
commonly used off-the-shelf test, which in turn is not customized to the needs and
curriculum of that particular ELL population.

This test construction is shown on a large scale when constructing a
standardized assessment such as this, which will impact the needs of ELL students in
an entire state. However, whether the State Department of Education decides to
pursue this endeavor alone or with the help of a Testing Company, again teachers need to be part of the process.

So as test developers, curriculum developers, item writers, and teachers would agree that it is critical to have item test specifications in place, I highlight the magnitude of the contribution of the panels or committee teachers (depending on each particular context and development situation) involved in test development. These teachers can collectively create test item specifications, based on their understanding of the test, the existing test items, and the standards and the curriculum that are used. In other cases, if the State Department/Testing Company have crafted some item writing guidelines, then such documentation needs to be evaluated by teachers to see if the recently developed items in fact are a true and valid representation (byproduct) of such specs. In addition, I content that there is a gap in the test development processes where no committees or panels exist in determining what Davidson and Lynch (2002) called “item-task-fit-to-spec,” in which there is actually a practice of verifying if indeed the test items match the specs and constructs. Typically, the contribution of content panels includes a quick verification of test item versus its allocated standard; but this does not include a formal review of a document (recipe) to create item test specifications against the produced item. I argue for allowing for the creation of a panel of teachers who can indeed verify the veracity of these documents and the test items.

In conclusion, I would like to end this section by posing a simple analogy that shows why we need test specifications and also an activity for verifying them.
Numerous times while working on various projects, I have seen testing companies engage in the questionable practice of having teachers write test items without any test specifications already in place. I ask how we can expect a group of teachers to put together a bicycle or bake a cake when there are not directions or a recipe available for doing so? This is the reason and argument for allowing panels of teachers work on the tasks of producing and reviewing test specs.

**Inherent participation of teachers in assessment panels.**

Educators from many teaching levels, contexts and backgrounds have historically collaborated in many ways to make educated decisions and recommendations for the crafting of assessments at large scale. They usually are called on to participate in a panel which specializes in different aspects of the assessment. In order to highlight the absolute importance of these participants in test development, I describe their purpose in various review settings.

**Item Writing Panel/Committee** – Teachers in these panels are usually in charge of writing the test items (questions) of the assessment. They take as reference the standards, the components of the construct, proficiency levels, and select appropriate materials (pictures, reading selections, diagrams, etc.) to produce them. Some item writing panels are performed in-house, in the state departments of education and are typically trained by the test and curriculum developers in charge of the project. Some other ways of item writing are performed online, where teachers receive instructions on how to create such items based on item writing guidelines (item test
Content or Item Review Panel/Committee – Educators in this panel have the duty to verify the “content validity” of the test. In other words, they make sure that test items and rubrics are aligned to both the construct of the test and the ELP (or any content) standards. Occasionally, their duties include reviewing the ELP/other subject standards and the selection of “assessable standards” for item writing purposes.

They also provide recommendations for editing test items if there is not a clear alignment between standards/construct vis-à-vis test items and rubrics. They review test items for content alignment, grade level difficulty, and level of second-language appropriateness. From the test developer perspective, this is typically the final stage after new field test items are developed, since items come from item writers and go through several content and editorial stages within a testing company. These items are brought to a committee or panel of teachers (representative of the state where the assessment resides) so that the newly developed field test items can be signed off by such committees.

Sensitivity & Bias Panel/Committee – After test items or rubrics have gone through the content panels, then they move on to the sensitivity panel. This panel is composed of educators from different ethnic backgrounds, parallel to the student population from across the state (or states where the assessment resides). The panel’s duty is to revise the finalized test items (including any pictures that go along with them) prior to
using them for a test trial. The objective is to make sure items are not offensive or inappropriate for the different ethnic and religious populations of test takers within the state. For example, they look at issues of ethnicity, religion, gender, inappropriate or harmful language, etc.

**Rubric/Range Finding Panel/Committee** – The rubric panel is in charge of creating the rubrics, or scoring guides, for the assessment. In this case, the rubrics are those for test items formatted as Short Answer 1-2 points and Extended Response 3-4 points under modalities such as speaking and writing in ELL assessments. Educators within this panel are in charge of listening to hundreds of student oral responses as well as reading written responses in order to create a standards-based rubric. At this stage of development, teachers are couched by a specialist who makes sure the rubrics are valid (aligned to the item, standard, and construct) and reliable (consistency of scoring).

**Panel/Committee of Raters** – Raters are a group of educators who are trained on how to score the oral and written student responses, based on the newly-created rubrics produced by the rubric panel. Due to the nature of the test items in ELL assessments, one rater can score hundreds of responses for a single item. Scoring can be done online and teachers can complete this working at home. Other models used by State Departments of Education and Testing Companies are simply training these educators in-house and then having them score students responses for an allocated amount of
time based on a contract.

**Special Education Panel/Committee** – The participants in this committee are a more specialized group of individuals with extensive background, knowledge, and understanding of special education issues, especially those affecting ELLs. Their duty is to look at a battery of test items and provide recommendations in the way the assessment is delivered to Special Ed ELL students. They also assist in the creation of a policy to outline these recommendations and changes that the assessment team needs to follow up on in the creation of the test.

**Standard Setting Panel/Committee** – Another set of teacher participants is the standard setting panel. Their duty is to establish a systematic selection of test items that belong to specific standards, or in the case of ELL assessments, the notion of language proficiency levels. Items are hierarchically rank-ordered by level of difficulty (least to most), based on the statistical results of pilot and field tests. Then, educators look at each proficiency level and empirically decide what items may constitute a “cut score” between one proficiency level and the next one. Usually they are couched by a statistician or a psychometrician who employs a model to standard setting (e.g. bookmark – as previously discussed, among other methods employed).

**Data Review Panel/Committee** – The individuals on this committee are a group of teachers who are called by the State Department of Education and the Testing
Company to review the resulting data of a battery of items that just underwent field testing. Since the items in question already contain statistics, the educators’ task is to select those items suitable for placement in an item bank that could be used in an operational test form. They look at a diverse set of statistics, such as the item mean (a.k.a.: item p-value, item difficulty), item total correlation (a.k.a.: discrimination, point bi-serial), option/score level total correlations, option/score level percentages, and differential item functioning (DIF) in order to make use of both statistics and content-educated decisions to use or reject a given test item.

**Final Forms Review Panel/Committee** – This specialized group of individuals is initially comprised by a group of test developers and state department representatives. Typically, the test developers put together a proposed set of test forms after a field test. Then such forms are sent to the state department content and statistics experts in charge of the final sign off. However, it is the last panel that provides their final eyes and points out any recommendations or discrepancies that were not caught by both the test developers and the state department representatives. This group of individuals is a set of well-informed and versed teachers in the field who are rather familiar with the test after its implementation.

**Instructional Impact Committee** – This is a similar group of teachers across all the fields (ELL, Special Education, etc.) within a given state who are called in order to take a look at a battery of sample items with a sole purpose of determining if the
assessment actually fits the intended instructional outcomes. This group often works under political considerations.

Conclusion

After having gone through this experiential process with the teacher participants in this study, and also with the teachers whom I interacted in the development of the Oregon ELPA, several valuable outcomes can be highlighted.

For example, teacher participants in this study performed a very principled process in their analysis of item formats. Consequently, teachers’ input has achieved an informed judgment where they were able to discriminate amongst the different item formats and sample items appropriately.

Another aspect to highlight is the strength of the sample used in order to bring validity to this study, but also to mention the notion of variability in the teacher responses. Because of this characteristic, it was proven that they were able to engage with and provide meaningful data in the process. As a result of this, educators in general should be involved in test construction due to their refined sense of the importance of these items as they can see item features that test developers cannot.

Final Remarks

In summary, having highlighted all the possible manners in which educators’ voices are used during different facets of test development in various states, it makes the argument stronger for the fact that test developers and state representatives cannot
undergo test development of this magnitude without the wise knowledge and experience of the primary stakeholders... the teachers who interact day-to-day with the test takers, who are the real customers of the testing companies’ contracts. Here I shall even argue that the entire purpose of test development in large testing corporations should not be driven to please the so-called customer “state departments” but the real customers who are being affected by the scores and the decisions that are made with such scores, including whether or not to pass or not pass a subject, to enter or deny entrance into the next level or a new school, and to provide or deny federal funds for a school based on group student performance. This is the reason why I contend that the real customers and the entire process of high quality should be motivated by how test takers and teachers are impacted... the real customers.

In closing, as we can observe the development of a large-scale assessment for ELLs is quite a complicated and time-consuming process that requires the knowledge of both the specialists who are experts in the theory and test construction and the educators who truly understand the needs of English language learners. As the Oregon ELPA (and all other state assessments) seeks to satisfy federal requirements, teachers should be free to make educated and research-based decisions on how to report ongoing language development. In addition, these teachers should be able to use this information to contribute to the larger context of language test development and to design their own customized assessments for their unique student communities.

Teachers’ voices need to be heard in the development of measurement
instruments of any magnitude so that validity can be fully achieved. I would like to end this thesis project with 4 principles that I believe constitute the heart of language test development at all levels (small and large scale).

These principles are derived from assessment experts such as Popham (1994) and Lynch and Davidson (2002). Thus, promoting test development in an Iterative, Flexible, Consensus-based, and Specification-driven fashion is critical. “Iterative” means that the test development process is responsive to feedback from peers’ comments (experts and educators), from trial and from operational useages. The second principle, “Flexible,” implies that the tests as well as test specifications are not set in stone, but they may change and adapt to shifting contexts. The third principle “Consensus-based,” refers to the unique benefit of test development and test specifications as a cooperative aspect of all interested parties in this process. Hence, decisions should not originate only from the top (e.g. specialists and theoreticians), but also come from the bottom (educators whose understanding of students is extremely critical). The last principle, “Specification-driven,” relies on the test development being based on test specifications. “A specification is an efficient generative recipe for a test that fosters dialogue and discovery at a higher, more abstract level than achieved by analysts of a simple item or task” (Davidson & Lynch, 2002, p. 7).
References.


Davidson, F. (2003). *Crafting a test of academic language proficiency: Some comments on reverse engineering and certain philosophical tensions in large-scale testing*. University of Illinois at Urbana-Champaign.


Hargrett, G. (2004). *Consolidated standards document*. Oregon Department of Education Item Writing Workshop. Training presented to develop English Language Proficiency (ELP) test items for listening, speaking, reading and writing domains to be used on Oregon’s English Language Proficiency Assessment, Salem, Oregon.


Oregon English Language Proficiency Standards (2004). Reading Grade 2. Listen to and Read Informational and Narrative Text.


(2004). Oregon English Language Proficiency Standards.


Perea, L. (2004, June). *North West Regional Educational Laboratory Report on Test Items*. Research report prepared for the NWREL in the creation of a taxonomy of language testing item formats for the K-12 Oregon English Language Proficiency Test, Portland, OR.
Perea, L. & Zaheed, F. (2004). Selecting Standards for Item Writing. Oregon Department of Education Item Writing Workshop. Training presented to develop English Language Proficiency (ELP) test items for listening, speaking, reading and writing domains to be used on Oregon’s English Language Proficiency Assessment, Salem, Oregon.


Scott, G. (2004). Defining and Operationalizing the English Language Proficiency Construct. Paper presented at the Oregon Department of Education Item Writing Workshop. Training presented to develop English Language Proficiency (ELP) test items for listening, speaking, reading and writing domains to be used on Oregon’s English Language Proficiency Assessment, Salem, Oregon.


Zaheed, F. (2004). Analysis of the Language Functions Presented by Dutro and
Moran Rethinking English Language Instruction: An Architectural Approach (2002). Research report prepared for the Northwest Regional Educational Laboratory, Portland, OR.
Appendix A.
Teacher Background Checklist
School District ______________
School Level currently teaching (circle one):   Elem. S, Middle S, High S
Please answer the following questions according to your background information. This is valuable information that will help the researcher code the data as means of data collection validation. All of this information will be kept confidential and will not be disseminated to anyone. In order to protect your confidentiality, this data will be coded with numbers and letters as your name and school where you work will be kept confidential.

(a) How many years of ESL teaching experience you have so far? __________

(b) What grade levels you have most experience teaching ELLs? circle one (K-1) (2-3) (4-5) (6-8) (9-12)

(c) What English proficiency levels have you taught? Circle those that apply Beginning - Early intermediate - Intermediate - Early advanced - Advanced - Proficient

(d) What ESL students with diverse ethnic background have you taught? (Hispanic) (European) (Native American Indian) (African) (Asian) Other _________

(e) Circle the settings that you have had more experience with (ESL pull-out) (Late exit transitional) (Early exit transitional) (Bilingual-dual language program) (Sheltered English) (Mainstream classroom) (Literacy center) (Community College – IELP, ESL, Immigrant, International students) Other _________

(f) Circle or write the relevant types of training that you have had so far in order to address English language learners (SIOP training) (CALLA training) (GLAD training) (Systematic ELD forms/functions) (TESOL Methods) (Assessment courses/training) which ones? ________ (Linguistics courses) which ones? ______
(g) Have you had experience proctoring or received training in any of the commercial tests to measure ELLs' language ability? If so which ones? (mark those that apply)
- ☐ (Woodcock-Muñoz Language Survey)
- ☐ (SOLOM - Student Oral Language Observation Matrix)
- ☐ (Stanford)
- ☐ (SLEP - Secondary Level English Proficiency)
- ☐ (IDEA (IPT) English Proficiency Test)
- ☐ LAS (Language Assessment Scales)
- ☐ LPTS (Language Proficiency Test Series)
- ☐ DRA (Developmental Reading Assessment)
- ☐ IRI (Individual Reading Inventory)
- ☐ English Language Proficiency Test for domestic students (by the College Board)
- ☐ TOEFL (Test of English as a Foreign Language)
- ☐ State Writing Rubric for work samples
- ☐ State Speaking Rubric for work samples
- ☐ State Reading Rubric for work samples
- ☐ Other ____________________________________________
Appendix B.
Levels of Proficiency

Taken from Scott (2004). Levels of Proficiency.

----Reference Material----

Entering =
Little to no communication
Emerging =
Limited communication
Developing =
Basic communication
Maturing =
Competent communication
Nearing =
Complex communication
Native-like =
Sophisticated communication
### Appendix C. Some Examples of Language Functions

<table>
<thead>
<tr>
<th>Language functions to convey information</th>
<th>Language functions to get things done</th>
<th>Language functions to learn about language</th>
<th>Language functions used to enjoy language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conveying information</td>
<td>Instrumental</td>
<td>Finding the meaning of words</td>
<td>Opening a story</td>
</tr>
<tr>
<td>Expressing needs and likes</td>
<td>(Session - Getting things done)</td>
<td>Finding out how to say something</td>
<td>Closing a story</td>
</tr>
<tr>
<td>Expressing emotional attitudes</td>
<td>Giving instructions on tasks</td>
<td>Using a dictionary to look up a word</td>
<td>Recognizing poetry or verse</td>
</tr>
<tr>
<td>Expressing moral attitudes</td>
<td></td>
<td>Using a thesaurus</td>
<td>Reciting and enjoying poetry</td>
</tr>
<tr>
<td>Describing people, places and things</td>
<td></td>
<td>Planning, organizing, composing</td>
<td></td>
</tr>
<tr>
<td>Describing spatial and temporal relations</td>
<td></td>
<td>and revising an essay</td>
<td></td>
</tr>
<tr>
<td>Describing actions</td>
<td></td>
<td>Demonstrating the meaning of an</td>
<td></td>
</tr>
<tr>
<td>Retelling / relating past events</td>
<td></td>
<td>utterance/sentence</td>
<td></td>
</tr>
<tr>
<td><strong>Finding information</strong></td>
<td></td>
<td><strong>Reference Material</strong></td>
<td></td>
</tr>
<tr>
<td>Asking informational questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asking clarifying questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Using information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expressing and supporting opinions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explaining</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persuading</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arranging information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrasting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Creating information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literary analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cause and effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drawing conclusions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defining</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Making predictions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothesizing and speculating</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Presenting information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalizing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpreting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summarizing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questionnaire for teachers (constructed response)
Instructions:
1) Look at the entire item formats with their definition.
2) Look at some sample items from which this item format was used to organize commonalities and skills reflected on each test item.
3) Make sure that the following questions are applied to each item format you are evaluating.
4) Circle the grade group that you are most experienced with.
5) Think about what are the characteristics of your ELL students in terms of age, grade level, and English level.
6) Reflect upon the fact that if you were going to write items within these item formats what characteristics can you change to enhance these formats (based on the grade group you are most experienced with). Think in terms of appropriateness of item format use for ELL students in the grade group you are most experienced with.
7) Reflect on the general idea of the purposes of this test and the sample items presented. Think of this general question: do these item formats fit the items presented?
8) Write your comments and be as specific and explicit as possible.

Grade groups (circle one only, the most experience you have with):
K-1 2-3 4-5 6-8 9-12

From the choices below, mark those attributes that apply to be tested in this item format. If no presence of any, explain why.—Give suggestions)

Item Format Name ______________________________________
Domain (e.g. speaking or reading, etc.) ______________________________

Grammar YES ☐ NO ☐ WHY?
Vocabulary YES ☐ NO ☐ WHY?
Morphology YES ☐ NO ☐ WHY?
Language functions to convey information YES ☐ NO ☐ WHY?
Language functions to get things done YES ☐ NO ☐ WHY?
Language functions to learn about language YES ☐ NO ☐ WHY?
Language functions used to enjoy language YES ☐ NO ☐ WHY?
Cohesion & Coherence in writing (text unity and relationship between different meanings) YES ☐ NO ☐ WHY?

A) Do you believe the task/item format is too difficult or ambiguous for this particular level of student? (If yes, why – Give suggestions)
B) Do you believe the items fit the item format? (If yes, how? – If not, why? What is missing?) Is the item format comprehensible, can the description be used to generate similar test items? (If not, why? – Give suggestions)
C) Given your teaching experience, do you believe the items and item format are appropriate for ESL students in this particular grade group? (If not, why? – Give suggestions)
D) Do you believe items and item format fit the selected proficiency level? (If not, why? – Give suggestions)
E) Do you believe the tasks represented in this item format match your typical instructional tasks on a regular basis? (If yes, how so? – If not, why? – Give suggestions)

F) Do you believe items/tasks from this item format need to be explicitly pre-taught to a given group of test-takers? (If yes, why?, which ones?)

G) Do you believe the procedure for test-taking needs to be explicitly pre-taught to a given group of test-takers? (If yes, how so?)

-------------------------------- Answer this question only once as applicable to all item formats. For instance, if there are 10 different item formats, select those that are applicable to test each one of the skills mentioned below-----------------

H) Given your teaching experience, what item formats do you believe match better the purposes of the test? (write the assigned numbers from each item format only)

Grammar __________________________
Vocabulary __________________________
Morphology __________________________
Language functions to convey information __________________________
Language functions to get things done __________________________
Language functions to learn about language __________________________
Language functions used to enjoy language __________________________
Cohesion & Coherence in writing (text unity and relationship between different meanings) __________________________

**Questionnaire for teachers (Likert-scale format)**

Instructions:
1) Look at the entire item format.
2) Look at some sample items from which this item format was used to organize commonalities and skills reflected on each test item.
3) Circle the grade group that you are most experienced with.
4) Think about what are the characteristics of your ELL students in terms of age, grade level, and English level.
5) Reflect upon the fact that if you were going to write items within these item formats what characteristics can you change to enhance these formats (based on the grade group you are most experienced with). Think in terms of appropriateness of item format use for ELL students in the grade group you are most experienced with.
6) Reflect on the general idea of the purposes of this test and the sample items presented. Think of this general question: do these item formats fit the items presented?

Choose only one option and state if you:

**Strongly disagree - Disagree – Agree – Strongly agree.**

(1) (2) (3) (4)

Grade groups (circle one only):
K-1 2-3 4-5 6-8 9-12

(a) The items contained in this item format assess the purpose of the test
Strongly disagree - Disagree – Agree – Strongly agree

(b) The task/item format is too difficult or ambiguous for this particular level of student
Strongly disagree - Disagree – Agree – Strongly agree

(c) The test items fit the item format
Strongly disagree - Disagree – Agree – Strongly agree

(d) The item format is comprehensible
Strongly disagree - Disagree – Agree – Strongly agree

(e) The items and item format are appropriate for ESL students in this grade group
Strongly disagree - Disagree – Agree – Strongly agree

(f) Items and item format fit the selected proficiency level
Strongly disagree - Disagree – Agree – Strongly agree

(g) Items and item format are applicable and practical to use for this grade level
Strongly disagree - Disagree – Agree – Strongly agree

(h) The tasks represented in this item format match your typical instructional tasks on a daily basis
Strongly disagree - Disagree – Agree – Strongly agree

(i) Items/tasks from the item format need to be explicitly pre-taught to a given group of test-takers
Strongly disagree - Disagree – Agree – Strongly agree
(j) The procedure for test-taking needs to be explicitly pre-taught to a given group of test-takers
Strongly disagree - Disagree – Agree – Strongly agree
(k) Items and item format are appropriate in age and cognitive level of students at this grade band
Strongly disagree - Disagree – Agree – Strongly agree

(l) Items contained in this item format test the purposes of the test in:
- Grammar - Strongly disagree - Disagree – Agree – Strongly agree
- Vocabulary - Strongly disagree - Disagree – Agree – Strongly agree
- Morphology - Strongly disagree - Disagree – Agree – Strongly agree
- Language functions to convey information - Strongly disagree - Disagree – Agree – Strongly agree
- Language functions to get things done - Strongly disagree - Disagree – Agree – Strongly agree
- Language functions to learn about language - Strongly disagree - Disagree – Agree – Strongly agree
- Language functions used to enjoy language - Strongly disagree - Disagree – Agree – Strongly agree
- Cohesion & Coherence in writing (text unity and relationship between different meanings) - Strongly disagree - Disagree – Agree – Strongly agree

(m) What grade groups are most appropriate for this item format? Circle only those that apply
K-1     2-3     4-5     6-8     9-12

(n) What proficiency levels are most appropriate for this item format? You can consider as reference the grade groups you selected above. Circle only those that apply
Entering (little to no communication)
Emerging (limited communication)
Developing (basic communication)
Maturing (competent communication)
Nearing (complex communication)
Native-Like (sophisticated communication)
### ITEM FORMAT - Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive, Compare/Contrast) Extended Response

Students are given a prompt asking for an anecdote, imaginative information, a persuasion, or a description. No quantity limit of writing is specified. No stimulus (picture) is necessary since it is a prompt standing by itself. However, in some cases it can contain quotes. The types of discourse in which the genre fall can be imaginative, informational, persuasive, descriptive, comparison & contrast. After having written the essay, students are graded on a 4-point scale that depicts the expected response attributes for that essay. Each attribute is represented in a holistic or analytical scale.

- **Maturing (2-3)**
  What is the first thing you do when you get to school? What comes next? Tell about a day at school.

- **Developing (4-5)**
  Write about an animal that you would like to have for a pet. Describe the animal and tell how you would take care of it.

- **Native-Like (6-8)**
  Compare two different jobs you might have in the future. Write about the positive and negative points of each job.

- **Nearing (9-12)**
  Some people believe that it takes hard work to be successful in life. Others believe it takes luck. What do you think? Write an argument explaining which you believe to be true and why.
ITEM FORMAT - Essay (Extended Response w/Picture, Diagram Stimulus)

Students first observe a picture, diagram, a video, or a description of a process (ex. demonstrating an experiment). They are given a prompt based on their analysis of the ‘stimulus’ in order to describe or imagine something about the picture, diagram or video. No quantity of writing is specified.

No stimulus per se is identified since it is a prompt standing by itself. However, we can consider as stimulus a picture or series of them. This picture is ‘key complement for understanding the stem or passage, that is the visual(s) used with the prompts are critical for the test-taker to aid in their understanding of what is expected from them. Pictures can be from people, people executing various activities (ex. Outdoor), places (ex. The playground), objects, animals, plants, processes of nature, seasons of the year, etc.

➢ Developing (9-12)

| Picture of a student in elementary school taking a test. |
| Another student sitting on her right side is looking at her answers. |

Imagine that you saw this happen in your classroom, what might you do about it? You have 30 minutes to write about your choices and explain why you took this decision.

➢ Maturing (4-5)

| Picture of a boy who just experienced an accident. |
| His knee is injured. Two other children run to help him. |

Here is a picture of an emergency. Write a short story that describes what happened and what will happen next.

➢ Nearing (2-3)

| Picture of three children wearing helmets and bicycle riding on the neighborhood. |

What these students might be doing? Where are they going? What will they do when they get there?

➢ Native-Like (6-8)

| Picture of two merchants selling vegetables. A female customer is choosing and buying some of these vegetables |

Explain what is happening in this picture
Appendix H (Item Format Description 3 Writing)

WRITING

ITEM FORMAT - Multiple Choice (Picture/Diagram Stimulus w/Sentence & Word Completion)

Students are first presented a picture or diagram as a stimulus. This picture(s) is ‘key complement for understanding the stem, that is the visual(s) used with the stems are critical for the test-taker to aid in their understanding of what is expected from them. Then, they complete a sentence with a missing blank that could be placed in the beginning, middle or end of the sentence. Four assigned letter choices are given to students. The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters. Students need to fill in the blank with the appropriate grammatical speech part given the nature/context of the sentence that can range from simple vocabulary (nouns/pronouns, prepositions, conjunctions) identification to the appropriate inflection of verbs (time frames) and adjectives (comparatives/superlatives), as well as modifiers (adverbs). In addition, given the properties of this item format, test-takers can be assessed in their ability to identify and/or place letters in the appropriate gap. Therefore, this skill is concerned with the notion of Mechanics.

➤ Entering (4-5)

Picture of a table. A bird is flying while a cat is sitting on the table. Under the table, a dog is barking to the cat.

Choose the correct answer

The dog is ____ the table
a. under
b. above
c. next to
d. on

➤ Maturing (2-3)

Picture of a student handing his completed homework to the teachers

Fill in the blank
Andy _____ his homework
a. done
b. finish
c. finished
d. do

➤ Native-Like (K-1)

Picture of a duck

What letter goes in each word?
D__ck
a. a
b. e
c. u

➤ Maturing (6-8)

Picture of a mother and daughter working in a garden

This mother and her daughter are working _________

a. in the garden
b. under the garden
c. over the garden
d. below the garden
Appendix I (Item Format Description 4 Writing)

ITEM FORMAT - Multiple Choice (Picture/Diagram Stimulus w/Answer Questions)

Students are first presented a picture or diagram as a stimulus. This visual or illustration is a key complement for understanding the stem, thus the visual(s) used is (are) critical to the test-takers to aid in their understanding of what is expected from them. Then, after having analyzed the illustration along with the presented prompt (stem) they choose a sentence that describes what the picture(s) is (are) portraying. Four assigned letter choices are given to students. The emphasis here on test-takers is to be able to describe with complete sentences (these can range from simple all the way to compound-complex sentences and even subordinate clauses) what the visuals/stems are portraying. Given this, aspects such as using appropriate sentence structure, verb time frames, and adjective inflections are critical to explain, describe, compare/contrast, answer questions and convey information in writing.

- Developing (2-3)
  Picture of a boy holding a mirror looking at himself.
  He uses the mirror to help paint himself on a canvas.
  On one side he has a container with paintbrushes and pencils.

  Look at the picture and choose the correct answer

  Which sentence best describes this picture?
  a. The children are running
  b. The children are singing
  c. The boy is reading
  d. The boy is drawing

- Maturing (4-5)
  Picture of a big old balance (weighing scale).
  One side of the scale is holding 3 men and 2 women.
  The other side is holding a giant pumpkin.
  The weight gives preference to the pumpkin.

  Which of these is true?
  a. The pumpkin is heavier that the people
  b. The pumpkin is less heavy than the people
  c. The people are as heavy as the pumpkin
  d. The people weight more than the pumpkin

- Maturing (6-8)
  Picture of woman setting up a table for dinner
  Picture of woman cooking in the kitchen
  Picture of woman dishing up food
  Picture of woman washing dishes in the kitchen

  Which of these sentences best describe the action in the pictures?
  a. Rebecca set the table and then cooked dinner. After she ate dinner, she washed the dishes.
  b. Before Rebecca set the table, she cooked dinner. After she ate dinner, she washed the dishes.
  c. Rebecca cooked dinner before she set the table. She ate dinner, and then she washed the dishes.
  d. Rebecca set the table and then cooked dinner. Before she ate dinner, she washed the dishes.
Appendix J (Item Format Description 5 Writing)

WRITING

ITEM FORMAT - Multiple Choice (Sentence Completion-Speech Parts Construction)

Students are first presented with a sentence as stem. Then, they complete that stem with a missing blank that could be placed in the beginning, middle or end of the sentence. Four assigned letter choices are given to students. The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters. Students need to fill in the blank with the appropriate grammatical speech part given the nature/context of the sentence that can range from simple vocabulary (nouns/pronouns, prepositions, conjunctions) identification to the appropriate inflection of verbs (time frames) and adjectives (comparatives/superlatives), as well as modifiers (adverbs).

- Entering (4-5)
  Do you want apples ___ bananas?
  a. to
  b. or
  c. for
  d. next

- Maturing (2-3)
  I went to three ______ this month
  a. parties
  b. part
  c. party
  d. pants

- Emerging (6-8)
  What goes in the blank?
  John____ to the store yesterday
  a. was
  b. will go
  c. goes
  d. went

- Nearing (K-1)
  My ______ are too big
  a. shoe
  b. shoe's
  c. shoes

- Developing (9-12)
  Choose the correct word
  The plane disappeared ______ the clouds.
  a. onto
  b. on
  c. into
  d. to

- Nearing (9-12)
  From now on, Marco and Paula ______ to the cafeteria together everyday for lunch
  a. have gone
  b. will go
  c. went
  d. go
Appendix K (Item Format Description 6 Writing)

WRITING

ITEM FORMAT - Multiple Choice (Simple Sentence Stimulus with Combining Sentences)

Students are first presented two or three simple sentences as stimulus. Then they are prompted with a question that requires identifying similar to equal meaning sentences to the stimulus sentences. Test-takers then are required to combine those ‘stimulus sentences’ in order to construct a more complex sentence (compound, complex, or complex-compound sentences). Four assigned letter choices are given to students. The emphasized criterion can be sentence construction or meaning through the combination of two or three sentences. Also, while performing this task, the ability to use conjunctions to indicate relationship of ideas, intra-sentential structure, appropriate vocabulary, verb tense, and cohesion are qualities of the criterion tested.

➤ Emerging (4-5)

a. Ms. Johnson teaches both math and English
b. Ms. Johnson teaches math or English
c. Ms. Johnson doesn’t teach math or English
d. Ms. Johnson doesn’t teach math and English

➤ Developing (2-3)
Mary likes school. John doesn’t like school.
Which sentence means the same?

a. Mary and John like school
b. Mary and John don’t like school
c. Mary likes school but John doesn’t
d. John likes school but Mary doesn’t

➤ Maturing (6-8)
Choose the sentence that means the same thing.

One woman is weighing vegetables. The other two women are waiting for her to finish.

a. The women are buying vegetables
b. One woman is waiting and two women are weighing the vegetables.
c. One woman weighs the vegetables while the other two waits.
d. One woman is weighing, one is waiting, and the other one is watching.
ITEM FORMAT - Multiple Choice (Simple Sentence Stimulus/Correct Sentence Identification)

Students are first presented two or three simple sentences as a stimulus. Then they are prompted with a question that requires identifying the correct meaning sentence to the stimulus sentences. Four assigned letter choices are given to students. The emphasized criterion can be sentence construction, meaning, or inference.

- Emerging (4-5)
If someone asks “When does she leave?” you can answer:

a. In the kitchen  
b. At four o’clock  
c. By the door  
d. For Mexico

- Developing (2-3)

| Picture of a sun, moon, and earth |

The sun is very large. The earth is smaller than the sun. the moon is smaller than the earth.  
Which sentence is correct?

a. The sun is smaller than the moon  
b. The moon is larger than the earth  
c. The earth is smaller than the sun  
d. The earth is smaller than the moon

- Maturing (6-8)
What sentence describes what happened yesterday?

a. The boy is going home  
b. The boy will go home  
c. The boy went home  
d. The boy is home
Appendix M (Item Format Description 8 Writing)

WRITING

ITEM FORMAT - Sentence Completion (Dichotomous-Short Answer 1) NO Picture

Students are first presented with a fragment of a sentence as a stimulus. Then they complete the sentence with a missing blank that is placed at the end of the sentence. The expected response must be in a right/wrong format with a short answer response. An underlined big stimulus can be used such as describing a procedure, phenomenon; so individual sentences can be generated from this long stimulus. If the complexity of the question is higher, the test-taker may produce a more complex & larger response. The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters.

➢ Developing (4-5)

Q1 Write a word that fits the sentence

Today, I am walking to class with Sarah. Yesterday, I ______ to class with Juan.

Q2 Read the following paragraph and fill in the blank

First, I get a piece of paper. __________, I write my name. Last, I write the date.
Appendix N (Item Format Description 9 Writing)

WRITING

ITEM FORMAT - Short Answer 1 or 2 (Sentence Structure/Single Sentence)

This item format does not need a stimulus since a stem is only needed in order to elicit the test-taker's response. Students are shown a question asking them to tell/describe something about a specific word/item/situation. Students must respond the prompt with a complete sentence or correct response. Depending on the nature of the prompt, test-takers may be also asked to explain, be able to use proper sentences according to time frames (Correct use of present progressive, simple present, simple past, simple future tenses in sentences) and even compare and contrast a particular aspect of something, given their background information and as depicted in the stem (prompt).

- Emerging (6-8)
  What is the best sport in the world?
  Explain why you think so

- Developing (6-8)
  Compare yourself with someone. It can be a brother, sister, or a friend. How are you different? How are you the same?

- Maturing (K-1)
  Write to tell what you like

- Nearing (K-1)
  If you only have one piece of paper, but you need two, what can you do?

- Native-Like (6-8)
  Write about a place you visited. Tell why you liked it.
Appendix O (Item Format Description 10 Writing)

WRITING

ITEM FORMAT - Short Answer 1 or 2 (Picture Stimulus-Single Word / Phrase / Letter Response)

Students are shown a picture in order to trigger some background information. This visual or illustration is a key complement for understanding the stem, thus the visual(s) used is (are) critical to the test-takers to aid in their understanding of what is expected from them. After having analyzed the picture or series of pictures, students are prompted to write (type) a complete phrase or word describing what mirrors best the picture(s). Also, Students are shown a question asking them to tell/describe something about a specific word/item/situation. Students must respond the prompt with a complete sentence or correct response. Depending on the nature of the prompt, test-takers may be also asked to identify specific vocabulary or recognize letters (upper & lower). This means to be able to use mechanics in writing properly.

- Entering (K-1)
  Picture of:
  \[m\]
  Write the uppercase letter (big letter) for the picture

- Emerging (2-3)
  Picture of a snake
  sleeping
  What is this?
  Picture of puppies
  Write the word that describes more than one puppy

(6-8)
  Picture of woman with a cart in a grocery store grabbing a box
  Write a sentence to tell what the person is doing

- Developing (2-3)
  Picture of foot ball
  Picture of tennis racket and ball
  Picture of basket ball
  What are these pictures about?

- Emerging (4-5)
  Look at the picture
  Picture of an apple
  This is an apple. Write a word that describes or tells about the apple

- Maturing (K-1)
  Picture of \[rAT\]
  Write the letter that is lowercase or small
Appendix P (Item Format Description 11 Writing)

WRITING

ITEM FORMAT - Short Answer 1 or 2 (Picture Stimulus-Sentence Structure / Single Sentence)

Students are shown a picture or series of pictures in order to trigger some background information and then prompted to write a complete sentence describing what is happening in the picture(s). This visual or illustration is ‘key complement for understanding the stem, thus the visual(s) used is (are) critical to the test-takers to aid in their understanding of what is expected from them. Depending on the nature of the prompt, test-takers may be also asked to explain, describe past experiences or future events. This means to be able to use proper sentences according to time frames (Correct use of present progressive, simple present, simple past, simple future tenses in sentences), and even compare and contrast a particular feature as depicted in the illustrations.

- **Entering (4-5)**
  
  Picture of a boy wearing tennis shoes and a school track uniform. The boy is pictured as in motion of running. The setting is a school track.

  Write about what the boy is doing

- **Maturing (2-3)**

  Picture of a recess situation in a school. A student is playing basketball while another one plays with a yoyo. Also, a girl is jumping the rope and a boy plays with a musical triangle.

  Some other students play with toys as well.

  Describe where the students are and what they are doing.

- **Nearing (4-5)**

  A picture of a girl in an art smock with paint, a paintbrush, and a piece of paper on an easel. The girl is standing in front of the easel.

  A picture of a girl standing in front of the easel using her paintbrush. An unfinished painting is on the easel.

  A picture of a girl in an art smock with a piece of finished painting on an easel. The girl has paint on her smock and cheek

  Look at the pictures and tell what happened in picture two.

- **Native-Like (K-1)**

  Picture of a girl reading

  Write about this picture

- **Nearing (9-12)**

  Look at the picture

  Picture of a human eye. Under the picture the word sight is written

  After you saw the picture, think about your favorite place and answer the question.

  What do you see?

- **Emerging (6-8)**

  Look at the picture

  Picture of a woman in the grocery store. She wears glasses, a stripped dress and earrings. She has a coat with various groceries (fruit, cereal, vegetables, etc.) she is grabbing a bar of cheese from a shelf that says: butter and cheese. Behind her, there is a sigh that depicts 'super market'

  Write what the person is doing?
Appendix Q (Item Format Description 12 Writing)

WRITING

ITEM FORMAT - Short Answer 1 or 2 (Picture Stimulus-Sentence Completion)

Students are first presented with a picture as stimulus. This picture(s) is ‘key complement for understanding the stem or passage, that is the visual(s) used with the stems are critical for the test-taker to aid in their understanding of what is expected from them. After having analyzed the picture or series of pictures, they complete a sentence (stem) with a missing blank that is placed at the beginning, middle or end of the sentence. The expected response must be in a right/wrong format with a short answer response. If the complexity of the question is higher, the test-taker may produce a more complex & larger response. The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters.

➢ Entering (K-1)
Look at the picture

<table>
<thead>
<tr>
<th>Picture of a boy on a slide</th>
<th>Picture of a dog sleeping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture of a girl on swings</td>
<td>Picture of a cat in the tree</td>
</tr>
</tbody>
</table>

Write the answer to complete each sentence

_______ is on the swings
The ________ is sleeping
The ________ is in the tree
_______ is playing on the slide

➢ Maturing (2-3)
Fill in the blank
This book belongs to Bill. It is ______ book.

➢ Developing (4-5)
Fill in the blank
The flag is ________ in the wind

➢ Maturing (6-8)
Look at the picture

| Picture of a man placing his hand by his ear as a sign to being able to hear something better. There is a tree close by and some birds singing. |

Complete the following sentence with the correct word

What is the man doing?
He is ____________ the birds sing

➢ Emerging (9-12)
Look at the picture

| Picture of a boy writing on a piece of paper |

Complete the sentence to describe the picture

The student is __________ on a piece of paper
### Appendix R. Writing Data Display with Item Formats with Positive Comments.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Item Formats with Positive comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF diff</td>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (K-12)</td>
</tr>
<tr>
<td></td>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (K-12)</td>
</tr>
<tr>
<td></td>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (K-1/6-12)</td>
</tr>
<tr>
<td></td>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (2-12)</td>
</tr>
<tr>
<td></td>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (K-12)</td>
</tr>
<tr>
<td></td>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (K-5/9-12)</td>
</tr>
<tr>
<td></td>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-1/6-12)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gram</th>
<th>Item Formats with Positive comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (2-5/9-12)</td>
</tr>
<tr>
<td></td>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (K-12)</td>
</tr>
<tr>
<td></td>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (K-1/6-12)</td>
</tr>
<tr>
<td></td>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (2-3/6-12)</td>
</tr>
<tr>
<td></td>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (K-12)</td>
</tr>
<tr>
<td></td>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (2-3/6-8)</td>
</tr>
<tr>
<td></td>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-12)</td>
</tr>
<tr>
<td>IF/TI↔Relationship</td>
<td>IF↔Comprehensibility</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Voc</td>
<td>Mop</td>
</tr>
<tr>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (2-3-9-12)</td>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (2-3-9-12)</td>
</tr>
<tr>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (2-3/6-12)</td>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (2-3/6-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (2-12)</td>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (2-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (2-12)</td>
<td>- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (2-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (9-12)</td>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (9-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus w/Combining Sentences) (K-3/6-12)</td>
<td>- Multiple Choice (Simple sentence stimulus w/Combining Sentences) (K-3/6-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (2-5)</td>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (K-3/6-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (2-8)</td>
<td>- Short Answer 1 or 2 (Sentence Completion-Dichotomous-No Picture) (K-5)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (K-12)</td>
<td>- Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (K-3/9-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus - Sentence Structure / Single Sentence) (2-5/9-12)</td>
<td>- Short Answer 1 or 2 (Picture Stimulus - Sentence Structure / Single Sentence) (2-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-3/6-12)</td>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-12)</td>
</tr>
<tr>
<td>RLF</td>
<td>IF→ESL Sts</td>
</tr>
<tr>
<td>-----</td>
<td>-----------</td>
</tr>
<tr>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (2-3/6-12)</td>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (K-3/6-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (6-12)</td>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (K-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (2-3/6-12)</td>
<td>- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (K-3/6-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (2-3)</td>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (K-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus w/Combining Sentences) (K-3/6-12)</td>
<td>- Multiple Choice (Simple sentence stimulus w/Combining Sentences) (K-3/6-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (2-8)</td>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (K-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (2-12)</td>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (K-1/4-5/9-12)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (K-3/6-12)</td>
<td>Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (2-3/6-12)</td>
</tr>
<tr>
<td>Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-3/6-8)</td>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-3/6-12)</td>
</tr>
</tbody>
</table>
- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (K-12)
- Essay (Extended Response w/Picture, Diagram Stimulus) (2-3/6-12)
- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (4-12)
- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (K-3/6-12)
- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (4-12)
- Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (K-1/6-8)
<table>
<thead>
<tr>
<th>ILF</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (K-12)</td>
</tr>
<tr>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (2-3/9-12)</td>
</tr>
<tr>
<td>- Short Answer 1 (Sentence Completion-Dichotomous-No Picture) (9-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (9-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus - Sentence Structure / Single Sentence) (9-12)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IF/TL+PreTeach</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (K-12)</td>
</tr>
<tr>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (K-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (K-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (2-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Sentence Completion-Speech Parts Construction) (K-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus w/Combining Sentences) (K-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (K-12)</td>
</tr>
<tr>
<td>- Short Answer 1 (Sentence Completion-Dichotomous-No Picture) (K-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (K-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (K-1/4-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus - Sentence Structure / Single Sentence) (K-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (2-12)</td>
</tr>
<tr>
<td>C&amp;C</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>- Essay (Narrative, Imaginative, Informational, Persuasive, Descriptive – Extended Response) (K-5/9-12)</td>
</tr>
<tr>
<td>- Essay (Extended Response w/Picture, Diagram Stimulus) (2-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Sentence Completion) (2-5/9-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Picture/diagram stimulus w/Answer Questions) (9-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus w/Combining Sentences) (K-3/6-12)</td>
</tr>
<tr>
<td>- Multiple Choice (Simple sentence stimulus – correct sentence identification) (4-8)</td>
</tr>
<tr>
<td>- Short Answer 1 (Sentence Completion-Dichotomous-No Picture) (9-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Sentence Structure / Single Sentence) (2-3/6-8)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus – Single Word / Phrase / Letter Response) (K-3/6-12)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus - Sentence Structure / Single Sentence) (2-3/6-8)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus - Sentence Completion) (K-1/6-8)</td>
</tr>
<tr>
<td>- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion) (K-1/6-8)</td>
</tr>
</tbody>
</table>
Appendix S. Sample RE Item Test Specification
Detailed English Language Proficiency Item Test Specification Model.

Mandate: NCLB, Oregon English Language Proficiency Standards, & Test Construct.

- Specification Number: Will coordinate with the ODE numbering scheme
- Title of Specification: A short title should be given that generally characterizes each spec. The title is a good way to outline skills across several specifications.
- Related Specification(s) if any: List the number and/or titles of specs related to this one. For example, in a reading test separate detailed specifications would be given for the passage and each item.

- General Description (GD): A brief general statement of the behavior to be tested. The GD is very similar to the core of a learning objective. The purpose of testing this skill may also be stated in the GD. The wording of this does not need to follow strict instructional objective guidelines.

For the ODE assessment, the GD will be a combination of the identified standard, the identified function and the form.

- Prompt Attribute (PA): A complete and detailed description of what the student will encounter.
- Response Attribute (RA): A complete and detailed description of the way in which the student will provide the answer; that is, a complete and detailed description of what the student will do in response to the prompt and what will constitute a failure or success. There are two basic types of RA. Selected Response (notes the choices must be randomly rearranged later in test development): Clear and detailed descriptions of each choice in a multiple-choice format. Constructed response: A clear and detailed description of the type of response the student will perform, including the criteria for evaluating or rating the response.

- Sample item (SI): An illustrative item or task that reflects this specification, that is, the sort of item or task this specification should generate.

- Specification Supplement (SS): A detailed explanation of any additional information needed to construct items for a given spec. In grammar tests, for example, it is often necessary to specify the precise grammar forms tested. In a vocabulary specification, a list of testable words might be given. A reading specification might list in its supplement the textbooks from which reading test passages may be drawn.
MCh & SA1-2 Points (Sentence Completion)
Speech Parts Construction

**Mandate**
NCLB, Oregon English Language Proficiency Standards, & ELP Test Construct

**Construct**
Writing

**Language Functions, Language Competence, & Language Forms**
Representational Functions, Grammatical Competence & Mechanics
S-V-A, Pronouns, Nouns, Adjectives, Adverbs (all), Modals, Comparative/Superlative
Adjectives, Verb & Verb Tense (all), Specific Vocabulary

**General Description**

**Identified Oregon ELP Assessable Standards for Item Writing**
Simple to varied sentences: Past tense, irregular verbs, adverbs, prepositions, conjunctions; uses vocabulary for simple and more complex sentences; prepositions, appositives, subordinate clauses.
Short to more complex sentences: Coordination and subordination; present perfect; past perfect; future perfect; compound subjects; active and passive voice; subject, object, possessive pronouns; comparative adjectives and adverbs, parts of speech, types and structures of sentences; regular past, future, past continuous; varied sentence types, simple, compound, complex, compound-complex; conjunctions that indicate relationship between ideas.
Past, future tense; conditional.
Variety of verbs, adverbs to indicate passage of time
Conjunctions, dependent and independent clauses, gerund, infinitive, participial phrases
Descriptive language—physical appearance, actions, feelings. Action verbs, adjectives, sensory details, modifiers.
Describe sights, sounds, smells, actions, movements, gestures, feelings.
Regular past, future, past continuous; varied sentence types, simple, compound, complex, compound-complex; conjunctions that indicate relationship between ideas.
Adjectives, action verbs, varied word choices. Subject-verb agreement, tense, pronouns, possessives.
Uses contractions, demonstrates understanding of components

English Language Learners need to manipulate Grammatical Competence, which deals with the knowledge of grammar and vocabulary at a sentence level. It also enables the building and recognition of well-formed, grammatically correct utterances, according to the rules of syntax, semantics, morphology, and phonology/graphology. Thus, as depicted in the above Oregon ELP Standards, the appropriate command of sentences and speech parts is critical in the development of Linguistic Competence. Test-takers will be tested in their ability to construct simple to more complex sentences by demonstrating the appropriate use of verbs, adjectives, adverbs, prepositions, nouns, pronouns, etc. Mechanics is another component (as part of linguistic competence) that allows knowledge of alphabet & letters while forming words.
In addition, while employing grammatical competence, speakers perform language functions in order to utilize language appropriately. Representational functions and speech parts allow describing, and conveying information needed to communicate concrete and abstract concepts. Further, these two components of language help English Language Learners build the basis to operate in an academic setting and employ strong written discourse in English.
**Prompt Attribute**

**Sentence Fragment Input (MCh)**

Students are first presented with a sentence as stem, no stimulus is needed. Then, they complete that stem with a missing blank that could be placed in the beginning, middle or end of the sentence. Four assigned letter choices are given to students. The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters. Students need to fill in the blank with the appropriate grammatical speech part given the nature/context of the sentence that can range from simple vocabulary (nouns/pronouns, prepositions, conjunctions) identification to the appropriate inflection of verbs (time frames) and adjectives (comparatives/superlatives), as well as modifiers (adverbs).

*One of the distractors* will contain the correct choice by accurately representing the appropriate speech part used to complete the sentence. That is the correct verb inflection (time frame: simple present, present perfect, etc.), the correct adjective inflection (comparatives & superlatives), the correct conjunction used to link two phrases or sentences, the correct adverb modifying a verb (adverbs if time, mood, quantity, etc.), the correct vocabulary selection, singular/plural structures (nouns, pronouns, reflexive adjectives), the correct use of prepositions and letters.

*A second distractor* will contain an opposite response but still this choice will be part of the same speech part category (e.g. Correct response: *rides* – Opposite incorrect response: *buys*).

*A third distractor* will contain a speech part within the same category (verbs, prepositions, nouns, etc.) but will not be extremely opposite in meaning to the right answer but incorrect (e.g. Correct response: *trees* – Incorrect response: *tree’s*).

Finally, *a fourth distractor* will contain a speech part in the same category as the other three choices but will be different on its morphology. This, leading to have a different meaning & function as required in the context of the sentence (e.g. Correct answer: *parties* or *traveled* – Different morphology answer: *part* or *traveling*).

When the items are created from this spec, the order of choices must be randomized.

**Characteristics of the prompts (stems)**
- Stem can be written in present progressive, simple future, present perfect, simple present, simple past
- Stems can be questions
- Stems can be structured as complex sentences (using conjunctions), or two simple sentences
- Can be a de-contextualized statements / quotations (ex. ‘The secretary asked: “why ____ you late?”’
- Blank spaces in the stem can contain two words (ex. *where is the sun?* or I have been in school for two years)
- Stems can have up to three simple sentences or two complex sentences
- Pre-instructions are used to guide test-takers how to answer the stem.

**Visual Aided Sentence Fragment Input (MCh)**

The criterion/skills reflected in this item format are the same as those described above. Students are first presented with a picture as stimulus. This picture(s) is ‘key complement for understanding the stem, that is the visual(s) used with the stems are critical for the test-taker to aid in their understanding of what is expected from them.

After having analyzed the picture or series of pictures, they complete a sentence (stem) with a missing blank that is placed at the beginning, middle or end of the sentence. Four assigned letter choices are given to students. The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters. Students need to fill in the blank with the appropriate grammatical speech part given the nature/context of the sentence that can range from simple vocabulary (nouns/pronouns, prepositions, conjunctions) identification to the appropriate inflection of verbs (time frames) and adjectives (comparatives/superlatives), as well as modifiers (adverbs). In addition, given the properties of this item format, test-takers can be assessed in their ability to identify and/or place letters in the appropriate gap. Therefore, this skill is concerned with the notion of Mechanics. See descriptions of the distractors from above item format.
**Characteristics of the Stimulus**
- Single picture or Series of pictures. Pre-instructions can be given to students (before the stimulus). Ex. Fill in the blank…, Look at the pictures & choose the correct answer…

**Characteristics of the prompts (stems)**
- Stated in simple present and past tenses.
- Single sentences: the dog is ______
- Affirmative sentences: I _____ a banana
- Imperative sentences: be ___________!
- Complex sentences: when it is cold, I put on a _________
- Can be a single word: d___ck
- Spaces can be filled at the beginning, middle or end of the sentence. They are one word or up to three words (ex. prepositions) or prepositional phrases (ex. below the garden)

**Sentence Fragment Input (SA1-2)**
Students are first presented with a fragment of a sentence as a stimulus. Then they complete the sentence with a missing blank that is placed at the end of the sentence. The expected response must be in a right/wrong format with a short answer response. An underlined big stimulus can be used such as describing a procedure, phenomenon; so individual sentences can be generated from this long stimulus. If the complexity of the question is higher, the test-taker may produce a more complex & larger response (SA2). The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters.

**Visual Aided Input (SA1-2)**
Students are first presented with a picture as stimulus. This picture(s) is ‘key complement for understanding the stem or passage (if any)’, that is the visual(s) used with the stems are critical for the test-taker to aid in their understanding of what is expected from them. After having analyzed the picture or series of pictures, they complete a sentence (stem) with a missing blank that is placed at the beginning, middle or end of the sentence. The expected response must be in a right/wrong format with a short answer response. If the complexity of the question is higher, the test-taker may produce a more complex & larger response (SA2). The emphasized criterions are parts of speech such as construction of verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, and letters.

Characteristics of the stimulus
- Pictures or Series of Pictures: from people executing specific actions, outdoor activities, sports, school supplies, flags, computers, natural places (ex. pond), houses, animals, bicycles, math symbols (X, 0)
- Includes instructions that focus on picture + stem

**Characteristics of the prompts (Stems)**
- Can be structured in simple present:
- Incomplete statement followed by set of instructions
- 1 sentence (as sample, making reference to a picture) + 1 incomplete sentence
- Exclamatory sentences
- Complete /Simple sentences
- WH questions

The gaps to be filled in the stem can be located at the beginning, middle or end of the sentence. Stems can come as an adjacency pair format (that is question + answer, command + response, etc.), or with a combination of two clauses
Item Formats Used
- Multiple Choice (Sentence Completion-Speech Parts) – (verbs, adjectives, adverbs, conjunctions, nouns, prepositions, reflexive adjectives, & letters)
- Multiple Choice (Picture/Diagram Stimulus w/Sentence & Word Completion)
- Short Answer 1 (Sentence Completion-Dichotomous-NO Picture)
- Short Answer 1 or 2 (Picture Stimulus-Sentence Completion)

Response Attribute
Sentence Fragment Input (MCh)
The test-takers will select the correct answer from the choices given by clicking the letter of the correct choice (if computerized testing) or by circling the right choice or bubble in the answer sheet (if paper & pencil delivery). Test-takers can refer to the question/statement (stem) as needed when responding.

Characteristics of the foils (response choices provided)
- Foils can be question tags, nouns, adjectives, verbs + auxiliary verb (ex. will miss), complex verb usage (ex. would have eaten), and other described speech parts.

Visual Aided Sentence Fragment Input (MCh)
The test-takers will select the correct answer from the choices given by clicking the letter of the correct choice (if computerized testing) or by circling the right choice or bubble in the answer sheet (if paper & pencil delivery). Test-takers can refer to the question/statement (stem) as needed when responding.

Characteristics of the foils (response choices provided)
- Foils can be single words, colors, pronouns, verbs, adjectives, prepositions, conjunctions, adverbs of time, and even single letters, possessive adjectives (her, them, him, etc.)
- Foils are stated in simple present, past, present progressive, present perfect.

Sentence Fragment Input (SA1)
After having read the stem (incomplete sentence or statement), test-takers will complete the sentences with the appropriate one, two or three word response depending on the characteristic of the criterion tested (e.g. One word correct response: then; Three word correct response: will be walking).
The expected response for this item will be incorrect response (zero points granted) or correct response (one point granted).

General rating scale:
0 → None/wrong word
1 → Correct word/misspelled

Visual Aided Input (SA1-2)
After having analyzed the visual stimulus and read the stem (incomplete sentence or statement), test-takers will complete the sentences with the appropriate short answer response (1 point-right/wrong answer) or (2 point-added value points) depending on the complexity of the stem and the criterion tested. If an item is targeted to a two-value point, test-takers may obtain full credit by reaching all the assessment tasks as portrayed on a general rating scale.

Nevertheless, some students may not fully obtain the expected answer but a partial response so in that case; one point value can be applied. If the response is totally incorrect or no response is observed at all, examinees will not get credit (zero points). Responses will vary depending on what speech part is used. Thus, it can range from a single word to a more complex use of a proper inflection (e.g. Auxiliary + main verb). The level of difficulty of the response will depend on the grade group and the language proficiency level in which the test taker performs.

The responses obtained will be stored, and rated at a later time by trained raters on a ‘distributed scoring’ method if an item reflects a 2-point short answer format. However it the item mirrors a right
or wrong answer only, the response will be stored and graded automatically by the computer.

Depicted below is a general rating scale on the expected responses considering both SA1 and SA2 responses. It will be applicable according to the nature of the item and the criterion selected.

**General rating scale:**
SA1 (Right – Wrong Answer)

<table>
<thead>
<tr>
<th>0 → Incorrect</th>
<th>1 → Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>Correct preposition, singular/plural noun, single adjective (other options are possible)</td>
</tr>
<tr>
<td>Incorrect response</td>
<td>Single verb + preposition</td>
</tr>
<tr>
<td></td>
<td>Appropriate time frame (action verb-ing, simple present, 3rd person)</td>
</tr>
<tr>
<td></td>
<td>Acceptable spelling</td>
</tr>
<tr>
<td></td>
<td>Appropriate inflections (base word) – listening-listen</td>
</tr>
<tr>
<td></td>
<td>Combination of vowels/consonants that resemble the word</td>
</tr>
</tbody>
</table>

SA2 (Value points given)

<table>
<thead>
<tr>
<th>0 → Incorrect</th>
<th>1 → Correct</th>
<th>2 → Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>Verb in present (write) vs. (writing)</td>
<td>Base word verb in past (picked) vs. (picked up, fished for)</td>
</tr>
<tr>
<td>Incorrect response</td>
<td>Single verb, adjective</td>
<td>Auxiliary + main verb (proper inflection)</td>
</tr>
<tr>
<td>Scribbles</td>
<td>Wrong word misspelled</td>
<td>Complete 2 adjective combination</td>
</tr>
<tr>
<td></td>
<td>Logical vocabulary</td>
<td>Appropriate comparative structure</td>
</tr>
<tr>
<td></td>
<td>Non-appropriate comparative to the picture</td>
<td>Correct word</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Appropriate spelling</td>
</tr>
</tbody>
</table>

**Sample Item**

**Sentence Fragment Input (MCh)**

**Beginning (4-5)**

Do you want apples ___ bananas?

a. to
b. or
c. for
d. next

**Early Intermediate (9-12)**

Fill in the blank

Joe is _________ to the dance tonight

a. going
b. goes
c. go
d. gone
Intermediate (9-12)

Choose the correct word

The plane disappeared _________ the clouds.
   a. onto
   b. on
   c. into
   d. to

Early Advanced (9-12)

Fill in the blank

“This is the ___________ day of my life”, cried the boy
   a. worse
   b. very bad
   c. worst
   d. bad

Advanced (9-12)

From now on, Marco and Paula ________ to the cafeteria together everyday for lunch
   a. have gone
   b. will go
   c. went
   d. go

Proficient (9-12)

Choose the correct response

Sue ate too much candy, and feels sick. She now wishes she _____ so much
   a. would have eaten
   b. should have eaten
   c. had not eaten
   d. should not have eaten

Visual Aided Sentence Fragment Input (MCh)

Beginning (4-5)

Picture of a table. A bird is flying while a cat is sitting on the table. Under the table, a dog is barking to the cat.

Choose the correct answer

The dog is ____ the table
   a. under
   b. above
   c. next to
   d. on
Early Intermediate (2-3)

Picture of a dog sleeping

Fill in the blank

The dog is ____
  a. climbing
  b. walking
eating
d. sleeping

Intermediate (3-4)

Picture of a girl with school books in her arms

She ____ some books
  a. have
  b. has
c. carry
d. look

Early Advanced (2-3)

Picture of a student handing his completed homework to the teachers

Fill in the blank

Andy ____ his homework
  a. done
  b. finish
c. finished
d. do

Advanced (2-3)

Picture of a boy

Fill in the blank

John ____ a girl
  a. isn’t
  b. can’t
c. won’t
d. didn’t

Proficient (K-1)

Picture of a duck

What letter goes in each word

D__ck
  a. a
  b. i
c. u

Sentence Fragment Input (SA1-2)

Intermediate (4-5)

Q1 Write a word that fits the sentence

Today, I am walking to class with Sarah. Yesterday, I ______ to class with Juan.
Q2 Read the following paragraph and fill in the blank

First, I get a piece of paper. __________, I write my name. Last, I write the date.

Visual Aided Input (SA1-2)
Beginning (K-1)
Look at the picture

| Picture of a boy on a slide | Picture of a girl on swings | Picture of a dog sleeping | Picture of a cat in the tree |

Write the answer to complete each sentence

__________ is on the swings

The ________ is sleeping

The ________ is in the tree

__________ is playing on the slide

Early Intermediate (2-3)

| Picture of two girls playing basketball |

The girls are __________ basketball

Intermediate (4-5)

| Picture of an US flag that is waving |

Fill in the blank

The flag is _______ in the wind

Early Advanced (2-3)

| Picture of a boy holding a math book |

Fill in the blank

This book belongs to Bill. It is ______ book.

(K-1)

| Picture of a frog hopping, fish swimming, bees buzzing, friends playing. There is a pond and a house as well. |

Write the missing word for each sentence

The frog __________

The fish _________

The friends _________
Proficient (K-1)
Look at the picture below

| Picture of a girl going down a slide in a park. |

Now fill in the blank

She ________ down the slide

**Contexts**
Some of the topics that can be used to shape the context of the items are talking about/describing animals, school situations, geography (talking about cities), comparing people’s appearance and personality, health, teacher instructions and commands, family situations, favorite school subjects, sports, types of food, talking about the alphabet, means of transportation, at the library, remembering past events, talking about present and future events, school schedules, museums, at the cafeteria, activities performed by people, at the lab, school equipment, at recess, national symbols, the five senses, professions, at the park, etc.

In order to produce set the context in which the items will take place, it is strongly encouraged to consider those contexts that pertain academic settings (math, language arts, social studies, and science). In terms of the language used, most of these items can fall under the Social Language in School and the Academic Language not related to Explicit Content. Further, Supplementary Contexts can be used in order the set the stage for the context of the items (art class, music, cafeteria, recess, sports, etc.), even tough if some of these items fall into the Social Language Outside School, it is important to relate the language or vocabulary to an academic context (math, social studies, etc.).

**Grade Groups**
The Sentence Fragment Input (MCh/SA1-2) prompts can be targeted mainly to grades in the following grade levels: 2-3, 4-5, 9-12, 6-8 (all proficiency levels)

The Visual Aided Sentence Fragment Input (MCh) can be targeted mainly to grades in the following grade levels: 2-3, 4-5, 9-12, 6-8 (all proficiency levels)

The Visual Aided Input (SA1-2) prompts can be targeted mainly to grades in the following proficiency levels: 2-3, 4-5, 9-12, 6-8 (all proficiency levels)

**Specification Supplement**
In this section (1) of the specification supplement are incorporated the Oregon English Proficiency Standards as used to select standards for item writing purposes. They represent what the content panel and the ELP test team arrived to a consensus as to assessable-manageable in a testing situation as well as the incorporation of the alignment to the Construct of the test.
Thus, Grammatical competence (X), Textual Competence (T), and Functional Competence are critical pieces that define language proficiency. Under Functional Competence, language functions include Representational (R), Manipulative (M), Heuristic (H), and Imaginative (I). Finally, in order to make item writing more tangible-convenient for item writers and educators, some standards reflect ‘unique’ academic concepts and rationales that could happen in one grade only but more than one CCG. Similarly, ‘consolidated’ standards representing a specific concept/rationale covering many grades and CCGs. They are framed in this way in order to decrease the load of standards amount for item writing purposes.

In section (2), (3), and (4) is presented a list of possible criterions that can be generated from the item format in place. They are posted as a suggestion or a window of what these items can produce, as a result of having been written to the standards. They can provide ideas to item writers in their crafting and improvement of future tasks/items as generated by this spec (considering their item format).
### 1) Oregon English Language Proficiency Standards for Item Writing

#### Writing – K-1

<table>
<thead>
<tr>
<th>Grade</th>
<th>CCG</th>
<th>Standard</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>2</td>
<td>1</td>
<td>Writes letters; left to right</td>
<td>Cohesion, Mechanics</td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>2</td>
<td>Writes upper and lower case letters</td>
<td>Cohesion, Mechanics</td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>3</td>
<td>Writes own name</td>
<td>Cohesion, Mechanics</td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>5</td>
<td>Writes using letters that represent sounds in right order</td>
<td>Cohesion, Mechanics</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>Writes 1-3 word description of own picture</td>
<td>Representational</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>Writes in simple sentences</td>
<td>Grammar</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>G-1</td>
<td>Writes singular and plural nouns</td>
<td>Grammar</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>N-1</td>
<td>Writes stories that describe an experience</td>
<td>Imaginative</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>E-1</td>
<td>Writes descriptions</td>
<td>Representational</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>E-2</td>
<td>Writes directions</td>
<td>Manipulative</td>
</tr>
</tbody>
</table>

#### Writing – Grade 2-3

<table>
<thead>
<tr>
<th>Grade</th>
<th>CCG</th>
<th>Standard</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>7</td>
<td>Sentence level structures: Uses correct word order; produces “more complex sentences”; simple to more complex sentences; subject-verb agreement, tense, pronouns, possessives</td>
<td>Grammar</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>Uses contractions, demonstrates understanding of components</td>
<td>Grammar</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5,6, 10</td>
<td>Uses descriptive words</td>
<td>Grammar</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>G-1,2,3,4</td>
<td>Subject-verb agreement, tense, pronouns, possessives</td>
<td>Grammar</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>6</td>
<td>Uses descriptive words</td>
<td>Grammar</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>Writes paragraph with internal structure</td>
<td>Representational</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5</td>
<td>Subject-verb agreement, tense, pronouns, possessives</td>
<td>Grammar</td>
</tr>
</tbody>
</table>

#### Writing – Grade 4-5

<table>
<thead>
<tr>
<th>Grade</th>
<th>CCG</th>
<th>Standard</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>10,11</td>
<td>Simple to varied sentences: Past tense, irregular verbs, adverbs, prepositions, conjunctions; uses vocabulary for simple and more complex sentences; prepositions, appositives, subordinate clauses</td>
<td>Grammar</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>G-1,2,3,4,5</td>
<td>Descriptive words, such as sensory details</td>
<td>Represetational</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>5</td>
<td>Descriptive words, such as sensory details</td>
<td>Represetational</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>11</td>
<td>Transitions</td>
<td>Textual</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>N-3</td>
<td>Chronological order</td>
<td>Representational</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>N-3</td>
<td>Sensory details</td>
<td>Representational</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>6</td>
<td>Textual</td>
<td>Textual</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>8</td>
<td>Conjunctions</td>
<td>Textual</td>
</tr>
</tbody>
</table>
## Writing – Grade 6-8

<table>
<thead>
<tr>
<th>Grade</th>
<th>CCG</th>
<th>Standard</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>Short to more complex sentences: Coordination and subordination; present perfect; past perfect; future perfect; compound subjects; active and passive voice; subject, object, possessive pronouns; comparative adjectives and adverbs, parts of speech, types and structures of sentences; regular past, future, past continuous; varied sentence types, simple, compound, complex, compound-complex; conjunctions that indicate relationship between ideas</td>
<td>Grammar</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>12</td>
<td>Sentences for chronological order, cause and effect, similarity and difference. Descriptive words:</td>
<td>Textual</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>Action verbs, sensory details, modifiers</td>
<td>Grammar</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
<td>Adjectives, adverbs</td>
<td>Grammar</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
<td>Sensory details</td>
<td>Grammar</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>Descriptive language—physical appearance, actions, feelings</td>
<td>Grammar</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>10</td>
<td>Adjectives, action verbs, varied word choices; adverbs, phrases, modifiers; descriptive language, comparisons</td>
<td>Grammar</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>G-1</td>
<td>Sentences for chronological order, cause and effect, similarity and difference</td>
<td>Grammar</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>G-2</td>
<td>Descriptive words</td>
<td>Grammar</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>G-3</td>
<td>Indefinite pronouns</td>
<td>Grammar</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>5</td>
<td>Adjectives, action verbs, varied word choices</td>
<td>Grammar</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>7</td>
<td>Adverbs, phrases, modifiers</td>
<td>Grammar</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>G-1</td>
<td>Action verbs, sensory details, modifiers</td>
<td>Grammar</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>1, 2, 3, 4</td>
<td>Sentences for chronological order, cause and effect, similarity and difference</td>
<td>Grammar</td>
</tr>
</tbody>
</table>

## Writing – Grade 9-10

<table>
<thead>
<tr>
<th>Grade</th>
<th>CCG</th>
<th>Standard</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
<td>Action verbs, sensory details, modifiers</td>
<td>Representational</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>G-3</td>
<td>Adjectives, adverbs</td>
<td>Grammar</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>3</td>
<td>Sensory details</td>
<td>Imaginative</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>5</td>
<td>Descriptive language—physical appearance, actions, feelings</td>
<td>Imaginative</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>Action verbs, adjectives, sensory details, modifiers</td>
<td>Rep.</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>3</td>
<td>Describe sights, sounds, smells, actions, movements, gestures, feelings</td>
<td>Rep.</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>G-4</td>
<td>Past, future tense; conditional.</td>
<td>Grammar</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>4</td>
<td>Variety of verbs, adverbs to indicate passage of time</td>
<td>Imaginative</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>G-2</td>
<td>Conjunctions, dependent and independent clauses, gerund, infinitive, participial phrases</td>
<td>Grammar</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>G-1</td>
<td>Varied types of clauses</td>
<td>Rep.</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>P-3</td>
<td>Phrases, complex phrases and sentences</td>
<td>Rep.</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>G-2</td>
<td>Conjunctions, dependent and independent clauses, gerund, infinitive, participial phrases</td>
<td>Grammar</td>
</tr>
</tbody>
</table>
2) Sample Expected Skills from Sentence Fragment Input (MCh & SA1) Prompts
Following is a list of the general expected responses that this type of format generates (criterion expected from the test-taker).
- Identify correct use of possessive pronouns, adjectives, adverbs of mood, prepositions (ex. Unless), demonstrative adjectives, phrasal verbs (ex. get out)
- Letter identification
- Use of noun plurals in a sentence
- Complete question-construction through conjunctions
- Identify proper use of appositives (ex. Jill gave the folder to Bill, her trusted friend, as soon left the meeting)
- Know how to connect two sentences
- Use of conjunctions (and, or, however)
- Past participle verbs use, past tense verbs
- Identify proper use of adjective inflection (comparatives, superlatives)
- Identify and apply proper use of verb inflection/time frame (present, past, future, present progressive, present perfect)
- Identify and apply proper use of auxiliary verbs
- Identify proper use of verb conjugation/plural-singular (be) was-were
- Identify nouns (singular-plural relationships) man-men, city-cities
- Proper use of third person present simple verbs, proper use of present tense negative 3rd person
- Identify time frame (days of week)
- Correct past tense (--ed)
- Proper use of question tags
- Construct questions appropriately (ex. Where is the sun?)
- Correct use of particle ‘to’ on infinitive verbs

3) Visual Aided Sentence Fragment Input (MCh) Prompts
Following is a list of the general expected responses that this type of format generates (criterion expected from the test-taker).
- Identify colors, identify specific letters in words (letter recognition for K-1 items)
- Describe attributes of a picture (description)
- Identify appropriate pronouns, verbs, adjectives, nouns, vocabulary
- Identify appropriate verb usage and time frame (present, past, present progressive, verb to be conjugations), identify verb inflection (3rd person present – have/has), identify proper use of negative sentences/verbs (to be)
- Identify proper use of modals in a sentence (negative and affirmative way)
- Identify proper use of apostrophes in possessives (Ann’s …)
- Identify proper use of adjective inflections (comparatives)
- Identify proper use of adverbs of time, prepositions, conjunctions, possessive adjectives (ex. Them, her, him, his)
- Understand commands through sentence completion
- Infer vocabulary used in a picture
- Identify number of items in a picture (naming items)

4) Visual Aided Input (SA1-2) Prompts
Following is a list of the general expected responses that this type of format generates (criterion expected from the test-taker).
- Correct past tense use verb from picture, identify simple present verb tenses, use of verbs in present continuous, correct use of passive voice, correct conjugation of verbs, identify actions, adjectives
- Interpret prepositions by filling in the right noun blank
- Identify objects used by characters in situations, identify vocabulary from a specific object, identify colors
- Compare and contrast 2 concepts through adjectives
- Identify correct use of apostrophes (personal pronouns) Bill’s-his
- Identify plurals in irregular nouns (child-children)
- Identify proper use of comparatives
- Construct noun phrases (the girl is on the swings)
- Identify nouns, prepositions
### Appendix T. Relationship between Writing Item Formats and the Test Construct

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Writing 1</th>
<th>Writing 2</th>
<th>Writing 3</th>
<th>Writing 4</th>
<th>Writing 5</th>
<th>Writing 6</th>
<th>Writing 7</th>
<th>Writing 8</th>
<th>Writing 9</th>
<th>Writing 10</th>
<th>Writing 11</th>
<th>Writing 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>90%</td>
<td>96%</td>
<td>93%</td>
<td>73%</td>
<td>93%</td>
<td>85%</td>
<td>86%</td>
<td>81%</td>
<td>90%</td>
<td>89%</td>
<td>93%</td>
<td>100%</td>
</tr>
<tr>
<td>No</td>
<td>10%</td>
<td>4%</td>
<td>7%</td>
<td>27%</td>
<td>7%</td>
<td>15%</td>
<td>14%</td>
<td>19%</td>
<td>10%</td>
<td>31%</td>
<td>7%</td>
<td>0%</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>93%</td>
<td>92%</td>
<td>83%</td>
<td>88%</td>
<td>79%</td>
<td>92%</td>
<td>93%</td>
<td>73%</td>
<td>86%</td>
<td>96%</td>
<td>90%</td>
<td>100%</td>
</tr>
<tr>
<td>No</td>
<td>7%</td>
<td>8%</td>
<td>17%</td>
<td>12%</td>
<td>21%</td>
<td>8%</td>
<td>7%</td>
<td>27%</td>
<td>14%</td>
<td>4%</td>
<td>10%</td>
<td>0%</td>
</tr>
<tr>
<td>Morphology</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>79%</td>
<td>77%</td>
<td>96%</td>
<td>62%</td>
<td>88%</td>
<td>73%</td>
<td>83%</td>
<td>69%</td>
<td>79%</td>
<td>69%</td>
<td>78%</td>
<td>73%</td>
</tr>
<tr>
<td>No</td>
<td>21%</td>
<td>23%</td>
<td>10%</td>
<td>38%</td>
<td>14%</td>
<td>27%</td>
<td>17%</td>
<td>31%</td>
<td>21%</td>
<td>31%</td>
<td>24%</td>
<td>27%</td>
</tr>
<tr>
<td>Representational</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>97%</td>
<td>96%</td>
<td>72%</td>
<td>92%</td>
<td>75%</td>
<td>96%</td>
<td>93%</td>
<td>81%</td>
<td>90%</td>
<td>92%</td>
<td>90%</td>
<td>96%</td>
</tr>
<tr>
<td>No</td>
<td>3%</td>
<td>4%</td>
<td>28%</td>
<td>8%</td>
<td>24%</td>
<td>4%</td>
<td>7%</td>
<td>19%</td>
<td>10%</td>
<td>8%</td>
<td>10%</td>
<td>4%</td>
</tr>
<tr>
<td>Manipulative</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>54%</td>
<td>50%</td>
<td>43%</td>
<td>35%</td>
<td>34%</td>
<td>42%</td>
<td>59%</td>
<td>36%</td>
<td>55%</td>
<td>42%</td>
<td>52%</td>
<td>54%</td>
</tr>
<tr>
<td>No</td>
<td>46%</td>
<td>50%</td>
<td>57%</td>
<td>68%</td>
<td>68%</td>
<td>41%</td>
<td>62%</td>
<td>45%</td>
<td>58%</td>
<td>43%</td>
<td>46%</td>
<td>46%</td>
</tr>
<tr>
<td>Heuristic</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>64%</td>
<td>65%</td>
<td>64%</td>
<td>56%</td>
<td>59%</td>
<td>58%</td>
<td>59%</td>
<td>50%</td>
<td>55%</td>
<td>58%</td>
<td>55%</td>
<td>55%</td>
</tr>
<tr>
<td>No</td>
<td>36%</td>
<td>35%</td>
<td>36%</td>
<td>44%</td>
<td>41%</td>
<td>42%</td>
<td>41%</td>
<td>50%</td>
<td>45%</td>
<td>42%</td>
<td>45%</td>
<td>42%</td>
</tr>
<tr>
<td>Imaginative</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>62%</td>
<td>58%</td>
<td>31%</td>
<td>27%</td>
<td>24%</td>
<td>19%</td>
<td>34%</td>
<td>19%</td>
<td>59%</td>
<td>38%</td>
<td>45%</td>
<td>42%</td>
</tr>
<tr>
<td>No</td>
<td>38%</td>
<td>42%</td>
<td>69%</td>
<td>73%</td>
<td>75%</td>
<td>81%</td>
<td>81%</td>
<td>41%</td>
<td>62%</td>
<td>55%</td>
<td>56%</td>
<td>56%</td>
</tr>
<tr>
<td>Cohesion/Coherence</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Yes</td>
<td>86%</td>
<td>96%</td>
<td>62%</td>
<td>62%</td>
<td>65%</td>
<td>77%</td>
<td>72%</td>
<td>62%</td>
<td>83%</td>
<td>73%</td>
<td>69%</td>
<td>73%</td>
</tr>
<tr>
<td>No</td>
<td>14%</td>
<td>4%</td>
<td>38%</td>
<td>38%</td>
<td>34%</td>
<td>23%</td>
<td>28%</td>
<td>36%</td>
<td>17%</td>
<td>27%</td>
<td>31%</td>
<td>27%</td>
</tr>
</tbody>
</table>
### Appendix U. Mean Results from the Writing Yes/No Response Format Used in the Quantitative Questionnaire.

<table>
<thead>
<tr>
<th>writing 1</th>
<th>writing 2</th>
<th>writing 3</th>
<th>writing 4</th>
<th>writing 5</th>
<th>writing 6</th>
<th>writing 7</th>
<th>writing 8</th>
<th>writing 9</th>
<th>writing 10</th>
<th>writing 11</th>
<th>writing 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item format too difficult?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>34%</td>
<td>66%</td>
<td>31%</td>
<td>69%</td>
<td>17%</td>
<td>83%</td>
<td>31%</td>
<td>69%</td>
<td>31%</td>
<td>69%</td>
<td>38%</td>
</tr>
<tr>
<td>Item format comprehensible?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>86%</td>
<td>14%</td>
<td>96%</td>
<td>4%</td>
<td>90%</td>
<td>10%</td>
<td>88%</td>
<td>12%</td>
<td>93%</td>
<td>7%</td>
<td>92%</td>
</tr>
<tr>
<td>Items &amp; item format appropriate for ESL students?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>79%</td>
<td>21%</td>
<td>88%</td>
<td>12%</td>
<td>83%</td>
<td>17%</td>
<td>81%</td>
<td>19%</td>
<td>90%</td>
<td>10%</td>
<td>85%</td>
</tr>
<tr>
<td>Items &amp; item format fit proficiency level?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>79%</td>
<td>21%</td>
<td>77%</td>
<td>23%</td>
<td>90%</td>
<td>10%</td>
<td>80%</td>
<td>20%</td>
<td>79%</td>
<td>21%</td>
<td>73%</td>
</tr>
<tr>
<td>Tasks from item format match instructional tasks?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>25%</td>
<td>73%</td>
<td>27%</td>
<td>83%</td>
<td>17%</td>
<td>58%</td>
<td>42%</td>
<td>86%</td>
<td>14%</td>
<td>46%</td>
</tr>
<tr>
<td>Items &amp; tasks need to be explicitly pre-taught?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>10%</td>
<td>77%</td>
<td>23%</td>
<td>72%</td>
<td>28%</td>
<td>58%</td>
<td>42%</td>
<td>72%</td>
<td>28%</td>
<td>73%</td>
</tr>
<tr>
<td>Procedure of test taking need to be explicitly pre-taught?</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
<td>No</td>
<td>Mean:</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>79%</td>
<td>21%</td>
<td>81%</td>
<td>19%</td>
<td>72%</td>
<td>28%</td>
<td>73%</td>
<td>27%</td>
<td>72%</td>
<td>28%</td>
<td>69%</td>
</tr>
</tbody>
</table>
### Appendix V. Grade groups most appropriate for this writing modality

<table>
<thead>
<tr>
<th></th>
<th>W1</th>
<th>W2</th>
<th>W3</th>
<th>W4</th>
<th>W5</th>
<th>W6</th>
<th>W7</th>
<th>W8</th>
<th>W9</th>
<th>W10</th>
<th>W11</th>
<th>W12</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-1</td>
<td>2%</td>
<td>0%</td>
<td>9%</td>
<td>11%</td>
<td>6%</td>
<td>1%</td>
<td>1%</td>
<td>7%</td>
<td>9%</td>
<td>17%</td>
<td>5%</td>
<td>8%</td>
</tr>
<tr>
<td>2-3</td>
<td>19%</td>
<td>18%</td>
<td>21%</td>
<td>22%</td>
<td>19%</td>
<td>20%</td>
<td>19%</td>
<td>15%</td>
<td>16%</td>
<td>25%</td>
<td>23%</td>
<td>20%</td>
</tr>
<tr>
<td>4-5</td>
<td>24%</td>
<td>22%</td>
<td>25%</td>
<td>22%</td>
<td>25%</td>
<td>22%</td>
<td>25%</td>
<td>26%</td>
<td>23%</td>
<td>19%</td>
<td>25%</td>
<td>21%</td>
</tr>
<tr>
<td>6-8</td>
<td>24%</td>
<td>30%</td>
<td>24%</td>
<td>22%</td>
<td>24%</td>
<td>26%</td>
<td>28%</td>
<td>24%</td>
<td>26%</td>
<td>20%</td>
<td>23%</td>
<td>24%</td>
</tr>
<tr>
<td>9-12</td>
<td>31%</td>
<td>30%</td>
<td>21%</td>
<td>23%</td>
<td>26%</td>
<td>31%</td>
<td>27%</td>
<td>26%</td>
<td>26%</td>
<td>19%</td>
<td>24%</td>
<td>27%</td>
</tr>
</tbody>
</table>
### Appendix W. Likert Scale Results for the Relationship between Test and Construct

<table>
<thead>
<tr>
<th></th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 - Grammar</strong></td>
<td>agree &amp; strongly agree</td>
<td>89%</td>
<td>96%</td>
<td>97%</td>
<td>73%</td>
<td>90%</td>
<td>88%</td>
<td>90%</td>
<td>81%</td>
<td>83%</td>
<td>76%</td>
<td>90%</td>
</tr>
<tr>
<td><strong>2 - Vocabulary</strong></td>
<td>agree &amp; strongly agree</td>
<td>89%</td>
<td>92%</td>
<td>83%</td>
<td>88%</td>
<td>83%</td>
<td>85%</td>
<td>87%</td>
<td>77%</td>
<td>90%</td>
<td>100%</td>
<td>87%</td>
</tr>
<tr>
<td><strong>3 - Morphology</strong></td>
<td>agree &amp; strongly agree</td>
<td>78%</td>
<td>77%</td>
<td>93%</td>
<td>61%</td>
<td>76%</td>
<td>65%</td>
<td>83%</td>
<td>65%</td>
<td>80%</td>
<td>69%</td>
<td>79%</td>
</tr>
<tr>
<td><strong>4 - Representational functions</strong></td>
<td>agree &amp; strongly agree</td>
<td>90%</td>
<td>96%</td>
<td>69%</td>
<td>73%</td>
<td>69%</td>
<td>88%</td>
<td>87%</td>
<td>81%</td>
<td>90%</td>
<td>92%</td>
<td>90%</td>
</tr>
<tr>
<td><strong>5 - Manipulative functions</strong></td>
<td>agree &amp; strongly agree</td>
<td>55%</td>
<td>50%</td>
<td>52%</td>
<td>39%</td>
<td>41%</td>
<td>42%</td>
<td>49%</td>
<td>35%</td>
<td>56%</td>
<td>42%</td>
<td>52%</td>
</tr>
<tr>
<td><strong>6 - Heuristic functions</strong></td>
<td>agree &amp; strongly agree</td>
<td>50%</td>
<td>65%</td>
<td>55%</td>
<td>38%</td>
<td>52%</td>
<td>54%</td>
<td>56%</td>
<td>39%</td>
<td>51%</td>
<td>57%</td>
<td>52%</td>
</tr>
<tr>
<td><strong>7 - Imaginative functions</strong></td>
<td>agree &amp; strongly agree</td>
<td>64%</td>
<td>54%</td>
<td>24%</td>
<td>31%</td>
<td>21%</td>
<td>19%</td>
<td>28%</td>
<td>19%</td>
<td>63%</td>
<td>43%</td>
<td>38%</td>
</tr>
<tr>
<td><strong>8 - Cohesion &amp; Coherence</strong></td>
<td>agree &amp; strongly agree</td>
<td>89%</td>
<td>92%</td>
<td>59%</td>
<td>65%</td>
<td>50%</td>
<td>81%</td>
<td>65%</td>
<td>62%</td>
<td>80%</td>
<td>69%</td>
<td>66%</td>
</tr>
</tbody>
</table>
### Appendix X. Likert Scale Results for the Remaining Variables in the Questionnaire

<table>
<thead>
<tr>
<th>(a) Items contained in item format assess the purpose of the test</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>93%</td>
<td>88%</td>
<td>83%</td>
<td>77%</td>
<td>86%</td>
<td>92%</td>
<td>80%</td>
<td>77%</td>
<td>83%</td>
<td>84%</td>
<td>83%</td>
<td>92%</td>
</tr>
<tr>
<td>strongly disagree &amp; disagree</td>
<td>67%</td>
<td>73%</td>
<td>90%</td>
<td>69%</td>
<td>69%</td>
<td>66%</td>
<td>63%</td>
<td>69%</td>
<td>69%</td>
<td>81%</td>
<td>72%</td>
<td>80%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) The task/item format is too difficult or ambiguous for this level of student</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>89%</td>
<td>100%</td>
<td>93%</td>
<td>81%</td>
<td>97%</td>
<td>100%</td>
<td>83%</td>
<td>77%</td>
<td>83%</td>
<td>88%</td>
<td>90%</td>
<td>96%</td>
</tr>
<tr>
<td>strongly disagree &amp; disagree</td>
<td>93%</td>
<td>92%</td>
<td>97%</td>
<td>88%</td>
<td>93%</td>
<td>96%</td>
<td>79%</td>
<td>81%</td>
<td>79%</td>
<td>92%</td>
<td>87%</td>
<td>92%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) Test items fit the item format</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>75%</td>
<td>73%</td>
<td>86%</td>
<td>65%</td>
<td>76%</td>
<td>65%</td>
<td>59%</td>
<td>69%</td>
<td>77%</td>
<td>73%</td>
<td>72%</td>
<td>77%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(d) The item format is comprehensible</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>89%</td>
<td>84%</td>
<td>83%</td>
<td>70%</td>
<td>69%</td>
<td>81%</td>
<td>83%</td>
<td>69%</td>
<td>80%</td>
<td>88%</td>
<td>76%</td>
<td>88%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(e) The items and item format are appropriate for ESL students in this grade group</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>75%</td>
<td>58%</td>
<td>65%</td>
<td>50%</td>
<td>72%</td>
<td>46%</td>
<td>69%</td>
<td>58%</td>
<td>83%</td>
<td>81%</td>
<td>69%</td>
<td>65%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(f) Items and item format fit the selected proficiency level</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>93%</td>
<td>88%</td>
<td>83%</td>
<td>77%</td>
<td>86%</td>
<td>92%</td>
<td>80%</td>
<td>77%</td>
<td>83%</td>
<td>84%</td>
<td>83%</td>
<td>92%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(g) Items and item format are applicable and practical to use for this grade level</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>75%</td>
<td>85%</td>
<td>80%</td>
<td>61%</td>
<td>69%</td>
<td>73%</td>
<td>79%</td>
<td>81%</td>
<td>62%</td>
<td>50%</td>
<td>73%</td>
<td>65%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(h) The tasks represented in this item format match your typical instructional tasks on a daily basis</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>93%</td>
<td>88%</td>
<td>83%</td>
<td>77%</td>
<td>86%</td>
<td>92%</td>
<td>80%</td>
<td>77%</td>
<td>83%</td>
<td>84%</td>
<td>83%</td>
<td>92%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(i) Items/tasks from the item format need to be explicitly pre-taught to a given group of test-takers</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>89%</td>
<td>84%</td>
<td>83%</td>
<td>70%</td>
<td>69%</td>
<td>81%</td>
<td>83%</td>
<td>69%</td>
<td>80%</td>
<td>88%</td>
<td>76%</td>
<td>88%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(j) The procedure for test-taking needs to be explicitly pre-taught to a given group of test-takers</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>75%</td>
<td>81%</td>
<td>68%</td>
<td>76%</td>
<td>69%</td>
<td>73%</td>
<td>76%</td>
<td>81%</td>
<td>69%</td>
<td>70%</td>
<td>69%</td>
<td>61%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(k) Items and item format are appropriate in age and cognitive level of students at this grade band</th>
<th>W 1</th>
<th>W 2</th>
<th>W 3</th>
<th>W 4</th>
<th>W 5</th>
<th>W 6</th>
<th>W 7</th>
<th>W 8</th>
<th>W 9</th>
<th>W 10</th>
<th>W 11</th>
<th>W 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree &amp; strongly agree</td>
<td>77%</td>
<td>84%</td>
<td>90%</td>
<td>85%</td>
<td>79%</td>
<td>77%</td>
<td>73%</td>
<td>73%</td>
<td>86%</td>
<td>92%</td>
<td>76%</td>
<td>88%</td>
</tr>
</tbody>
</table>