Improvements in Automated Capturing of Psycho-linguistic Features in Reading Assessment Text

Makoto Sano

Prometric

Correspondence may be sent to:

Makoto Sano

Prometric

makoto.sano@prometric.com

Paper presented at the annual meeting of

National Council on Measurement in Education (NCME)

Washington, D.C.

April, 2016

Abstract

This study explores psycho-linguistic features associated with reading passage MC item types that can be used to predict item difficulties of these item types. The effectiveness of new functions on an NLP tool, PLIMAC (Sano, 2015) was evaluated in use of 60 items from the NAEP Grade 8 Reading assessment in 2002 through 2013. Results indicated that 12 psycho-linguistic features captured by PLIMAC accounted for 52% of the variance of item difficulty using a multiple linear regression. It also indicated that seven psycho-linguistic features accounted for 83% of the variance of item difficulty using a tree-based regression. It was a remarkable improvement from 30% and 56% of the variances accounted by Sano (2015) who also used multiple linear and tree-based regressions, respectively.

Improvements in Automated Capturing of Psycho-linguistic Features in Reading Assessment Text

Passage based multiple-choice (MC) reading comprehension test items are commonly used on high stakes assessments as these types of items have the ability to measure higher order language functioning, they can be easily machine or hand scored across large groups of examinees, and if well-constructed, have more robust statistical qualities in comparison to open ended constructed response (CR) item types.

Research has extensively explored the cognitive processes, skills, and knowledge that examinees are required to have in order to successfully answer these types of questions (see Leighton & Gierl, 2007) and several cognitive based item difficulty models have been developed (Drum, Calfee & Cook, 1981; Embretson & Wetzel, 1987; Freedle & Kostin, 1992; Gorin, 2005; Gorin & Embretson, 2006; Kirsch & Mosenthal, 1988, 1990; Mosenthal & Kirsch , 1991, Rupp, Garcia, & Jamieson, 2001; Sheehan, 1997; Sheehan, Kostin, & Persky, 2006).

The primary goal of the research is to explore and identify, through the uses of natural language processing (NLP) techniques, the psycho-linguistic features associated with reading passage MC item types that can be used to predict item difficulty levels of these item types. The NLP tool created and used for this study is the Psycho-Linguistic Measures of Assessment Content (PLIMAC, Sano, 2015). The PLIMAC was developed in Python and its Natural Language Tool Kit (NLTK, Bird, Klein, & Loper, 2009) to capture psychological and linguistic features automatically from reading passage and MC item text.

The Original Study

The original version of PLIMAC (Sano, 2015) included the following 10 functionalities: (1) an Assessment Content Text Parser, (2) a Tokenizer and Part-Of-Speech (POS) tagger, (3) a

Lemmatizer, (4) the ability to retrieve synonyms from Wordnet (Miller, 1995), (5) the ability to count of overlapping lemmas and synonyms in reading passage, item stem and options, (6) the ability to retrieve lemma frequencies from lexical resources (The Open American National Corpus; Reppen, Ide, & Suderman, 2005), (7) the ability to calculate the distance of lemmas and synonyms in reading passage, (8) the ability to reformat psycho-linguistic measures and given item stats/information, (9) the ability to produce summary statistical reports, and (10) The ability to perform multiple liner and tree-based regression.

For the initial evaluation of the PLIMAC functionalities, content from the NAEP 2011 Grade 8 Reading assessment was used (Sano, 2015). The assessment content included four reading passages with 21 MC items. Results indicated that three psycho-linguistic features captured by PLIMAC (see below) accounted for 30% and 56% of the variance of item difficulty (*p*-value) using multiple linear and tree-based regression techniques, respectively. The regression models indicated that the following variables significantly predicted item difficulty:

1) The standardized total part-of-speech count in an item stem
2) The number of synonyms in a keyed option, overlapping with the lemmas in an item stem
3) The total part-of-speech[1] count in a keyed option (for linear regression), or

   The standardized mean of locations of lemmas in a passage, overlapping with the lemmas in a keyed option (for tree-based regression)

Improvements in the Current Study

Two major issues are aimed to be improved upon in the current study. Firstly, the variance accounted for by the regression models was considerably smaller when compared to the 89%

[1] Part-of-speech is a word class or a lexical category. In this study, a part-of-speech also means an individual word tokenized and classified into a lexical category

result found by previous related research carried out by Kirsch and Mosenthal (1990). Secondly the original study was hampered by the small number of items used (21 items).

Method

For the second evaluation of the PLIMAC functionalities, the NAEP Grade 8 Reading assessment in 1992 through 2013 (retrieved from http://nces.ed.gov/NationsReportCard/nqt/) was used. The assessment content included 12 reading passages with 60 MC items. When Sheehan, Kostin, and Persky (2006) studied the NAEP Grade 8 Reading assessment, they separately modeled two of the three stimulus types which represent three broad purposes of reading: reading to gain information, reading for literary experience, and reading to perform a task (excluded by them). The 60 items chosen for the PLIMAC study are all the *reading to gain information* stimulus type following the example of Sheehan et al. (2006) who focused on this type for the tree-based regression analysis. Sheehan et al. (2006) used delta values as the item difficulty indices, these values (expected to be equated across the years) are not available for the public use. Therefore, in the PLIMAC study here, the average scale score of the candidates who answered an item correctly was used as the item difficulty index so that the item difficulties across the years could be compared.

To achieve much higher amount of variance accounted, the new version of PLIMAC was developed including the following one enhancement and three newly implemented functionalities together with the existing ones (see Figure 1):

10. **Regression analysis (Enhanced):** An enhanced graphical representation of the tree-based regression (Sheehan, 1997; Sheehan, Kostin, & Persky, 2006) including a new pruning the tree capability. The enhanced tree-based regression graphics help model item difficulty more interactively using the average difficulty of items in each node on the

horizontal axis and accommodating the capability to switch psycho-linguistic features and evaluate the improvement in the resulting accounted variance of item difficulty.

11. **Regexp parser and chunker:** A new NLP processor was implemented to help distinguish between *Given* and *Requested* information for a correct answer (Kirsch & Mosenthal, 1988/1990; Mosenthal & Kirsch,1991). The new processor provides a noun and verb phase chunking function. The chunking function segments and labels multi-token sequences as potential *Given* or *Required* information. The previous version of PLIMAC (Sano, 2015) just can handle single token (single part-of-speech) without identifying noun/verb phrases. In use of NLTK regular expression parser, the phrase structures of the passages, item stems, and options are analyzed and noun/verb phrases are parsed as chunks. In order to detect *Given* and *Requested* information more precisely, the regular expression grammar for the phrase parsing can be modified.
12. **Calculate lemma discrimination:** The function 3. *Lemmatizer* provides a process of grouping together the different inflected forms of a word (e.g. run, runs, ran and running) into a lemma (e.g. run) so they can be analyzed as a single word. The new function 12. *Calculate lemma discrimination* was implemented to calculate the Pearson correlation between lemma frequency (by item) and the item difficulty as the discrimination index of lemma under an assumption that the presence or absence of a particular word affects the item difficulty.
13. **PCA:** This function provides Principal Component Analysis to explain the maximum amount of the variance of psycho-linguistic features with a minimum number of principal components. The results can be used to interpret and analyze how the psycho-linguistic features work as predictors of item difficulty.

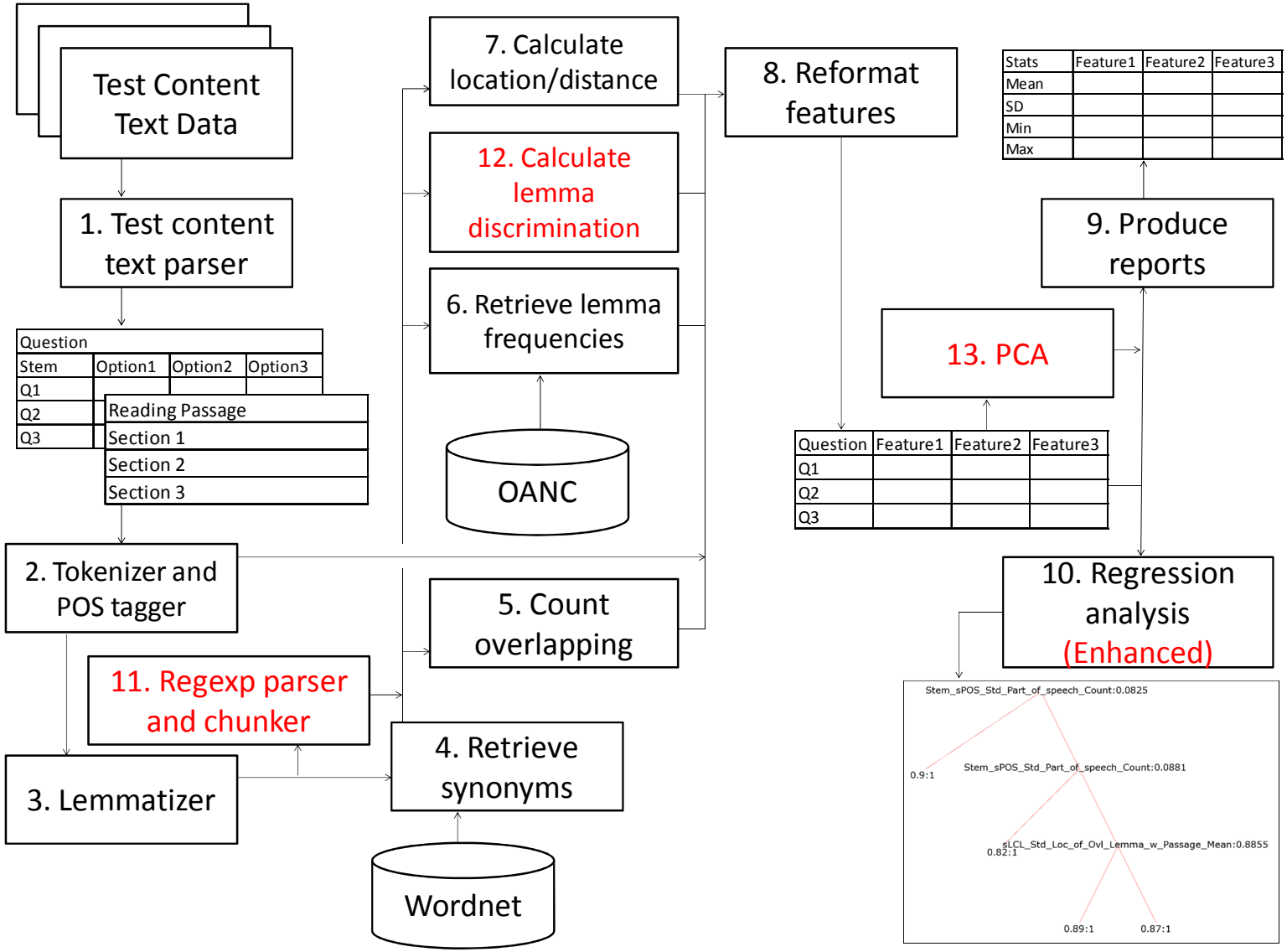

Figure 1. PLIMAC Function Diagram

Results

Captured Psycho-linguistic Features

The following psycho-linguistic features were automatically captured by PLIMAC:

1) The number of lemmas/synonyms **in a keyed option**, overlapping with the lemmas in a passage/item stem/distractor options [OLP/OSP/OLS/OSS/OLO/OSO],
2) The number of lemmas/synonyms **in an item stem**, overlapping with the lemmas in a passage/all options [OLP/OSP/ OLO/OSO],
3) The total part-of-speech count **in a passage/item stem/keyed option** [POS],
4) The mean/SD of locations of lemmas **in a passage**, overlapping with the lemmas/synonyms in an item stem/keyed option [LCL/LCS],
5) The distance between lemmas **in a passage**, overlapping with the lemmas/synonyms in an item stem, and overlapping with the lemmas/synonyms in a keyed option [DTL/DTS],
6) The mean/SD/min/max of the frequency of lemmas i**n a passage/item stem/keyed option** [FRQ],
7) The standardized (divided by the total part-of-speech count in a passage) number of lemmas/synonyms **in an item stem/keyed option**, overlapping with the lemmas in a passage [sOLP/sOSP],
8) The standardized total part-of-speech count **in an item stem** [sPOS],
9) The standardized mean of locations of lemmas **in a passage**, overlapping with the lemmas/synonyms in an item stem/keyed option [sLCL/sLCS],

10) The standardized distance between lemmas **in a passage**, overlapping with the lemmas/synonyms in an item stem, and overlapping with the lemmas/synonyms in a keyed option [sDTL/sDTS].

11) The number of noun chunks (noun phrases) **in a keyed option**, overlapping with the noun chunks in a passage/item stem/distractor options [ONCP/ONCS/ONCO], and

12) The number of noun chunks (noun phrases) **in an item stem**, overlapping with the noun chunks in a passage/all options [ONCP/ONCO].

In addition to the psycho-linguistic features above, two highly discriminating lemmas, "passage" and "author" in item stems were identified through the function 12. *Calculate lemma discrimination*. The lemma "passage" has the strongest negative discrimination (-0.27; as the correlation between the lemma count by item stem and the item difficulty) across the lemmas whose overall counts are more than 5. It means that the questions which have higher counts of the lemma "passage" in their item stems are relatively easier (i.e. the questions have lower average scale scores of the candidates who answered them correctly). The lemma "author" has the strongest positive discrimination (0.34) across the lemmas whose overall frequencies are more than 5. It means that the questions which have higher count of the lemma "author" in their item stems are relatively harder (i.e. the questions have higher average scale scores of the candidates who answered them correctly).

Table 1 shows the descriptive statistics of captured psycho-linguistic features in a keyed option, an item stem and a reading passage. The acronyms [POS], [FRQ], [LCL], [LCS], [DTL], [DTS], [OLP], [OSP], [OLS], [OSS], [OLO], [OSO], [sOLP], [sOSP], [sLCL], [sLCS], [sDTL], [sDTS], [sPOS], [ONCP], [ONCS], and [ONCO] shown at the beginning of the psycho-linguistic

features correspond to the PLIMAC psycho-linguistic features listed above. Note that the mean/SD/min/max of the frequency of lemmas [FRQ] in a passage/item stem/keyed option is calculated based on the word (lemma) frequency information retrieve from The Open American National Corpus (Reppen, Ide, & Suderman, 2005). If a particular word included in a passage/item stem/keyed option is not retrieved from the corpus (and the length of the word is more than 2) the frequency of the word is recognized as zero. Also each NAEP Grade 8 Reading assessment passage has multiple sections divided by the page breaks. If an item stem explicitly specifies a particular section (page) of the reading passage, the structural relationships between the question and passage are captured just from the specified section(s). Some of the questions share the same section of the reading passage. However, PLIMAC calculates the psycho-linguistic features in the corresponded section(s) for each question independently. Thus some of the reading passage sections were double-counted across questions (see the last section of the Table 1: Psycho-linguistic Features in A Passage).

Pearson correlations between each feature variable and item difficulty (average scale score) were calculated in Table 2. The features which have the absolute value of correlations at least 0.30 are shown as bold styled. The strongest negative correlation shown over the all 60 questions was [POS] The total part-of-speech count in a passage (-0.49) followed by [FRQ] The mean of lemma frequency in a stem (-0.16). That means much longer passages with more common words in the item stems make questions easier to answer. The strongest positive correlation to the average scale score for all questions was [sPOS] The standardized total part-of-speech count in a stem (0.35) followed by [POS] The total part-of-speech count in a stem (0.31) and [sOLP] The standardized number overlapping lemma w/ passage in a stem (0.29).

Table 1. Descriptive Statistics of Psycho-Linguistic Features

	Average Scale Score	Psycho-linguistic Features in A Keyed Option				
		[POS] Part-of-speech Count	[FRQ] Lemma Frq. Mean	[FRQ] Lemma Frq. SD	[FRQ] Lemma Frq. Min	[FRQ] Lemma Frq. Max
Count	60	60	60	60	60	60
Mean	272.57	6.62	27923.89	35733.12	1152.27	99446.60
SD	3.98	2.81	27762.89	37046.49	2286.44	102201.41
Min	262.00	1.00	23.00	0.00	0.00	23.00
Max	281.00	13.00	104959.38	139908.21	15719.00	339971.00

	[LCL] Location of Overlapping Lemma w/ Passage Mean	[LCL] Location of Overlapping Lemma w/ Passage SD	[LCS] Location of Overlapping Synonym w/ Passage Mean	[LCS] Location of Overlapping Synonym w/ Passage SD	[DTL] Distance of Overlapping Lemmas in Passage	[DTS] Distance of Overlapping Synonyms in Passage
Count	50	50	52	52	46	52
Mean	463.10	174.75	435.36	179.24	88.54	140.51
SD	272.71	100.67	343.71	102.39	105.12	224.88
Min	123.00	0.00	25.00	0.00	1.86	0.79
Max	1416.75	419.40	2505.00	346.12	492.83	1408.43

	[OLP] Number Overlapping Lemma w/ Passage	[OSP] Number Overlapping Synonym w/ Passage	[OLS] Number Overlapping Lemma w/ Stem	[OSS] Number Overlapping Synonym w/ Stem	[OLO] Number Overlapping Lemma w/ Distractor Options	[OSO] Number Overlapping Synonym w/ Distractor Options
Count	60	60	60	60	60	60
Mean	3.75	4.37	0.53	0.23	3.02	0.32
SD	2.38	3.94	0.81	0.53	2.68	0.68
Min	0	0	0	0	0	0
Max	9	20	3	2	9	3

	[sOLP] Std. Number Overlapping Lemma w/ Passage	[sOSP] Std. Number Overlapping Synonym w/ Passage	[sLCL] Std. Location of Overlapping Lemma w/ Passage Mean	[sLCS] Std. Location of Overlapping Synonym w/ Passage Mean	[sDTL] Std. Distance of Overlapping Lemmas in Passage	[sDTS] Std. Distance of Overlapping Synonyms in Passage
Count	60	60	50	52	46	52
Mean	0.0057	0.0064	0.6420	0.5852	0.1257	0.1567
SD	0.0039	0.0050	0.1796	0.2178	0.1236	0.1533
Min	0.0000	0.0000	0.3672	0.0329	0.0021	0.0026
Max	0.0162	0.0192	1.1528	1.0487	0.5620	0.5541

	[ONCP] Number Overlapping NChunk w/ Passage	[ONCS] Number Overlapping NChunk w/ Stem	[ONCO] Number Overlapping NChunk w/ Distractor Options
Count	60	60	60
Mean	0.18	0.02	0.15
SD	0.39	0.13	0.55
Min	0	0	0
Max	1	1	3

Table 1(continued). Descriptive Statistics of Psycho-Linguistic Features (continued)

Psycho-linguistic Features in A Stem

	[POS] Part-of-speech Count	[FRQ] Lemma Frq. Mean	[FRQ] Lemma Frq. SD	[FRQ] Lemma Frq. Min	[FRQ] Lemma Frq. Max
Count	60	60	60	60	60
Mean	15.47	35931.73	56312.77	288.48	180512.98
SD	8.84	24685.38	38523.87	508.67	123449.61
Min	5.00	3046.00	3708.64	0.00	12090.00
Max	60.00	135831.86	154697.24	2150.00	339971.00

	[OLP] Number Overlapping Lemma w/ Passage	[OSP] Number Overlapping Synonym w/ Passage	[OLO] Number Overlapping Lemma w/ All Options	[OSO] Number Overlapping Synonym w/ All Options
Count	60	60	60	60
Mean	7.37	6.92	2.07	0.72
SD	4.38	4.54	2.72	1.03
Min	0	1	0	0
Max	20	20	10	5

	[sPOS] Std. Part-of-speech Count	[sOLP] Std. Number Overlapping Lemma w/ Passage	[sOSP] Std. Number Overlapping Synonym w/ Passage
Count	60	60	60
Mean	0.0299	0.0139	0.0109
SD	0.0305	0.0134	0.0087
Min	0.0035	0.0000	0.0011
Max	0.1754	0.0585	0.0409

	[ONCP] Number Overlapping NChunk w/ Passage	[ONCO] Number Overlapping NChunk w/ All Options
Count	60	60
Mean	0.65	0.07
SD	0.76	0.41
Min	0	0
Max	3	3

Psycho-linguistic Features in A Passage

	[POS] Part-of-speech Count	[FRQ] Lemma Frq. Mean	[FRQ] Lemma Frq. SD	[FRQ] Lemma Frq. Min	[FRQ] Lemma Frq. Max
Count	60	60	60	60	60
Mean	765.55	36320.73	67942.15	0.20	344751.80
SD	426.93	8089.95	9512.96	0.88	3936.45
Min	227.00	23190.98	49647.07	0.00	339971.00
Max	2542.00	61003.25	94436.59	4.00	347939.00

Table 2. Correlations Between Psycho-Linguistic Features and Average Scale Score

Psycho-linguistic Features in A Keyed Option	Correlation to Average Scale Score
[POS] Part-of-speech Count	0.20
[FRQ] Lemma Frq. Mean	0.22
[FRQ] Lemma Frq. SD	0.13
[FRQ] Lemma Frq. Min	0.09
[FRQ] Lemma Frq. Max	0.11
[LCL] Location of Overlapping Lemma w/ Passage Mean	**-0.50**
[LCL] Location of Overlapping Lemma w/ Passage SD	-0.01
[LCS] Location of Overlapping Synonym w/ Passage Mean	-0.19
[LCS] Location of Overlapping Synonym w/ Passage SD	0.23
[DTL] Distance of Overlapping Lemmas in Passage	-0.22
[DTS] Distance of Overlapping Synonyms in Passage	**-0.34**
[OLP] Number Overlapping Lemma w/ Passage	0.07
[OSP] Number Overlapping Synonym w/ Passage	0.08
[OLS] Number Overlapping Lemma w/ Stem	0.01
[OSS] Number Overlapping Synonym w/ Stem	0.06
[OLO] Number Overlapping Lemma w/ Distractor Options	0.09
[OSO] Number Overlapping Synonym w/ Distractor Options	0.15
[sOLP] Std. Number Overlapping Lemma w/ Passage	0.26
[sOSP] Std. Number Overlapping Synonym w/ Passage	0.26
[sLCL] Std. Location of Overlapping Lemma w/ Passage Mean	**-0.30**
[sLCS] Std. Location of Overlapping Synonym w/ Passage Mean	0.18
[sDTL] Std. Distance of Overlapping Lemmas in Passage	-0.03
[sDTS] Std. Distance of Overlapping Synonyms in Passage	-0.17
[ONCP] Number Overlapping NChunk w/ Passage	0.04
[ONCS] Number Overlapping NChunk w/ Stem	0.08
[ONCO] Number Overlapping NChunk w/ Distractor Options	0.02

Psycho-linguistic Features in A Stem	Correlation to Average Scale Score
[POS] Part-of-speech Count	**0.31**
[FRQ] Lemma Frq. Mean	-0.16
[FRQ] Lemma Frq. SD	-0.07
[FRQ] Lemma Frq. Min	0.18
[FRQ] Lemma Frq. Max	0.04
[OLP] Number Overlapping Lemma w/ Passage	0.20
[OSP] Number Overlapping Synonym w/ Passage	-0.11
[OLO] Number Overlapping Lemma w/ All Options	-0.01
[OSO] Number Overlapping Synonym w/ All Options	0.16
[sPOS] Std. Part-of-speech Count	**0.35**
[sOLP] Std. Number Overlapping Lemma w/ Passage	0.29
[sOSP] Std. Number Overlapping Synonym w/ Passage	0.25
[ONCP] Number Overlapping NChunk w/ Passage	0.08
[ONCO] Number Overlapping NChunk w/ All Options	0.08

Psycho-linguistic Features in A Passage	Correlation to Average Scale Score
[POS] Part-of-speech Count	**-0.49**
[FRQ] Lemma Frq. Mean	0.26
[FRQ] Lemma Frq. SD	0.22
[FRQ] Lemma Frq. Min	0.20
[FRQ] Lemma Frq. Max	-0.11

Table 3. 12 Best Predictors of Item Difficulty in Multiple Linear Regression Analysis

Variables	Coefficient	Standard Error	*t*-value	*p*-value
Psycho-linguistic Features in A Keyed Option				
[POS] Part-of-speech Count	0.2434	0.1655	1.4700	0.1481
[FRQ] Lemma Frq. Max	-0.0001	0.0000	-2.0400	0.0474
[FRQ] Lemma Frq. SD	0.0002	0.0001	2.1600	0.0362
[OSO] Number Overlapping Synonym w/ Distractor Options	1.6591	0.5985	2.7700	0.0080
Psycho-linguistic Features in A Stem				
[OLP] Number Overlapping Lemma w/ Passage	0.1993	0.1165	1.7100	0.0937
[OLO] Number Overlapping Lemma w/ All Options	-0.3394	0.1671	-2.0300	0.0479
[FRQ] Lemma Frq. Mean	-0.0001	0.0000	-3.3700	0.0015
[FRQ] Lemma Frq. SD	0.0001	0.0000	3.2800	0.0020
[FRQ] Lemma Frq. Min	0.0019	0.0009	2.0700	0.0444
Psycho-linguistic Features in A Passage				
[POS] Part-of-speech Count	-0.0053	0.0011	-5.0000	0.0000
[FRQ] Lemma Frq. Max	0.0003	0.0001	2.4500	0.0179
Highly Discriminating Lemma				
Count of lemma “author” in a stem	2.4197	1.0727	2.2600	0.0288
Intercept	180.4939	37.1262	4.8600	0.0000

That means if the (relative) length of an item stem (compared to the corresponded reading passage) is longer and more overlapping words in the item stem with the reading passage, the question becomes harder.

Multiple Linear Regression Analysis

Multiple linear regression analyses were performed in use of the feature variables captured from all 60 questions as the predictors of item difficulty. A result of regression analysis identified the 12 best predictors of item difficulty in Table 3. The amount of variance accounted by the 12 predictors was 0.52 (as the adjusted R^2). This 52% shows a remarkable improvement from Sano's (2015) 30% of the variance accounted using multiple linear regression analysis. But it is still smaller when compared to the 89% of the variance result found by Kirsch and Mosenthal (1990). Since the PLIMAC result shows relatively smaller correlations than Kirsch and Mosenthal's (1990; the maximum of correlation was 0.85), PLIMAC still has some room of improvement in capturing the psycho-linguistic features.

Tree-based Regression Analysis

Tree-based regression analyses were performed using the feature variables captured from all 60 questions as the predictors of item difficulty. In this study, the Classification and Regression Trees (CART) algorithm (Breiman, Friedman, Olshen, & Stone, 1984) was used as the same manner of Gao and Rogers (2011), Rupp et al. (2001), Sheehan (1997), and Sheehan et al. (2006). The CART algorithm forms clusters of questions which have similar psycho-linguistic feature values, by successively splitting the questions into subsets called nodes. As the first step of the recursive partitioning, all possible splits of the question groups are evaluated by deviance. In this study, the deviance D is the sum of squared differences between an item average scale score and the mean of average scale score of all questions belonging to a single node.

$$D(y, \hat{y}) = \sum (y_i - \hat{y})^2 \quad (1)$$

where $\hat{y}$ is the mean of average scale score and y_i is the average scale score of item i.

Then the best split (by a particular psycho-linguistic feature value) is found as maximizing the difference in deviance ΔD between the parent node and the sum of the two child nodes.

$$\Delta D = D(y, \hat{y}) - D_{split}(y, \hat{y}_L, \hat{y}_R) \quad (2)$$

$$D_{split}(y, \hat{y}_L, \hat{y}_R) = \sum (y_i - \hat{y}_L)^2 + \sum (y_i - \hat{y}_R)^2 \quad (3)$$

where $\hat{y}_L$ is the mean of average scale score in the left child node and $\hat{y}_R$ is the mean of average scale score in the right child node.

Figure 2 shows a result of tree-based regression analysis using seven psycho-linguistic features. These seven features were automatically identified by the algorithm coupled with the new feature of pruning the tree. The result accounted 83% of the variance of item difficulty. In the root node, the best splitter can be found as [sPOS] The standardized total part-of-speech count in a stem at the value of 0.0049. The left hand side of the child node shows a group of questions whose feature values of [sPOS] (in a stem) are below 0.0049, while the right hand side of the childe node shows a group of questions whose feature values of [sPOS] are equal to or greater than 0.0049. That means, if the relative length of an item stem is longer in comparison with the corresponded reading passage, the question becomes harder and then located right hand side of the difficulty scale. The second split (in the second node) applies [FRQ] The mean of lemma frequency in a passage at 52044.5885 followed by further splitting variables of [FRQ] The minimum of lemma frequency in a stem, [OSO] The number overlapping synonym w/ distractor options (in a keyed option), [FRQ] The mean of lemma frequency (in a keyed option), [FRQ] The SD of lemma frequency (in a keyed option), and [sOSP] The standardized number overlapping synonyms w/ passage in a stem.

Figure 2. A Tree-based Regression Analysis with Seven Psycho-linguistic Features

In Figure 2, a leaf node shows the mean of the average scale scores of the question(s) belonging to the leaf node. The leaf node located the horizontal scale of the item difficulty. The leaf node also indicates the number of question(s) after the colon. For example, a leaf located on the far left side shows the values of 264.75: 4. This means that the leaf node has four questions whose mean of the average scale scores is 264.75. Right under the right third node ([OSO] Number Overlapping Synonym w/ Distractor Options), crossing branches can be found. This means that the questions in the right child node (i.e. the true branch) have an equal or greater psycho-linguistic value of the threshold value (as 1.0) but the mean of the average scale scores of the node is lower than the left childe node (i.e. the false branch). That's why, even the branch begins from the right hand side of the third node, it goes down to the left hand side to the fourth node crossing the false branch.

If there is a lot of crossing branches observed from the top to the bottom of the tree except such a type of the crossing branches mentioned above, there might be an overfitting to the specified input data and the estimated item difficulties may be fluctuated on the scale in every step of the recursive partitioning. In order to avoid such a fluctuation, the pruning the tree feature works by checking these overfitted pairs of nodes that have a common parent and verifies if the pruning (merging) the nodes would increase the deviance just within a certain amount of acceptable extent. In Figure 2, the threshold of the acceptable level was 30.0.

Principal Component Analysis

A principal component analysis was performed for the feature variables captured from all 60 questions. Figure 3 shows the scree plot of eigenvalues from the principal component analysis. The first six components accounted 67% of the variance. Table 4 shows the component loadings, these are the correlations between the features and the components. The component loadings

whose absolute value are at least 0.70 are shown as bold styled (as the strong correlations indicated). The first principal component is strongly correlated with five features in a stem and the second is strongly correlated with other five features in a keyed option. The third principal component is also strongly correlated with two features in a passage. This suggests that the captured features are characterized more by the structure of questions (i.e. passage, item stem and keyed option) than by the type of features as the number of overlapping lemmas/synonyms, the total part-of-speech count, and the frequency of lemmas. The fourth and fifth principal components do not show strong correlations but some features in a stem regarding the lemma frequencies have moderate correlations. The sixth principal component is strongly correlated with [OSP] The number overlapping synonym w/ passage in a stem but it does not appear as an independent variable in regression analyses.

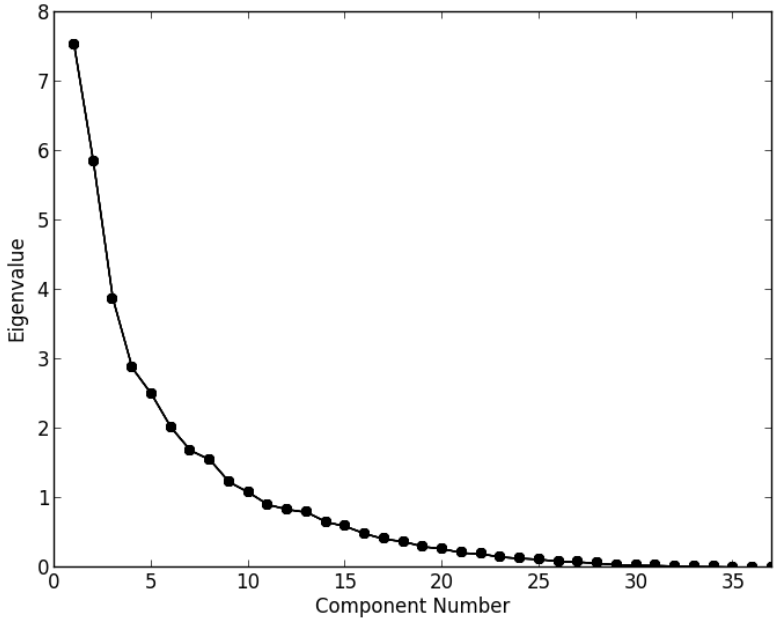

Figure 3. The Scree Plot of Eigenvalues from PCA

Table 4. Component Loadings of PCA

Psycho-linguistic Features in A Keyed Option	Comoponet1	Comoponet2	Comoponet3	Comoponet4	Comoponet5	Comoponet6
[POS] Part-of-speech Count	-0.166	**0.782**	-0.012	0.094	0.120	0.090
[FRQ] Lemma Frq. Mean	0.043	0.674	-0.351	-0.325	-0.358	-0.039
[FRQ] Lemma Frq. SD	0.042	**0.718**	-0.220	-0.365	-0.375	-0.087
[FRQ] Lemma Frq. Min	-0.050	0.019	-0.366	-0.077	-0.133	0.070
[FRQ] Lemma Frq. Max	0.036	**0.748**	-0.208	-0.379	-0.316	-0.097
[OLP] Number Overlapping Lemma w/ Passage	-0.113	**0.877**	0.041	-0.006	0.152	-0.041
[OSP] Number Overlapping Synonym w/ Passage	-0.138	**0.717**	-0.017	0.178	0.323	-0.092
[OLS] Number Overlapping Lemma w/ Stem	-0.626	0.282	0.050	-0.206	0.104	0.099
[OSS] Number Overlapping Synonym w/ Stem	-0.556	-0.033	-0.110	0.287	0.203	0.069
Options	-0.421	0.694	0.122	0.020	-0.249	0.094
[OSO] Number Overlapping Synonym w/ Distractor Options	-0.104	0.305	-0.154	0.439	0.323	-0.050
[sOLP] Std. Number Overlapping Lemma w/ Passage	-0.568	0.584	0.103	0.052	-0.092	0.293
[sOSP] Std. Number Overlapping Synonym w/ Passage	-0.522	0.558	-0.077	0.298	0.185	0.134
[ONCP] Number Overlapping NChunk w/ Passage	0.093	0.263	0.409	0.409	0.199	-0.138
[ONCS] Number Overlapping NChunk w/ Stem	-0.492	0.046	0.401	0.316	-0.055	-0.148
[ONCO] Number Overlapping NChunk w/ Distractor Options	-0.254	0.311	0.405	0.141	-0.213	0.006
Psycho-linguistic Features in A Stem						
[POS] Part-of-speech Count	**-0.709**	-0.382	-0.264	-0.019	-0.157	-0.255
[FRQ] Lemma Frq. Mean	-0.362	-0.078	0.108	-0.509	0.519	0.411
[FRQ] Lemma Frq. SD	-0.446	-0.076	0.184	-0.584	0.477	0.314
[FRQ] Lemma Frq. Min	0.264	0.056	0.061	0.416	0.042	0.378
[FRQ] Lemma Frq. Max	-0.615	-0.180	0.150	-0.515	0.371	0.198
[OLP] Number Overlapping Lemma w/ Passage	**-0.763**	-0.245	-0.090	-0.213	-0.090	-0.426
[OSP] Number Overlapping Synonym w/ Passage	-0.335	-0.135	-0.090	-0.088	0.275	**-0.703**
[OLO] Number Overlapping Lemma w/ All Options	-0.695	0.265	0.019	-0.071	0.099	0.018
[OSO] Number Overlapping Synonym w/ All Options	-0.471	0.163	-0.208	0.456	0.293	-0.116
[sPOS] Std. Part-of-speech Count	**-0.769**	-0.398	-0.138	0.048	-0.230	0.151
[sOLP] Std. Number Overlapping Lemma w/ Passage	**-0.845**	-0.346	0.025	-0.046	-0.216	0.088
[sOSP] Std. Number Overlapping Synonym w/ Passage	**-0.729**	-0.327	-0.132	0.138	0.001	-0.193
[ONCP] Number Overlapping NChunk w/ Passage	-0.585	0.037	0.281	-0.228	-0.144	-0.244
[ONCO] Number Overlapping NChunk w/ All Options	-0.467	0.092	0.387	0.377	-0.007	-0.204
High Discriminating Lemma: "author"	-0.117	-0.157	-0.698	0.127	-0.039	0.221
High Discriminating Lemma: "passage"	0.363	-0.054	0.181	0.053	0.392	0.173
Psycho-linguistic Features in A Passage						
[POS] Part-of-speech Count	0.467	0.137	0.114	-0.207	0.514	-0.442
[FRQ] Lemma Frq. Mean	0.028	0.087	**-0.894**	0.032	0.145	-0.017
[FRQ] Lemma Frq. SD	0.075	0.096	**-0.879**	0.067	0.232	-0.054
[FRQ] Lemma Frq. Min	-0.383	-0.198	-0.599	0.271	0.132	0.182
[FRQ] Lemma Frq. Max	0.270	0.278	-0.179	-0.294	0.382	-0.352
Eigenvalue	7.53	5.84	3.87	2.88	2.50	2.02
Percent of Variance Accounted	20.36	15.79	10.47	7.77	6.75	5.45

Discussion

Comparison with the Original Study

In comparison with the initial evaluation of the PLIMAC functionalities (Sano, 2015), a couple of differences can be found from the correlations between psycho-linguistic features and item difficulties. First, item difficulty indices are different between the two studies, *p*-value in Sano's (2015) and the average scale score in the current study. Higher *p*-values mean the questions are easier while higher average scale scores mean the questions are harder. Thus, the same polarity of correlations (between the features and the item difficulty) across the two studies has opposite meaning. If the correlations are both positive (negative) in Sano's (2015) and this study, the psycho-linguistic feature may influence the item difficulties completely differently. Take into consideration the difference, the second finding was that the feature [sPOS] The standardized total part-of-speech count in a stem has completely different effects on the item difficulty. That is, if the relative length of an item stem is longer (in comparison with the corresponded reading passage), the question becomes easier in Sano's study (2015) as opposed to this study. In this study, if the relative length of an item stem is longer, the question becomes harder. The following is an interpretation of the effect in Sano's study (2015):

Five out of the seven questions from the easier question group (the mean *p*-value: 0.80) have repetitive phrases in the item stem quoted from the reading passages (in order to accommodate the distinction of the *Given* and *Required* information). The repetitive phrases make the average length of the item stems in the easier question group much longer than the harder question group (the mean *p*-value: 0.64) and affect the feature values of [sPOS] The standardized part-of-speech count in a stem. These repetitive phrases may accommodate more information given by the item

stems and make questions easier. Finally, the questions which have much greater feature values of [sPOS] have resulted in much higher *p*-values.

In this study, three out of the five questions originally used for Sano's study (2015) from 2011 NAPE Grade 8 Reading assessment were displaced since they were not categorized as the *reading to gain information* stimulus type, therefore just eight (including two 2nd hardest questions as these average scale scores are 280) out of the 50 newly added questions from 2002 to 2009 and 2013 NAEP Grade 8 Reading assessment have such repetitive phrases in the item stems. They are no longer a major part of the questions. A similar completely different effect can be found by the feature [POS] The total part-of-speech count in a passage. In this study, the questions which have longer passages tend to be easier (the correlation between the feature and item difficulty is -0.49) as opposed to Sano's study (2015). This is another effect by newly added items. They have much shorter item stems corresponded to longer passages and these questions tend to be easier.

The third finding through the comparative study of correlations was that the strongest negative correlation shown in Sano's study as [OSS] The number overlapping synonym w/ stem in a keyed option (-0.42) is currently just as 0.06. Because five out of the six questions used for Sano's study (2015) which have overlapping synonyms in the item stems were not the *reading to gain information* stimulus type, and they were displaced in this study. Thus, the correlation between this feature and item difficulty is no longer observed. All these changes and the highest discriminating lemma (as mentioned below) contributed gaining the variance of item difficulty accounted by the multiple linear regression as 52%.

Newly Added Psycho-Linguistic Features

The new function 11. *Regexp parser and chunker* enables capturing five new psycho-linguistic features as the number of noun chunks (noun phrases) in a keyed option, overlapping with the lemmas in a passage/item stem/distractor options [ONCP/ONCS/ONCO], and the number of noun chunks (noun phrases) in an item stem, overlapping with the lemmas in a passage/all options [ONCP/ONCO]. Against the initial expectation, these five features did not show higher correlations in Table 2 because the total count of captured noun chunks was very limited and the maximum of these feature values are small, 1 to 3 as shown in Table 1. Table 5 shows a list of the parsed noun chunks whose total count across all questions are at least 3. Note that some of the nouns as a part of the chunks are lemmatized and have different inflected forms from the original words in the questions (e.g. “specie” was originally “species”, and “unite state” was “United States”). One of the noun phrase “invasive specie(s)” appears on all the options of a question and it makes the count is the highest. The original purpose of implementing the new function was to distinguish *Given* and *Requested* information more precisely in chunking the pieces of information into noun and verb phrases both. However, very little number of verbs which can work as a part of chunk and a trigger of identifying *Given* and *Requested* information were found (e.g. assume, mean, suggest, and so on) especially in item stem, then just the result of capturing none phrases were used in this study.

In compensation for non-captured verb phrases, high discriminating lemmas in item stems were identified through function 12. *Calculate lemma discrimination*. The lemma “author” which shows the strongest positive discrimination (0.34) across the lemmas contributed the best prediction of item difficulty in the multiple linear regression. For example, the lemma “author” captured from the stems was followed by the words “tells”, “assumes”, or “mentions” in their

sentences. This type of question asking author's assumptions or thoughts were harder than the other type of question to locate the *Requested* information, and then the lemma "author" was identified as the highest discriminating lemma.

Table 5. Noun Chunks and the Count across All Questions

Noun Chunk	Count
invasive specie	6
unite state	6
page 3	5
cane toad	5
white shark	5
oregon trail	4
alien invasion	4
mental illness	4
main purpose	3

Comparison with Sheehan et al.'s Study (2006)

In comparison with the study by Sheehan, Kostin, & Persky (2006), several distinctions can be found. First, most of the task features in their study were hand-coded and a limited number of automated task features were used. In the current study all the features were captured automatically. Second, the total number of items in their study (N= 104) was larger than this study (N= 60). In general, the lager number of items for regression analysis makes it much harder to find the best fit regression model which gains higher adjusted R^2. In their study, the highest adjusted R^2 observed was 48% by the tree-based regression analysis. Whereas the best adjusted R^2 in the current study was 83% even though much higher, it's hard to say that this study achieved a much better model fit than theirs. Third, their items include 59 CR items and 45 MC items while only MC items were included in the current study. They categorized items into

three categories of search strategies for *Requested* information as a) *Right There Items*, b) *Think and Search*, and c) *On My Own Items*. The items in the last category were identified as the hardest, as asking examinees to state and support an opinion, or to relate concepts discussed in the text (passage) to their own personal experiences (Sheehan et al. 2006). The *On My Own Items* are not included in this study since there is no CR item. Thus, the tree-based regression model identified here is completely different from Sheehan et al's (2006). But at least, the hardest item group, in this study, observed on the far right hand side of the difficulty scale in the tree-based regression (see Figure 2) can be categorized as b) *Think and Search* items. They are harder than a) *Right There Items* according to Sheehan et al's study (2006) and the hardest item tasks are described as "Recognize implicit supporting idea in persuasive essay", "Recognize meaning of word as used in persuasive essay", and "Recognize generalization of main idea of persuasive essay based on one paragraph", respectively. These answers to the questions cannot be found in the passage texts literally as opposed to the nature of the a) *Right There Items*. They also include the highest discriminating lemma "author" in item stems and ask to find answers regarding author's thoughts.

Conclusion

The effectiveness of new functions on the NLP tool, PLIMAC (Sano, 2015) was evaluated using 60 items from the NAEP Grade 8 Reading assessment in 2002 through 2013. The key factors of the remarkable improvements from the previous study in multiple linear (30% to 52% as the adjusted R^2) and tree-based regressions (56% to 83% as the adjusted R^2) were, a) focusing on the *reading to gain information* stimulus type as following Sheehan et al. (2006) and observed much better fit to the item difficulty prediction models in use of the captured psycho-linguistic features, b) the pruning the tree feature in the tree-based regression contributing to

avoid overfitting to a particular data set, and c) the two new functions of Regexp parser and chunker, and PCA enabling explanatory analyses to find how the psycho-linguistic features work for item difficulty prediction and leading the finding of the highest discriminating lemma "author" as an important feature of the item difficulty prediction.

Future Research

Even though remarkable improvements in item difficulty prediction were observed, these regression models are still just the best fit models for the 60 particular items. Thus, it might be worthwhile in future research to try predicting new item difficulties of NAEP Grade 8 Reading assessment in 2015 once the item contents are released. In this study, some of the features regarding the (standardized) mean/SD of locations and the (standardized) distance between lemmas were not used because these features were not captured from the all 60 items. There is some room of improvement how to capture and use these features more efficiently considering possible imputations of the missing values. It's also worth considering much better use of WordNet to capture the synonymous overlapping and measuring semantic similarity as suggested by Belov and Knezevich (2008).

Acknowledgements

I would like to express my deepest appreciation to our corporate advisor, Ric Luecht and my colleague, Ian Clifford, whose contribution in stimulating suggestions helped me to coordinate my research especially in writing this paper.

References

Belov, D. I., and Knezevich, L. (2008). *Automatic Prediction of Item Difficulty Based on Semantic Similarity Measures*. Law School Admission Council Research Report 08-04, Newtown, PA: Law School Admission Council, Inc.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth, Inc.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 16, 486-514.

Embretson, S. E. & Wetzel, C. D. (1987). Component latent models for paragraph comprehension. *Applied Psychological Measurement*, 11, 175-193.

Freedle, R., & Kostin, I. (1992).*The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main ideas, inferences, and explicitstatements*. ETS Research Report RR-91-59, Princeton, NJ: Educational Testing Service.

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28, 77-104.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351-373.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 3, 394-411.

Kirsch, I. S., & Mosenthal, P. B. (1988). *Understanding document literacy: Variables underlying the performance of young adults*. ETS Research Report RR-88-62, Princeton, NJ: Educational Testing Service.

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, 25, 5-30.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26, 3-16.

Miller,G. A. (1995). WordNet: A Lexical database for English. *Communications of the ACM*, 38, 39-41.

Mosenthal, P.B., & Kirsch, I.S. (1991). Toward an explanatory model of document literacy. *Discourse Processes*, 14, 147-180.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) Second Release LDC2005T35*, Web Download, Philadelphia, PA: Linguistic Data Consortium.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1, 185-216.

Sano, M. (2015). *Automated Capturing of Psycho-linguistic Features in Reading Assessment Text*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement*, 34, 333-352.

Sheehan, K. M., Kostin, I., & Persky, H. (2006). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP grade 8 reading assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.