

Original Test Development Project

Timothy Brockley

EDU570: Assignment 3

Professor Kathleen Bailey

October 5, 2009

Word Count: 3350

**Please view the online test and scoring instrument as a stimulus activator for this paper**

## [The Language Works Module Three Test One](#)

[The Language Works Module Scoring Instrument](#) (Ctrl + click to follow the link).

The test will take 20 ~ 50 seconds to open depending on your server speed.

### **Introduction**

The idea of creating and piloting a test for use in the curriculum I've been developing for young learners here in Korea has been with me for quite some time. The Language Works Online English Curriculum (referred to as [TLWOEC](#) Ctrl + click to follow link) employs a flash file format for each individual activity and a system of modules. In each level, these modules can be followed sequentially, but there is allowance and encouragement for variation in the selection and use of individual activities. In this way, student need drives the lesson planning, course organization and objectives, which are seen as dynamic and progressive. The focus is on the speaking and listening macro-skills, although the reading skill is necessary to participate fully in the lessons and tends to develop in consequence. The target student population is young learners of English (roughly between the ages of 7 and 15) in both ESL and EFL environments. As this process of assessment development evolves, more tests will be added to the curriculum.

### **Rationale**

The choice of a target audience coincides with the learners I'm presently teaching. Test methods grew out of the lesson files themselves which have been in development over the past six years. It's a process of trial and error. If the students are engaged and are using the language to communicate with each other and the teacher, the format and/or content is further

developed; on the other hand, if the format and/or content fail to engage the learners, the lessons are changed or abandoned.

The constructs for this pilot test present the vocabulary and grammatical structures as well as general content of the lessons that make up the curriculum. I will set out to determine the actual validity, reliability, washback and practicality of this test in the following section.

### **Report on the Administered Test**

Brown (2005) discusses factors leading to test score variance from two points of view: meaningful and error variance. Meaningful variance involves factors "directly attributable to the testing purposes" (p. 170) while error variance in test scores are "not directly related to the purposes of the test" (p. 171). I will first discuss meaningful variance in relation to the statistical information gathered in the tables below and then draw conclusions regarding error variance. The frequency polygons for the test (Table 1) as a whole and the three subtests all show similar ranges in test scores. This indicates meaningful variance is consistent in all subtests. Indeed the overlapping variance for the three subtests (Table 2) are all favorable (subtests one and two: 0.99, subtests one and three: 0.95, subtests two and three: 0.97). See also the trendlines and scatterplots related to this data (Tables 3—5). In general, the test succeeds in relation to overlapping variance.

Brown (2005) posits five general categories attributable to test scores and error variance: the environment, the administration procedures, the examinees, the scoring procedures and the test and test items (pp. 171—175). If we further speculate on the factors described in Brown's second source of variance, it's possible to claim there were no such

factors impinging on the test results. The positive results when applying overlapping variance are then a strong point of the test as a whole. The weakness in this study is that too few students (six total) were tested. The statistics in this case are not as reliable as the paradigm in which larger numbers of test-takers had participated. This brings up the issue of credibility. Larger numbers of examinees and the resulting data will be required in future analyses.

I will now focus on validity. Bailey (1998) explains it this way: "If a test actually measures what it is intended to measure, we say it is a valid test" (p. 2). This is important in order to match activities, lessons and syllabi to the actual content and format of the test itself. In developing the curriculum (TLWOEC), I've created a channel between lesson activities and test, such that the content and format in each case reflect one another. This match, between activities in lessons and on tests, is an attempt to achieve an acceptable level of validity.

Regarding content validity, Brown (2005) points out that "...a test can only be considered reliable and valid for a particular context (or for contexts that are very similar) and purpose (p. 226). Further, Brown (2005) states:

...validity is not about the test itself so much as it is about the test when the scores are interpreted for some specific purpose. In fact, it is much more accurate to refer to the validity of the scores and interpretations that result from a test than to think of the test itself as being valid. (p. 221).

Following this line of reasoning, I will assess the two objectively scored portions of the test. Using Wesche's (1983) four components, the first subtest—a criterion-referenced direct

test of vocabulary knowledge—can be described as follows (Appendix A). The students are given the following directions. Match the correct words to the correct definitions in the following manner:

1 is \_\_\_\_\_, 2 is \_\_\_\_\_, 3 is \_\_\_\_\_, 4 is \_\_\_\_\_, 5 is \_\_\_\_\_ .

You may match the words to their definitions in any order. You have one minute to complete each set. (In this subtest, there are seven stems, five keys and two distractors per set. There are ten sets.)

A stimulus activator in the form of a cartoon picture and audio/spoken vocabulary items is presented first. On the following slide, seven word choices are given together with five blanks for matching. The learner must recognize and comprehend the meaning of words used and studied in past lessons and therefore find the corresponding definitions. Learners must match the correct words to the correct definitions and produce the answers orally. The learner must answer correctly as determined by the test creator. The answers are either right or wrong and can be found on the corresponding key.

If we assess the results of this subtest according to item facility (IF) and item discriminability (ID) we can refer to the "...validity of the scores and interpretations that result from a test..." as Brown (2005, p. 221) mentions.

Item Facility simply "...represents the portion of the people who got the item right..." (Bailey, 1998, p. 132). According to Oller (1979, p. 247) "items falling somewhere between about 0.15 and 0.85 are usually preferred" (Bailey, 1998, p. 133). With just six examinees, the IF for my results showed a mean value of 0.80 with 20 of the 35 items falling into the preferred range (Table 6). Since the number of testees is low, we could consider this a good result. The

ID for this subtest is perhaps disappointing, where 21 of the 35 items had a value of 0.00 (Table 6). On the other hand, Bailey (1998) notes:

If you are designing a criterion-referenced test, there are situations in which ID values of 0.00 would not be problematic. For instance, if all the test-takers missed an item prior to instruction, there would be no variance, so the ID value of 0.00 would indicate a need for instruction. Likewise, if all the test-takers got an item right after instruction, the ID value would be 0.00, but this could indicate their mastery of the item's content (p. 138).

Once again, with the low number of testees, a larger population of students may be needed in order to gather sufficient data from the ID results. In fact, the matching format is elusive to this type of analysis since every option (in this case 6 of 7 options) CAN be interpreted as a possible distractor for each key, with five total keys per set. A multiple-choice format is a better candidate for this type of analysis.

Using Wesche's (1983) four components, the second subtest—a criterion-referenced direct test of listening and reading comprehension—can be described as follows (Appendix A). The testees are given the following directions. Listen to the audio dialogue and read along with it. Select the best answer from the choices given (A, B, C or D). Read BOTH the letter AND the written answer together. For example: "C) A probably likes Art class". There will be one or two questions for each dialogue. You may listen to the dialogue twice. You have one minute to complete each set (with one or two items per set).

A stimulus activator in the form of a scripted dialog which is both written on the screen and spoken as an audio recording is presented. There are ten multiple choice questions referring to seven different scripted dialogs (three dialogues have two questions each). The learner must comprehend the meaning of the dialogues as a whole and infer the correct answer from there interpretations. Learners must select the correct answers and produce them orally as they are written. They should read BOTH the letter and the written answer together. The learner must answer correctly as determined by the test creator. The answers are either right or wrong and can be found on the corresponding key.

Bailey (1998) notes "...that distractors that don't distract anyone serve no purpose in the norm-referenced approach to assessment" (p. 134). The question is then whether or not such results are acceptable in criterion-referenced tests. The results on this subtest are questionable. I'm sure the numbers are too low (only six test-takers) to get an accurate idea of which questions should be changed (Table 7). The mean IF at 0.87 is out of the preferred range of 0.15 to 0.85 (Oller, 1979, p. 247) so some serious consideration needs to be taken in regards to these ten questions. Also 6 of the 10 questions had a 0.00 ID. This part of the pilot could be redone in order to secure better results. This format may indeed be forfeited in future.

Bailey (2005) states: "Intrarater reliability is determined by having the same person evaluate the same data (usually writing samples or recordings of student speech) on two different occasions and comparing the results to see how similar they are" (p. 178). In fact, this is exactly what I employed on the third subtest, a direct analytically scored test of speaking (Appendix A). First, using Wesche's (1983) four components, this subjective test can be described as follows. The examinees are given the following directions. There are seven

questions. Listen to the teacher. Read question one and answer the question to the best of your ability. In the same manner, answer each of the other six questions. Try to answer fully and try to use the language to communicate with the teacher. Take care to NOT make your answers too long or too short. You have five minutes to complete the seven questions.

The stimulus material consists of seven open-ended questions (defined in the curriculum as questions eliciting original student responses with no prior teacher knowledge of said responses). These questions are embedded in the module three activities. The task for the learner is to comprehend the meaning of the questions posed and communicate face-to-face with the teacher. The learners must respond to the questions to the best of their ability and with the intention of communicating meaning. The scoring criteria were developed according to my own experience with learners of this age and level and according to the nature of the curriculum itself (online medium):

- 1) The discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).
- 2) The lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).
- 3) The sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).
- 4) The relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).
- 5) The student's degree of comprehension of the given questions (0 ~ 5 points).

Before I discuss the intrarater reliability of this subtest, I'd like to address the issue of subjective scoring and how the above scale came into being. I first considered using a holistic

approach (Appendix B) but decided that the criteria were too vague to accurately detail the elements I considered essential for discourse. Here is the original holistic scale:

**Level 4 ~ 5:** The speaker uses the question as an opportunity to use the language. Lexical and syntactic elements are acceptable. The student is neither long-winded nor too brief. Pronunciation is intelligible. Student comprehension is perfect (range: 29 ~ 35 points).

**Level 3 ~ 4:** The speaker uses the question as an opportunity to use the language but is hesitant and not fully confident. Lexical and syntactic elements are acceptable, but there is a noticeable error. The student is a bit long-winded or a bit brief. Pronunciation is intelligible but with potential to stray toward non-native speaker error. Student comprehension is very good (range: 22 ~ 28 points).

**Level 2 ~ 3:** The speaker regards the question as an exercise more so than as a tool for communication. Lexical and syntactic elements are acceptable but there are a few errors. The student is somewhat long-winded or somewhat brief. Pronunciation is less than intelligible. Student comprehension is fair to good (range: 15 ~ 21 possible points).

**Level 1 ~ 2:** The speaker regards the question as an imposition. There are many lexical and syntactic errors. The student is either long-winded or too brief. Pronunciation is relatively unintelligible. Student comprehension is poor (range: 8 ~ 14 possible points).

**Level 0 ~ 1:** The speaker doesn't use the question to communicate or the speaker is silent (range: 0 ~ 7 possible points).

Perhaps with higher level students, where we might carry out a conversational activity or a dynamic assessment (topics out of the range of this paper) the holistic scale may be appropriated. Simply stated, the more teacher-learner interaction we have, the more data we have to work with and, in my view, the more value we can place on a holistically scored speaking test. As it worked out on this subtest, the analytic scoring system proved a good match with the younger and lower level learners. It was easier to specify their weaknesses in their discourse.

Let's move on to reliability. The students were recorded answering and communicating using the seven questions mentioned (Appendix C). I rated the learners' recordings twice, each time in a different order, with a three-day interval between the two ratings. I performed *Cronbach's alpha* (Bailey, 1998, pp. 178—182) and came out with a 0.91 reliability rating (Table 8). This shows a favorable result, but once again, since the number of test-takers is quite low, the statistics cannot claim total credibility. The results then are mixed in nature: strong data but too few individual scores.

Finally, I'd like to discuss the piloted test in regards to Swain's (2004) four principles of communicative language testing. 'Starting from somewhere' is a concept that involves more than a single aspect of language. Swain first proposes: "...it saves reinventing the wheel" (2004, p. 188). This more superficial interpretation relates directly to the practicality issue in my case.

Since I'm developing a curriculum, a lot of time and effort must be placed on designing, building and trialing activities that make up the lessons and syllabi. If I had to reinvent tests and subtests for each module and sub-module, I would likely fail to complete the curriculum before my dying days. Just to add a further note on practicality, the online nature of the test (as a flash file) makes it quite easy to administer.

Regarding the four competences—grammatical, sociocultural, discourse and strategic—I can claim a portion of each of these on the speaking subtest. The examinee must be able to produce properly constructed utterances. These utterances must be of appropriate length according to basic social norms (as established subjectively by the examiner). Further, the discourse must be language in use. Also, the examinee must communicate meaning and must strategically make use of the five-minute time limit.

According to Swain (2004) the content of the test should be "motivating, substantive, integrated and interactive for the testee" (p. 190). The stimulus activators (cartoon pictures and audio) are meant to motivate the learners to engage with the material. The teacher-learner-online/onscreen test structure is also motivating simply due to triadic interaction: "The students are not speaking face-to-face bridging some sort of information gap, but they are working side by side, with a joint focus of activity..." (van Lier, 2002, pp. 147,148).

The substantive issue, "... new information ensures that "real" communication can occur (Swain, 2004, p. 192) is realized in the second and third subtests. All the multiple-choice questions are novel and the format of the speaking test is a first time experience. Less so with the first subtest, though the distractors in the matching activity are all first time items.

Swain (2004) defines integrated content as "...one theme around which *all* information and activities are centered" (p. 193). The thread that runs through the test is the use of the simple present tense. In order to improve upon the test from a communicative point of view, the entire curriculum would have to be redone from a theme-related point of view. The advantages and disadvantages of this approach cannot be dealt with in this paper, although this is an interesting topic in curriculum development. On the final aspect of content, the test exhibits complete interaction as was mentioned earlier in regards to triadic interaction. The teacher-learner-screen interface as a joint focus on the online activity assures this possibility (van Lier, 2002, pp. 147, 148).

Bias for the best is described by Swain (2004) as doing "... everything possible to elicit the learners' best performance. Indeed, I used the higher of the two speaking subtest scores (taken from the intrarater reliability data) in each case in order to give the student the benefit of my doubt. Also, the examiner can be kind and considerate whilst administering the test to further bias for the best, and this is exactly what I did

Finally, there is washback or, according to Swain (2004), "... the effect a test has on teaching practices" (p. 196). Here, I'm concerned with positive washback. Since the examiner/teacher participates with the student in completing the test, there is ample opportunity to record specific problems the learner is having with vocabulary as well as listening and reading comprehension, the reading skill itself and speaking-related issues that can be reviewed by way of the audio recordings. I found the test very helpful in assessing the progress and problems of each of my students.

I will end my paper on some suggestions for improving the test. The obvious point that comes to mind is that there is no assessment of writing. While TLWOEC does not intend to address writing of any kind, it would be necessary in future to do so in order to include all the macro-skills. In relation to Swain's (2004) promotion of substantive content, the test could present novel stimulus activators as the same activators are used in both lesson activities and subtests. On the other hand, this would require endless hours of work and raises the specter of practicality. The multiple-choice questions could also be improved upon using Mehren's (1978) criteria for proper creation of such questions. On the other hand, they could be replaced with a different construct.

### **Conclusion**

Overall, I'm quite satisfied with how the test turned out on the whole. The task was daunting, but the payoff in terms of learning and being motivated to carry out a pilot of this nature is truly recognizable. I thank you Professor Bailey for giving me this opportunity...

Tim. October 4, 2009.

## Appendix A: Test Specifications

### *The Language Works Module Three Test for EFL/ESL Young Learners*

#### **Description of Learners**

The learners in this pilot study range in age from 10 to 15 years. They are Korean elementary and middle school students in an ESL (Korean) environment. They were all at a similar level and studying in module three of The Language Works Online English Curriculum ([TLWOEC](#) Ctrl + click to follow link) at the time of testing. I tutor them privately and administered the test to each of them during their regular class time.

#### **General Statement of Purpose**

This criterion-referenced progress test measures the students' knowledge of vocabulary, their listening and reading comprehension skills and their general speaking skills as outlined in the analytic scoring criteria. The content of the test relates directly to the material they have been presented with in their previous lessons. The purpose is to provide an analysis of student progress as they engage in TLWOEC. A further purpose is to ensure the piloted test actually sets out to assess what the students have studied and learned in their lessons: "If a test actually measures what it is intended to measure, we say it is a valid test" (Bailey, 1998, p. 2). The students themselves study English in this program in order to improve their oral communication skills; that is, the focus is speaking and listening.

## **The Test Battery**

There are a total of three subtests: time allowed 25 minutes

- I. Vocabulary Matching (With Stimulus Activator) (10 minutes)
- II. Listening and Reading Comprehension (Multiple-Choice) (10 minutes)
- III. Speaking/Oral Production (Short Answer) (5 minutes)

Each subtest has a separate score. The combination of the three subtests gives the student a score out of a possible 100 points (x/100). The specifications for subtests one and two include objectively scored criteria while subtest three includes an original set of criteria for subjective scoring.

### **I. Vocabulary Matching**

A criterion-referenced, objectively scored vocabulary matching exercise is the first subtest. It is an indirect, discrete-point subtest with seven sets of five matching items worth one point each and attempts to assess the vocabulary knowledge of the students participating in The Language Works Online English Curriculum (<http://eslenglishclassroom.com>). The students have had prior exposure to the vocabulary being tested and there is a stimulus activator in the form of a cartoon picture and an audio reading of some of the vocabulary matching choices. The students are given a maximum of one minute per set (seven sets) and the entire subtest shall not exceed 10 minutes (including the use of the stimulus activator).

The test itself has immediate washback: listening (audio aspect of the stimulus activator), reading (the options must be read aloud) and speaking (the answers must be spoken).

None of these macro-skills are assessed here. The test itself is a flash file included in the curriculum and can be found here: <http://www.lang-works.com/TLW/T3-1/shell.swf> (copy and paste to a browser to open the link). It can also be accessed in the reference section at the end of the paper.

The explanation using Wesche's framework is as follows:

Directions for Learners: In the first matching activity, match the correct words to the correct definitions in the following manner:

1 is \_\_\_\_\_, 2 is \_\_\_\_\_, 3 is \_\_\_\_\_, 4 is \_\_\_\_\_, 5 is \_\_\_\_\_ .

You may match the words to their definitions in any order. You have one minute to complete each set: seven stems, five keys and two distractors per set. There are ten sets.

Stimulus Material: A stimulus activator in the form of a cartoon picture and audio/spoken vocabulary items is presented first. On the following slide, seven word choices are given together with five blanks for matching.

Task Posed: The learner must recognize and comprehend the meaning of words used and studied in past lessons and therefore find the corresponding definitions.

Learner's Response: Learners must match the correct words to the correct definitions and produce the answers orally.

Scoring Criteria: The learner must answer correctly as determined by the test creator. The answers are either right or wrong and can be found on the corresponding key.

## **II. Listening and Reading Comprehension**

The second subtest is a criterion-referenced direct test of reading and listening comprehension. These multiple-choice questions refer to a series of seven scripted dialogues

(three of the seven sets will have two questions each). An 'A/B/A/B' format (four lines of dialogue) will be available as stimulus material in both written and spoken (audio recording) form along with the multiple-choice questions.

To clarify: the dialogue has an audio element (my own voice recorded onto a flash file format) The students will also read along and then must select the best answer from four possible choices according to the content of the dialogue.

The explanation using Wesche's framework is as follows:

Directions for Learners: Listen to the audio dialogue and read along with it. Select the best answer from the choices given (A, B, C or D). Read BOTH the letter AND the written answer together. For example: "C) A probably likes Art class". There will be one or two questions for each dialogue. You may listen to the dialog twice. You have one minute to complete each set (with one or two items per set).

Stimulus Material: A stimulus activator in the form of a scripted dialog which is both written on the screen and spoken as an audio recording is presented. There are ten multiple choice questions referring to seven different scripted dialogs (three dialogues have two questions each).

Task Posed: The learner must comprehend the meaning of the dialogues as a whole and infer the correct answer from there interpretations

Learner's Response: Learners must select the correct answers and produce them orally as they are written. They should read BOTH the letter and the written answer together.

Scoring Criteria: The learner must answer correctly as determined by the test creator. The answers are either right or wrong and can be found on the corresponding key.

### III. Oral Production

The third and final subtest is a Criterion-referenced direct test of oral production/speaking. Simple questions are used as stimulus material to elicit an original student response. These will be 'open-ended' questions (e.g., "Who do you have for breakfast with?") as defined in the TLW Curriculum: "where only the student knows or can create the answer". A 'closed' question, by contrast, is defined: "the answer can be known by all" (Brockley, 2009, Curriculum Overview).

This is the subjective scoring aspect of the test. It's based on an analytic scoring system which was chosen over a holistic scoring scale: as this is a test of beginners, I've concluded that there is no need to assess a general proficiency level; rather, the elements considered important in communicating meaning will be analyzed. This kind of assessment will also allow for more specific forms of feedback and potential positive washback in the process of test review. There are 35 total points possible and five criteria. The rater will assess:

- 1) the discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).
- 2) the lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).
- 3) the sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).
- 4) the relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).
- 5) student comprehension of the given questions (0 ~ 5 points).

Variability in perception (mostly due to dialect and socio-cultural differences) will be inevitable and an attempt at norming the criteria just mentioned would be necessary if this scale were to be used with multiple raters.

The explanation using Wesche's framework is as follows:

Directions for Learners: There are seven questions. Listen to the teacher read question one and answer the question to the best of your ability. In the same manner, answer each of the other six questions. Try to answer fully and try to use the language to communicate with the teacher. Take care to NOT make your answers too long or too short. You have five minutes to complete the seven questions.

Stimulus Material: Seven open-ended (described above) questions embedded in the curriculum.

Task Posed: The learner must comprehend the meaning of the questions posed and communicate face to face with the teacher

Learner's Response: Learners must answer the questions to the best of their ability and with the intention of communicating meaning

Scoring Criteria: The learner must answer correctly as determined by the test creator. The answers are either right or wrong

## Appendix B

*Revisions made after pre-piloting the test***Revision One: Multiple Choice (Section Two) Question One:**

Dialog One: Choose the best answer for each question

Jim) Do you like ice cream?

Alex) Of course... It's my favorite food.

Jim) Really... What flavor do you like best?

Alex) Well... hard question... maybe blueberry.

| Original Question ( <b>bold</b> = correct answer)  | Revised Question   |
|--|--|
| <u>MC1</u> : Alex likes blueberry ice cream.<br>A) Yes<br>B) No<br><b>C) Probably</b><br>D) Not sure | <u>MC1</u> : Alex's favorite ice cream is blueberry.<br>A) Yes<br>B) No<br><b>C) Probably</b><br>D) Not sure |

**Revision Two: Short Questions for Speaking Subtest (Section Three) Scoring Criteria:***Original Holistic Scoring Criteria:*

**Level 4 ~ 5:** The speaker uses the question as an opportunity to use the language. Lexical and syntactic elements are acceptable. The student is neither long-winded nor too brief. Pronunciation is intelligible. Student comprehension is perfect (range: 29 ~ 35 points).

**Level 3 ~ 4:** The speaker uses the question as an opportunity to use the language but is hesitant and not fully confident. Lexical and syntactic elements are acceptable, but there is a noticeable error. The student is a bit long-winded or a bit brief. Pronunciation is intelligible but with

potential to stray toward non-native speaker error. Student comprehension is very good (range: 22 ~ 28 points).

**Level 2 ~ 3:** The speaker regards the question as an exercise more so than as a tool for communication. Lexical and syntactic elements are acceptable but there are a few errors. The student is somewhat long-winded or somewhat brief. Pronunciation is less than intelligible. Student comprehension is fair to good (range: 15 ~ 21 possible points).

**Level 1 ~ 2:** The speaker regards the question as an imposition. There are many lexical and syntactic errors. The student is either long-winded or too brief. Pronunciation is relatively unintelligible. Student comprehension is poor (range: 8 ~ 14 possible points).

**Level 0 ~ 1:** The speaker doesn't use the question to communicate or the speaker is silent (range: 0 ~ 7 possible points).

*Revised Analytic Scoring Criteria:*

1) The discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).

2) The lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).

3) The sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).

4) The relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).

5) The student's degree of comprehension of the given questions (0 ~ 5 points).

### Appendix C

*Actual Piloted Test* (Ctrl + click to link: <http://www.lang-works.com/TLW/T3-1/shell.swf> )

**Subtest One: Vocabulary Matching** (35 points/ 0 ~ 5 points per set/ 7 sets)

Directions for Learners: In the first matching activity, speak the correct words for the correct definitions in the following manner:

"1 is \_\_\_\_\_, 2 is \_\_\_\_\_, 3 is \_\_\_\_\_, 4 is \_\_\_\_\_, 5 is \_\_\_\_\_."

You may match (by speaking) the words in any order. Choose only five of the seven words per set. You have one minute to complete each set.

#### Vocabulary Matching One:

|               |                  |                  |
|---------------|------------------|------------------|
|               | <b>Lunch</b>     | <b>Delicious</b> |
| <b>Flavor</b> | <b>Ice Cream</b> | <b>Snack</b>     |
|               | <b>Freezer</b>   | <b>Candy Bar</b> |

- \_\_\_ 1) a cold, sweet dessert made with cream
- \_\_\_ 2) pleasing to taste, tasty
- \_\_\_ 3) a very cold place for storing food
- \_\_\_ 4) a special kind or type of taste

\_\_\_ 5) food eaten between meals

**Vocabulary Matching Two:**

|                  |                |
|------------------|----------------|
| <b>Breakfast</b> | <b>Bull</b>    |
| <b>Cream</b>     | <b>Farm</b>    |
|                  | <b>Low Fat</b> |
| <b>Cow</b>       | <b>Calf</b>    |

- \_\_\_ 1) a place where animals are kept and raised  
 \_\_\_ 2) a meal served in the morning  
 \_\_\_ 3) a female animal that eats grass  
 \_\_\_ 4) a very rich form of milk  
 \_\_\_ 5) a male animal with horns

**Vocabulary Matching Three:**

|                 |               |
|-----------------|---------------|
| <b>Barbecue</b> | <b>Cook</b>   |
| <b>Pancake</b>  | <b>Many</b>   |
|                 | <b>Bake</b>   |
| <b>Practice</b> | <b>Hungry</b> |

- \_\_\_ 1) use heat to make food on a stove  
 \_\_\_ 2) use heat to make food in an oven  
 \_\_\_ 3) a round, sweet, flat cooked food  
 \_\_\_ 4) food cooked on a grill over fire  
 \_\_\_ 5) doing something many times to improve skill

**Vocabulary Matching Four:**

|                 |                |
|-----------------|----------------|
| <b>Teeth</b>    | <b>Dentist</b> |
| <b>Floss</b>    | <b>Drill</b>   |
|                 | <b>Clean</b>   |
| <b>Check-up</b> | <b>Sink</b>    |

- \_\_\_ 1) a person who checks people's teeth  
 \_\_\_ 2) what we use to chew food  
 \_\_\_ 3) to make shiny and new  
 \_\_\_ 4) to clean between the teeth  
 \_\_\_ 5) an appointment to help your health

**Vocabulary Matching Five:**

**Chase**                      **Afraid**  
**Restroom**                  **Tease**                      **Bully**  
**Race**                          **Angry**

- \_\_\_\_ 1) when you feel mad or upset  
 \_\_\_\_ 2) when you want to run or hide  
 \_\_\_\_ 3) when someone calls you names and laughs  
 \_\_\_\_ 4) a bathroom without a bath or shower  
 \_\_\_\_ 5) to run after someone or something

**Vocabulary Matching Six:**

**Watch**                      **Hide**  
**Whistle**                  **Hit**                          **Wait**  
**Walk**                          **Teacher**

- \_\_\_\_ 1) to pause and delay an action or event  
 \_\_\_\_ 2) to be invisible to everyone  
 \_\_\_\_ 3) to punch or strike someone/something  
 \_\_\_\_ 4) to fix your eyes on an action or event  
 \_\_\_\_ 5) to blow air and make sound

**Vocabulary Matching Seven:**

**Laugh**                      **Rest**  
**Pet**                          **Dream**                      **Best**  
**Smile**                      **Think**

- \_\_\_\_ 1) relief from work, a feeling of quiet  
 \_\_\_\_ 2) to make a sound when something is funny  
 \_\_\_\_ 3) to see images when you are sleeping  
 \_\_\_\_ 4) an animal kept at home as a companion  
 \_\_\_\_ 5) to raise the corners of your mouth

**Subtest Two: Multiple Choice** (30 points/ 3 points each/ 10 sets)

Directions for Learners: Listen to the audio dialogue and read along with it. Select the best answer from the choices given (A, B, C or D). Read BOTH the letter AND the written answer together. For example: "C... **A probably likes Art class**". There will be one or two questions for each dialogue. You may listen to the dialog twice. You have one minute to complete each set (with one or two items per set).

**Dialog One: Choose the best answer for each question**

**Jim) Do you like ice cream?**

**Alex) Of course... It's my favorite food.**

**Jim) Really... What flavor do you like best?**

**Alex) Well... hard question... maybe blueberry.**

**MC1: Alex's favorite ice cream is blueberry.**

- A) Yes**
- B) No**
- C) Probably**
- D) Not sure**

**MC2: What is "flavor" most similar to?**

- A) Touch**
- B) Taste**
- C) Sound**
- D) Sight**

**Dialog Two: Choose the best answer for each question**

**Jan) Do you drink milk for breakfast?**

**Sam) No, I usually drink orange juice.**

**Jan) Orange juice gives me a stomachache.**

**Sam) Well... O.J. has lots of vitamin C.**

**MC3: Who usually drinks orange juice for breakfast?**

- A) Jan**
- B) Sam**
- C) Both Jan and Sam**

D) Neither Jan nor Sam

**Dialog Three:** Choose the best answer for each question

Eve) Why does that girl chase you?

Bill) Well... I like to tease her!

Eve) You shouldn't do that. She gets angry...

Bill) I know. That's why I like to do it!

**MC4:** What does Eve tell Bill to do?  
She tells him \_\_\_\_\_ the girl.

- A) to tease
- B) not to tease
- C) to chase
- D) not to chase

**MC5:** Why does the girl get angry?  
Because Bill \_\_\_\_\_ her.

- A) chases
- B) likes
- C) knows
- D) teases

**Dialog Four:** Choose the best answer for each question

Jill) What kind of food can you cook?

Ben) Nothing really... I can't cook at all.

Jill) Really... I can cook Chinese food...

Ben) Maybe if I practice, I can do it...

**MC6:** What kind of food can Ben cook?

- A) Chinese
- B) Many kinds
- C) A few kinds
- D) Nothing

**Dialog Five:** Choose the best answer for each question

Ed) How often do you go to the dentist?

Liz) Oh... a few times a year... Why?

Ed) Well... I'm going tomorrow...

Liz) That's terrible. I feel sorry for you!

**MC7:** What does the dentist do? He \_\_\_\_\_ .

- A) cleans people's toes
- B) feels sorry for us
- C) goes a few times a year
- D) checks our teeth

**Dialog Six:** Choose the best answer for each question

Jack) Who are you waiting for?

Lia) I'm not waiting. I'm just taking a rest.

Jack) Oh, I see. Can I join you?

Lia) Sure... What's your name?

**MC8:** What is Lia doing? She is \_\_\_\_\_ .

- A) taking a rest
- B) waiting for a friend
- C) joining a friend
- D) sitting on a bench

**Dialog Seven:** Choose the best answer for each question

Pam) Why are you smiling?

Ted) Well... I got an 'A' on my English test.

Pam) Unbelievable!... That test was scary.

Ted) I'm kind of in shock... I guess I was lucky.

**MC9:** What is unbelievable? | **MC10:** Why is Ted in shock? Because he \_\_\_\_\_ .

- A) Ted is in shock.  
 B) Pam is scary.  
 C) Ted got an 'A'.  
 D) Ted is lucky.

- A) is afraid of Pam  
 B) is a lucky person  
 C) has an 'A' on the test  
 D) is unbelievable

**Subtest Three: Short Answer Speaking** (35 points/ analytic scoring/ 7 questions)

Directions for Learners: There are seven questions. Listen to the teacher read question one and answer the question to the best of your ability. In the same manner, answer each of the other six questions. Try to answer fully and try to use the language to communicate with the teacher. Take care to NOT make your answers too long or too short. You have up to five minutes to complete the seven questions.

**Short Questions:**

- Q1) When do you eat ice cream?  
 Q2) Who do you have breakfast with?  
 Q3) Where do people barbecue?  
 Q4) What does the dentist do?  
 Q5) What do you and your friends talk about?  
 Q6) Who do you sometimes wait for?  
 Q7) What do you do in English class?

Appendix D

*Checklist for writing multiple-choice items:*

Table 10-1 CHECKLIST FOR WRITING MULTIPLE-CHOICE ITEMS

| Factor | Yes | No |
|--------|-----|----|
|--------|-----|----|

1. Has the item been clearly presented? Is the main problem in the stem? Has excess verbiage been eliminated? YES
2. Has the item been cast so that there is no repetition of key words or phrases for each option? YES
3. Do the options come at the end of the stem? YES (I see MC 4 and 5 as completions)
4. Have the responses been arranged in some systematic fashion, such as alphabetic or length of response? YES
5. Are all distractors plausible? Are the number of distractors related to the examinees' age level? To the subject matter? To the time available for testing? YES
6. Have all irrelevant clues been avoided (grammatical, rote verbal association, length of correct answer, etc.)? YES
7. Are the correct answers randomly assigned throughout the test with approximately equal frequency? NO (My theory is that there is no need given the format of the subtest: spoken NOT written. There is no way for the testee to track this.)
8. Has an "I don't know" option been considered? YES
9. Is there only one correct (or best) answer? YES
10. Has "all the above" been avoided? YES
11. Has the "none of the these" option been used sparingly? Only when appropriate? YES
12. Have overlapping options been avoided? YES
13. Have negative statements been avoided? If used, has the negative been underlined or written in capital letters? NO (See M4... the two negatives are in the front of the options... in this case I didn't deem it necessary... What do you think?)
14. Have the items been reviewed independently? By you? YES (Myself and two different pre-piloters)

From Mehrens, W.A. & Lehman, I.J. (1978). *Measurement and evaluation in education and psychology* (2<sup>nd</sup> edition). Holt, Rinehart and Winston. (From week 5 EDU570)

## Appendix E

*Key for scoring subtest one: matching*

### **M1 Vocabulary Answers:**

- 1) Ice Cream: a cold, sweet dessert made with cream
- 2) Delicious: pleasing to taste, tasty
- 3) Freezer: a very cold place for storing food
- 4) Flavor: a special kind or type of taste
- 5) Snack: food eaten between meals

### **M2 Vocabulary Answers:**

- 1) Farm: a place where animals are kept and raised
- 2) Breakfast: a meal served in the morning
- 3) Cow: a female animal that eats grass
- 4) Cream: a very rich form of milk
- 5) Bull: a male animal with horns

### **M3 Vocabulary Answers:**

- 1) Cook: use heat to make food on a stove
- 2) Bake: use heat to make food in an oven
- 3) Pancake: a round, sweet, flat cooked food
- 4) Barbecue: food cooked on a grill over fire
- 5) Practice: doing something many times  
to improve skill

### **M4 Vocabulary Answers:**

- 1) Dentist: a person who checks people's teeth
- 2) Teeth: what we use to chew food
- 3) Clean: to make shiny and new
- 4) Floss: to clean between the teeth
- 5) Check-up: an appointment to help your health

### **M5 Vocabulary Answers:**

- 1) Angry: when you feel you mad or upset
- 2) Afraid: when you want to run or hide
- 3) Tease: when someone calls you names and laughs
- 4) Restroom: a bathroom without a bath or shower
- 5) Chase: to run after someone or something

**M6 Vocabulary Answers:**

- 1) Wait: to pause and delay an action or event
- 2) Hide: to be invisible to everyone
- 3) Hit: to punch or strike someone/something
- 4) Watch: to fix your eyes on an action or event
- 5) Whistle: to blow air and make sound

**M7 Vocabulary Answers:**

- 1) Rest: relief from work, a feeling of quiet
- 2) Laugh: to make a sound when something is funny
- 3) Dream: to see images when you are sleeping
- 4) Pet: an animal kept at home as a companion
- 5) Smile: to raise the corners of your mouth

## Appendix F

*Key for scoring subtest two: multiple-choice*

|      |       |
|------|-------|
| 1) C | 2) B  |
| 3) B | 4) B  |
| 5) D | 6) D  |
| 7) D | 8) A  |
| 9) C | 10) C |

## Appendix G

*Key for scoring subtest three: short answer*

- 1) The discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).
- 2) The lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).
- 3) The sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).

4) The relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).

5) The student's degree of comprehension of the given questions (0 ~ 5 points).

## Appendix H

### *Specifications for test writers (Alderson et. al., 1995)*

1. What is the purpose of the test? (How will the information you gather be used? Are you measuring achievement or progress? Are you placing students in a program?)

=This test is part of a series of ongoing assessments... there is no definite 'end' to the program. In this paper, the test pilots students at a private academy but could be used in any second language learning environment (ESL/EFL settings in either public or private sectors) along with the curriculum from which it was born (see TLW Online English Curriculum in the reference section or Ctrl and click to link: <http://eslenglishclassroom.com/index.html> ). The information gathered will be used to develop further such tests for the curriculum.

2. What sort of learners will be taking the test?

=The target group for this piloted test is Korean young learners in an EFL context (mid-elementary to mid-middle school, roughly ages 9 to 13 ) at beginner to intermediate level on average. The purpose for studying English is mixed (some are forced by parents to do so while others enjoy learning a second language but the root element is that they perceive a vague need learn English for future high stakes tests (such as middle, high school and university entrance exams).

3. What language skills should be tested (reading, writing, speaking and/or listening)?

=reading, speaking and listening (no writing)

4. What language elements should be tested?

=vocabulary, discourse, grammar and pronunciation

5. What target language situation is envisaged for the test, and is this to be simulated in some way in the test content and method?

= This piloted test is one of nine tests planned for the beginner modules in The Language Works online English Curriculum. The curriculum is laid out to promote speaking, reading and listening in a classroom environment. The test reflects the content and method of study in the lessons themselves.

6. What text types should be chosen as stimulus materials -- written and/or spoken?

=written, spoken and aural (matching: reading, speaking and listening in #3).

7. What sort of tasks are required -- discrete point, integrative, simulated 'authentic', objectively assessable? (That is, what will the test-takers actually do?)

=It will be integrative, the matching subtest requires listening in the stimulus activator and reading and speaking in the answer sets. The multiple-choice follows the same format. Both of these subtests are objectively scored. The third subtest, employs a question-style format (e.g. "Who do you have breakfast with?"). It will require listening comprehension (the teacher elicits answers orally) and will require reading, listening and speaking (the teacher will read the question but it is also available for the student to see and read silently). There is a discourse element as the teacher and student may elaborate on the given questions, for example:

T: Who do you have breakfast with?

S: I have breakfast with my sister.

T: Really... Not with your mom and dad.

S: No... my dad go to work... My mom cooks...

The speaking test will be graded subjectively using an analytic scoring system.

8. What test methods (what item formats) are to be used? (One multiple-choice subtest is required.)

=subtests: 1) matching 2) multiple choice 3) subjectively scored responses to short questions.

9. How many sections should the test have, how long should they be and how will they be differentiated? (There will be at least three sections – more if you are working with another student.)

=three sections: 35 points (ten minutes), 30 points, (ten minutes) 35 points (five minutes) with a total of 100 points (a 25-minute test). They will be differentiated by sections on a flash file format which is part of a greater online curriculum (see the test here, Ctrl and click to link: <http://www.lang-works.com/TLW/T3-1/shell.swf> ).

10. How many items are required for each section? What is the relative weight for each item?

=three sections: 30 items (1 point each) 10 items (3 points each) 7 items (analytic scoring with 35 possible points)

11. What rubrics are to be used as instructions for candidates? (That is, what instructions and guidance are printed in the test and/or announced by the test administrator?)

=students will be familiar with the format since these same exercises/activities appear in the lessons themselves... but guidance will be available as the test is administered orally by the teacher and the directions are clearly presented in the flash file test as well.

12. Which criteria will be used for assessment by markers? (In other words, describe how the answer key will be developed for the objectively scored portion, and explain the rating system for the subjectively scored portion.)

=The answers for the objectively scored portions of the test are included in the test itself. There is a scoring instrument that is available to the teacher as a download (word doc format):

<http://eslenglishclassroom.com/Assessment-Instrument-Modular.doc> (Ctrl and click to link).

The teacher fills in the results as the test takes place. The speaking portion (section three) must be recorded and scored at a later time using the prescribed original analytic criteria as follows:

- 1) The discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).
- 2) The lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).
- 3) The sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).
- 4) The relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).
- 5) The student's degree of comprehension of the given questions (0 ~ 5 points).

=Intra Rater Reliability: two different attempts at scoring the speaking test will be carried out on two occasions. The interval between scoring sessions will not be less than two days.

=Correlation: All three tests will employ interval data. Pearson's correlation coefficient will be employed in the manner of comparing sections 1 and 2, sections 1 and 3 and sections 2 and 3

From: Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. (pp. 9-39). Cambridge: Cambridge University Press.

## Appendix I

### *Construct Chart for use with TLW Online Tests*

| <b>Subtest</b> | <b>Definitions of Construct(s) Assessed (including citations)</b> | <b>Test Method (Item Types)</b> | <b>Points Possible</b>          |
|----------------|---|---------------------------------|---------------------------------|
|                | Criterion-referenced, objectively scored vocabulary matching      | Matching                        | 35 points<br>(1 point per item) |

| Subtest 1      | <p>exercise.</p> <p>It is an indirect, discrete-point subtest with seven sets of five matching items worth one point each.</p> <p>The test measures recall and comprehension of these items that are part of the lesson content in the curriculum.</p> <p>They will be related to a cartoon which will be available as stimulus material (as was the case in lesson exercises) along with the matching vocabulary-style questions.</p>          |                                 | with 35 items)                                 |
|----------------|---|---------------------------------|--|
| <b>Subtest</b> | <b>Definitions of Construct(s) Assessed (including citations)</b>   | <b>Test Method (Item Types)</b> | <b>Points Possible</b>                         |
| Subtest 2      | <p>Criterion-referenced direct test of reading and listening comprehension</p> <p>These multiple-choice questions refer to a series of seven scripted dialogues (three of the seven sets will have two questions each). An 'A/B/A/B' format (four lines of dialogue) will be available as stimulus material in both written and spoken (audio recording) form along with the multiple-choice questions.</p> <p>To clarify: the dialogue has</p> | Multiple Choice                 | 30 points<br><br>(3 points each with 10 items) |

|                |  |                                 |  |
|----------------|--|---------------------------------|--|
|                | <p>an audio element (my own voice recorded onto a flash file format) The students will read along and then must select the best answer from four possible choices according to the content of the dialogue.</p> <p>Mehrens, W.A. &amp; Lehman, I.J. (1978). <i>Measurement and evaluation in education and psychology</i>. Holt, Rinehart and Winston.</p>   |                                 |  |
| <b>Subtest</b> | <b>Definitions of Construct(s) Assessed (including citations)</b>  | <b>Test Method (Item Types)</b> | <b>Points Possible</b>   |
| Subtest 3      | <p>Criterion-referenced direct test of oral production/speaking</p> <p>This is the subjective scoring aspect of the test with an analytic scoring scale: the students are assessed to their performance in relation to five criteria, after the fact. The rater will assess:</p> <p>Simple questions are used as stimulus material to elicit an original student response</p> <p>These will be 'open-style' questions (e.g., "Who do you have breakfast with?") meaning that the answer must be constructed by the student.</p> <p>35 total points possible:</p> | Short Questions                 | <p>35 points</p> <p>Analytic scoring system with five categories and a range of 0 to 35 points</p> |

|  |   |  |  |
|--|---|--|--|
|  | <p>1) The discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).</p> <p>2) The lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).</p> <p>3) The sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).</p> <p>4) The relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).</p> <p>5) The student's degree of comprehension of the given questions (0 ~ 5 points).</p> <p>Variability in perception (mostly due to dialect and socio-cultural differences) will be inevitable and an attempt at norming the criteria just mentioned would be necessary if this scale were to be used with multiple raters.</p> <p>(Bailey, K (1998) Learning About Language Assessment. Heinle and</p> |  |  |
|--|---|--|--|

|  |                     |  |  |
|--|---------------------|--|--|
|  | Heinle Publishers.) |  |  |
|--|---------------------|--|--|

## Appendix J

*Scoring Instrument for use with TLW Online Tests*

|   |         |           |          |                           |         |           |    |    |     |
|---|---------|-----------|----------|---------------------------|---------|-----------|----|----|-----|
| <b>Assessment Instrument Using the Language Works (TLW) Test Files (TF):</b>                                    |         |           |          |                           |         |           |    |    |     |
| Student name/ID: _____ Class: _____ Date: _____   |         |           |          |                           |         |           |    |    |     |
| Section One: Vocabulary Matching  |         |           |          | Total Points Section Two: |         |           |    |    |     |
| _____/35  |         |           |          |                           |         |           |    |    |     |
| Directions: Record the number of correct matches per set, then tally the total in the space above.              |         |           |          |                           |         |           |    |    |     |
| Set One   | Set Two | Set Three | Set Four | Set Five                  | Set Six | Set Seven |    |    |     |
| _____/5   | _____/5 | _____/5   | _____/5  | _____/5                   | _____/5 | _____/5   |    |    |     |
| Section Two: Multiple-Choice for Comprehension  |         |           |          | Total Points Section Two: |         |           |    |    |     |
| _____/30  |         |           |          |                           |         |           |    |    |     |
| Directions: Check the box if the answer is correct (three points each) then tally the total in the space above. |         |           |          |                           |         |           |    |    |     |
| Q1  | Q2      | Q3        | Q4       | Q5                        | Q6      | Q7        | Q8 | Q9 | Q10 |
|   |         |           |          |                           |         |           |    |    |     |

Section Three: Short Questions for speaking. Rate the student using the criteria below. Total Points:

\_\_\_\_\_/35

- 1) The discourse aspect of the answers, that is, the degree to which the learner attempts to communicate meaning in response to questions (0 ~ 12 points).
- 2) The lexical and syntactic qualities of the utterances, that is, the degree to which the learner uses form to communicate meaning (0 ~ 7 points).
- 3) The sufficiency of length of the responses, that is, the degree to which learners' are neither long-winded nor unresponsive in communicating meaning (0 ~ 6 points).
- 4) The relative intelligibility in relation to pronunciation, intonation, stress and rhythm (0 ~ 5 points).
- 5) The student's degree of comprehension of the given questions (0 ~ 5 points).

|                                 |                                 |                                   |                            |
|---------------------------------|---------------------------------|-----------------------------------|----------------------------|
| Total Section One:<br>_____/ 35 | Total Section Two:<br>_____/ 30 | Total Section Three:<br>_____/ 35 | Grand Total:<br>_____/ 100 |
|---------------------------------|---------------------------------|-----------------------------------|----------------------------|

Also available online as a printout (Ctrl + Click to follow link):

<http://eslenglishclassroom.com/Assessment-Instrument-Modular.doc>

Table 1.

Frequency Polygons for TLW Module Three Subtests and Total Scores:

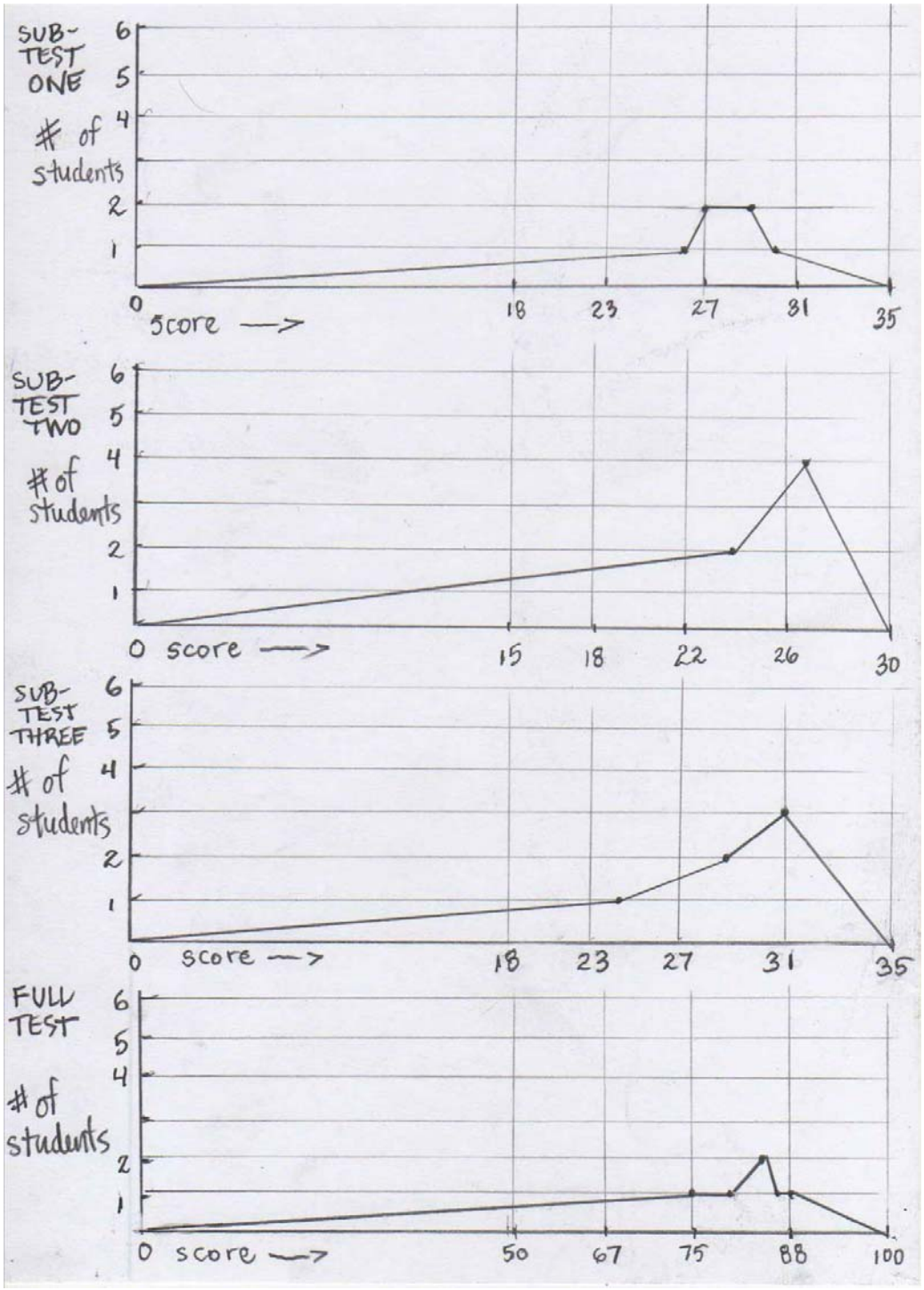


Table 2.

*Pearson's Correlation and Overlapping Variation for the Sub-tests*

|                                       | I.<br>Vocabulary Matching<br>and<br>II. Comprehension<br>Multiple Choice | I.<br>Vocabulary Matching<br>and<br>III. Short Answer<br>Speaking | II. Comprehension<br>Multiple Choice<br>and<br>III. Short Answer<br>Speaking |
|---------------------------------------|--|---|--|
| Correlation - r                       | 0.995  | 0.975   | 0.987  |
| Overlapping variance - r <sup>2</sup> | 0.990  | 0.951   | 0.974  |

Table 3.

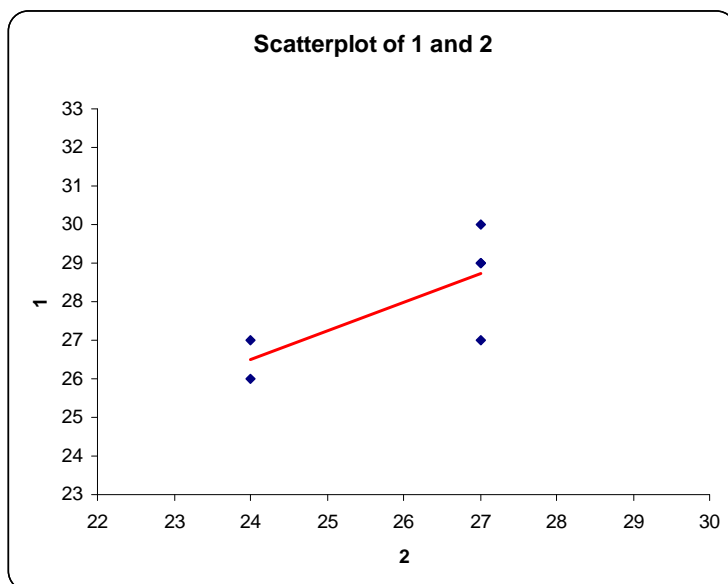
*Trendline and Scatterplot for Subtests One and Two*

Table 4.

*Trendline and Scatterplot for Subtests One and Three*

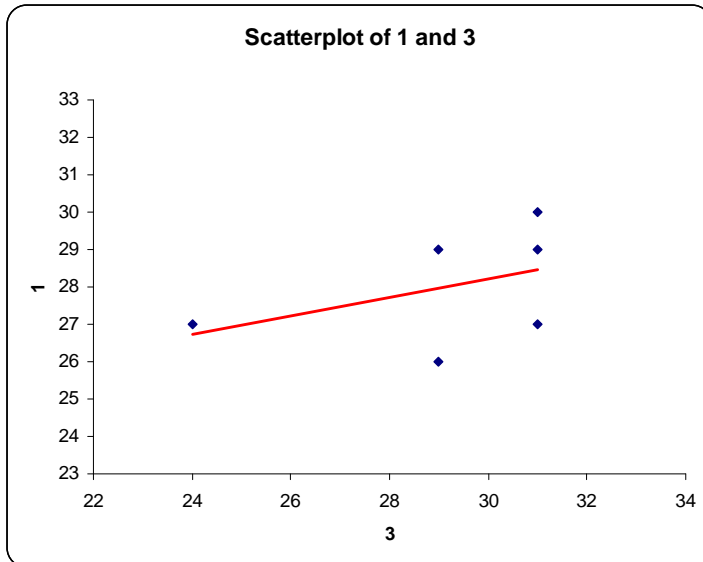


Table 5.

*Trendline and Scatterplot for Subtests Two and Three*

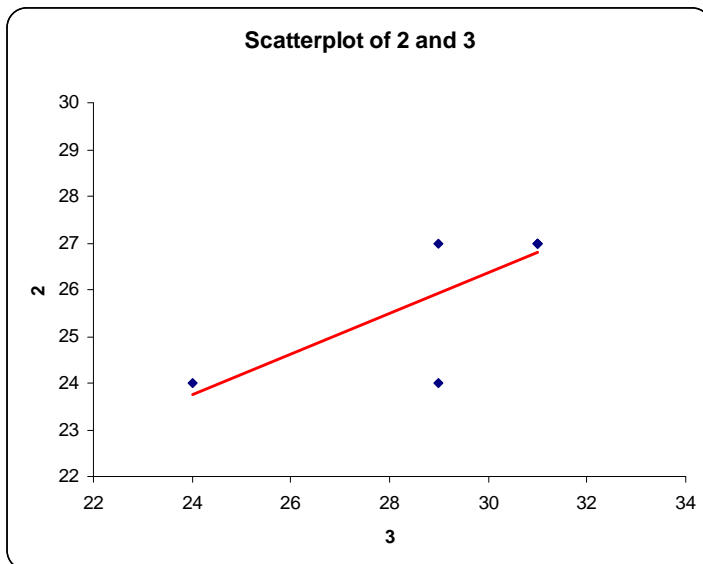


Table 6.

*Item Facility and Discriminability for Subtest One: Matching*

|      | I.F.  | I.D.  |      | I.F.  | I.D.  |
|------|---|---|------|---|---|
| Item | (quantity<br>of same<br>numerical<br>results) | (quantity<br>of same<br>numerical<br>results) | Item | (quantity<br>of same<br>numerical<br>results) | (quantity<br>of same<br>numerical<br>results) |
| 1    | 1   | 0   | 19   | .50   | 1.01(1)                                       |
| 2    | .66   | .51   | 20   | .66   | 0   |
| 3    | .83   | .51   | 21   | 1   | 0   |
| 4    | .50   | 0   | 22   | .33(1)  | -.51(4)                                       |
| 5    | .83   | .51   | 23   | .66   | 0   |
| 6    | .83   | 0   | 24   | 1   | 0   |
| 7    | 1   | 0   | 25   | .66   | -1.01(1)                                      |
| 8    | .83   | -.51  | 26   | .66   | 0   |
| 9    | 1   | 0   | 27   | .50(3)  | .51   |
| 10   | .17(1)  | -.51  | 28   | 1   | 0   |
| 11   | .66   | .51   | 29   | .66   | .51(8)  |
| 12   | .66   | .51   | 30   | 1   | 0   |
| 13   | .83   | .51   | 31   | .66(9)  | 0   |
| 14   | 1   | 0   | 32   | 1   | 0   |
| 15   | 1   | 0   | 33   | 1   | 0   |

|      |        |      |    |       |       |
|------|--------|------|----|-------|-------|
| 16   | 1      | 0    | 34 | 1     | 0     |
| 17   | 1      | 0    | 35 | 1(15) | 0(21) |
| 18   | .83(6) | -.51 |    |       |       |
| Sum  |        |      |    | 27.92 | x     |
| Mean |        |      |    | 0.80  | x     |

Table 7.

*Distractor Analysis, Item Facility and Discriminability for Subtest Two: Multiple Choice*

| Item | A  | B  | C  | D  | I.F.  | I.D. |
|------|----|----|----|----|-------|------|
| 1    | 1  | 0  | 5* | 0  | .83   | .51  |
| 2    | 0  | 6* | 0  | 0  | 1     | 0    |
| 3    | 1  | 5* | 0  | 0  | .83   | .51  |
| 4    | 2  | 3* | 0  | 1  | .50   | -.51 |
| 5    | 0  | 0  | 0  | 6* | 1     | 0    |
| 6    | 0  | 0  | 0  | 6* | 1     | 0    |
| 7    | 0  | 0  | 0  | 6* | 1     | 0    |
| 8    | 6* | 0  | 0  | 0  | 1     | 0    |
| 9    | 0  | 0  | 6* | 0  | 1     | 0    |
| 10   | 0  | 3  | 3* | 0  | .50   | .51  |
| Sum  |    |    |    |    | 8.66  | x    |
| Mean |    |    |    |    | 0.866 | x    |

Table 8.

*Intra-rater Reliability*

| Student | First Rating (9/24/09) | Second Rating (0/27/09) |
|---------|------------------------|-------------------------|
| Pink    | 23                     | 24                      |
| Jenny   | 31                     | 31                      |
| Jerin   | 29                     | 29                      |
| Min     | 30                     | 31                      |
| Suzy    | 29                     | 27                      |
| Rich    | 28                     | 31                      |

Calculate intrarater values as follows:

$$2 \left[ \frac{1 - (\text{rating one variance} + \text{rating two variance})}{\text{variance of combined scores of one and two}} \right]$$

Results:

- 1) rating one variance: 7.869
- 2) rating two variance: 8.168
- 3) combined variance: 29.366
- 4) intrarater reliability: .908 rounded to .91

Table 9.

*Descriptive Statistics for TLW Module Three: Test One* (100 points possible, 3 Sections)

| Student (age) | Score | Mean  | Distance | Distance <sup>2</sup> |
|---------------|-------|-------|----------|-----------------------|
| Pink (10)     | 75    | 83.18 | -8.18    | 66.9124               |
| Jenny (11)    | 85    | 83.18 | 1.82     | 3.3124                |
| Jerin (13)    | 85    | 83.18 | 1.82     | 3.3124                |

|                    |              |       |       |          |
|--------------------|--------------|-------|-------|----------|
| Min (13)           | 88           | 83.18 | 4.82  | 23.2324  |
| Suzy (14)          | 79           | 83.18 | -4.18 | 17.4724  |
| Rich (15)          | 87           | 83.18 | 3.82  | 14.5924  |
| Sum                | 499          |       |       | 128.8344 |
| Median             | 85           |       |       |          |
| Mode               | 85           |       |       |          |
| Range              | 75 ~ 88 (13) |       |       |          |
| df n-1             | 5            |       |       |          |
| Variance           | 25.766       |       |       |          |
| Standard deviation | 5.076        |       |       |          |

Table 10.

*Descriptive Statistics for Subtest One: Matching (35 points possible, 1 point each)*

| Student    | Score | Mean | Distance | Distance <sup>2</sup> |
|------------|-------|------|----------|-----------------------|
| Pink (10)  | 27    | 28   | -1       | 1                     |
| Jenny (11) | 27    | 28   | -1       | 1                     |
| Jerin (13) | 29    | 28   | 1        | 1                     |
| Min (13)   | 30    | 28   | 2        | 4                     |
| Suzy (14)  | 26    | 28   | -2       | 4                     |
| Rich (15)  | 29    | 28   | 1        | 1                     |
| Sum        | 168   |      |          | 12                    |
| Median     | 28    |      |          |                       |

|                    |             |
|--------------------|-------------|
| Mode               | 27          |
| Range              | 26 ~ 30 (4) |
| df n-1             | 5           |
| Variance           | 2.399       |
| Standard deviation | 1.549       |

Table 11.

*Descriptive Statistics for Subtest Two: Vocabulary* (30 points possible, 3 points each)

| Student            | Score       | Mean | Distance | Distance <sup>2</sup> |
|--------------------|-------------|------|----------|-----------------------|
| Pink (10)          | 24          | 26   | -2       | 4                     |
| Jenny (11)         | 27          | 26   | 1        | 1                     |
| Jerin (13)         | 27          | 26   | 1        | 1                     |
| Min (13)           | 27          | 26   | 1        | 1                     |
| Suzy (14)          | 24          | 26   | -2       | 4                     |
| Rich (15)          | 27          | 26   | 1        | 1                     |
| Sum                | 156         |      |          | 12                    |
| Median             | 27          |      |          |                       |
| Mode               | 27          |      |          |                       |
| Range              | 24 ~ 27 (3) |      |          |                       |
| df n-1             | 5           |      |          |                       |
| Variance           | 2.399       |      |          |                       |
| Standard deviation | 1.549       |      |          |                       |

Table 12.

*Descriptive Statistics for Subtest Three: Short Answer (35 points possible, analytic scoring)*

| Student (age)      | Score       | Mean   | Distance | Distance <sup>2</sup> |
|--------------------|-------------|--------|----------|-----------------------|
| Pink (10)          | 24          | 29.167 | -5.167   | 26.698                |
| Jenny (11)         | 31          | 29.167 | 1.833    | 3.360                 |
| Jerin (13)         | 29          | 29.167 | .167     | .028                  |
| Min (13)           | 31          | 29.167 | 1.833    | 3.360                 |
| Suzy (14)          | 29          | 29.167 | .167     | .028                  |
| Rich (15)          | 31          | 29.167 | 1.833    | 3.360                 |
| Sum                | 175         |        |          | 36.834                |
| Median             | 30          |        |          |                       |
| Mode               | 31          |        |          |                       |
| Range              | 24 ~ 31 (7) |        |          |                       |
| df n-1             | 5           |        |          |                       |
| Variance           | 7.366       |        |          |                       |
| Standard deviation | 2.714       |        |          |                       |

Brown, J.D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill. p. 108

## References

- Alderson, J.C., Clapham, C. and Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Bailey, K.M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Boston, MA: Heinle & Heinle Publishers.
- Brockley, T. (2009) [The Language Works online English classroom](http://eslenglishclassroom.com/) *A flash file curriculum for teachers of young learners*, <http://eslenglishclassroom.com/> (press Ctrl + click to follow the link).
- Brockley, T. (2009) [The Language Works Module Three Test One](http://www.lang-works.com/TLW/T3-1/shell.swf) *A flash file curriculum for teachers of young learners*, <http://www.lang-works.com/TLW/T3-1/shell.swf> (Ctrl + click to follow the link).
- Brockley, T. (2009) [The Language Works Module Scoring Instrument](http://eslenglishclassroom.com/Assessment-Instrument-Modular.doc) *A flash file curriculum for teachers of young learners*, <http://eslenglishclassroom.com/Assessment-Instrument-Modular.doc> (Ctrl + click to follow the link).
- Brown, J.D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Mehrens, W.A. & Lehman, I.J. (1978). *Measurement and evaluation in education and psychology* (2<sup>nd</sup> edition). Holt, Rinehart and Winston.

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

Swain, M. (1984). Large-scale communicative language testing. In S. J. Savignon, & M. S. Berns (Eds.), *Initiatives in communicative language teaching*. Reading, MA: Addison-Wesley.

van Lier, L. (2002). *An ecological-semiotic perspective on language and linguistics*. In Kramsch (2002) *Language acquisition and language socialization*. Claire Kramsch and contributors.

Wesche, M. B. (1983). Communicative testing in a second language. *The Modern Language Journal*, 67, 41-55.

### **Final Checklist**

Name(s): Timothy Brockley

Date: October 6, 2009

### **EDU 570: CLASSROOM-BASED EVALUATION**

### **CHECKLIST FOR ASSIGNMENT THREE: ORIGINAL LANGUAGE TEST DEVELOPMENT PROJECT**

This worksheet should serve as a checklist for you as you complete assignment three -- your original test development project. It should be copied and pasted into the end of your report. Please complete it as a self-assessment, using the following symbols:

+ = **good to excellent work; no questions or doubts in these areas**  
 √ = **fair to good work; some doubts and/or some confusion here**  
 - = **poor to fair work; many doubts and/or much confusion**  
 NA = **not applicable in this case**

These entries will not be “used against you.” The information will be used solely to help me improve the preparation for this assignment and to guide you in completing it. Please turn in this checklist with your completed project.

**1. I have developed test specifications for my original language test using the guidelines from Chapter 2 of Alderson, Clapham and Wall (1995).**

- + The test has been written with some real audience and/or purpose in mind.
- + I have explained the purpose(s) of my test in the report.
- + I have included the test specifications and the construct chart as appendices and have referred to them in the body of my paper.
- + I have added Alderson, Chapman and Wall (1995) to my reference list.

**2. I have drafted a language test with at least three subsets (following the specifications).**

- + The test includes both discrete point and integrative components.
- + At least one part of the test is objectively scored and at least one is subjectively scored.
- + I have included a multiple-choice subtest.
- + I have included a key for the objectively scored portion(s) of the test as an appendix.

**3. I have designed (or adapted) scoring procedures for the subjectively scored portion.**

- + I have written (or adapted) clear descriptors for the scoring levels.
- NA I have trained and normed my additional rater(s) using benchmarks and/or the descriptors.
- NA I have properly justified any adaptations I've made to an existing scale, and cited the sources that influenced my scoring procedures (ORIGINAL SCALE EMPLOYED).

**4. I have developed a key for the objectively scored portion(s).**

- + I have taken the test myself.
- + I have made certain that there is one and only one correct answer to the items in the objectively scored section(s)

**5. I have pre-piloted my test**

- + with at least two native or near-native speakers of the target language (My wife and two Korean English teachers... note that the test is for BEGINNER EFL students).
- + I have checked their responses against my predicted answer key and obtained their feedback about the instructions.
- + I have revised the test as needed and duplicated it (at my own expense).
- NA If I designed the test for a class which I myself am not teaching, I have gotten the teacher's feedback on the draft.

**6. I have administered my revised test in order to pilot test this version.**

- √ I piloted my test with at least twelve learners of the target language (There were only six learners available at this level and using this curriculum).
- + On the basis of the students' performance and direct feedback, I have determined the

clarity of the instructions, stimulus material(s) and tasks.

**7. I scored the test and analyzed the results.**

- + I have drawn frequency polygons representing the students' scores for each subtest and for the total test.
- + I have calculated the descriptive statistics (mean, mode, median, range, variance, and standard deviation) for each subtest and for the total test and included them in a table, along with the total points possible for the total test and each subtest.
- + I have evaluated the students' performance on the subjectively scored part of the test using at least two raters (Yes... Intrarater).
- + I have computed ID and IF for each item in the objectively scored subtest(s), as well as average ID, and average IF for the objectively scored subtest(s).
- + I have reported the results in table form, following APA format.
- + I have correctly interpreted and discussed the results of each of these analyses.
- + I have reported the data in tables following APA format.
- + I have computed and interpreted the inter-rater reliability (or intra-rater reliability) for the subjectively scored portion(s) of the test using Cronbach's alpha or correlation (as appropriate).
- √ I have reported the correlations, the r-squared, degrees of freedom, and probability levels.
- + If it was appropriate to use Pearson's r to calculate the correlations, I have computed the r-square to determine what information the various subtests are providing. (If not, I've explained why not.)
- + I have correctly interpreted and explained the results of the statistics from the pilot testing.

**8. I have written a coherent and well documented report of my project.**

- + The body of my report is approximately 10 to 12 pages long, typed, double-spaced, in twelve-point font (but not counting the title page, appendices or reference list).
- + Based on my analyses, I have discussed the test's strengths and weaknesses.
- + I have analyzed my test in terms of the four traditional criteria (reliability, validity, practicality, and washback) and have cited Brown's (2005) work in doing so.
- + I have included appropriate suggestions as to how my test could be improved in the future.
- √ My report clearly locates my work relative to the literature covered in the course and other appropriate research which I have found.
- + I have analyzed my test in terms of Wesche's (1983) four components of a language test
- + I have discussed my test in terms of Swain's (1984) four principles of communicative language testing.
- + I have properly cited Bailey, Brown, Wesche, Swain and other authors in my reference list.
- + My report includes a rationale explaining my choice of target audience, my choice of test methods, and the choice of constructs I measured, as well as the reasoning behind the choices.
- + The test specifications, the rating scale, and the test itself are included as appendices to the report.
- + I have provided a complete and accurate reference list (using APA format) citing at least ten appropriate items.
- + I have personally checked the reference list for completeness and accuracy.
- + I understand that the grade is final and that I may not resubmit this paper to improve the grade.
- + I have learned something in the process of developing this original test and am proud of the work that I have done.