CaMLA Working Papers

2014–02

Predicting Listening Item Difficulty with Language Complexity Measures: A Comparative Data Mining Study

Vahid Aryadoust
Centre for English Language Communication
National University of Singapore

Christine C. M. Goh
National Institute of Education
Nanyang Technological University

Predicting Listening Item Difficulty with Language Complexity Measures: A Comparative Data Mining Study

Authors

Vahid Aryadoust
Centre for English Language Communication, National University of Singapore

Christine C. M. Goh
National Institute of Education, Nanyang Technological University

About the Authors

Vahid Aryadoust received his PhD in applied linguistics from the National Institute of Education, Nanyang Technological University, Singapore. He is currently working as a lecturer at the Center for English Language Communication of the National University of Singapore where he teaches academic writing and oral communication skills to science students. His research focuses on the application of item response theory, structural equation modelling, and latent class models in pedagogy and assessment. He also has hands-on research experience in Hybrid Intelligent Models and Genetic Programming. His latest book (with Janna Fox of Carleton University) is an edited volume entitled Current Trends in Language Testing in the Pacific Rim and the Middle East.

Christine C. M. Goh is Professor of Linguistics and Language Education at the English Language & Literature Academic Group of the National Institute of Education, Nanyang Technological

Table of Contents

University, Singapore. Her teaching and research interests are in the development and assessment of listening, speaking, and pronunciation of second language and bilingual learners, discourse intonation features of speakers of English as an international language, the role of oracy in language learners' thinking development and academic learning, metacognition and language learning, and language teacher cognition. She has authored many peer-refereed journal articles, books and book chapters, as well as research technical reports on these topics.

Abstract

Modelling listening item difficulty remains a challenge to this day. Latent trait models such as the Rasch model used to predict the outcomes of test takers' performance on test items have been criticized as "thin on substantive theory" (Stenner, Stone, & Burdick, 2011, p.3). The use of regression models to predict item difficulty also has its limitations because linear regression assumes linearity and normality of data which, if violated, results in a lack of fit. In addition, classification and regression trees (CART), despite their rigorous algorithm, do not always yield a stable tree structure (Breiman, 2001).

Another problem pertains to the operationalization of dependent variables. Researchers have relied on content specialists or verbal protocols elicited from test takers to determine the variables predicting item difficulty. However, even though content specialists are highly competent, they may not be able to determine precisely the lower-level comprehension processes used by low-ability test takers just by reading test items. Furthermore, verbal protocols elicited during test-taking may interfere with the cognitive task (Sawaki & Nissan, 2009).

Previous reading research uses CART to investigate item difficulty, but despite being competently conducted, the resultant regression trees have been inconsistent across test forms (Gao, 2006). In the current proposed study, two classes of Artificial Neural Networks (i.e., Multilayer Perceptron ANN and the Adaptive Neuro-Fuzzy Inference System or ANFIS) are used to explore the effect of lexical and syntactic complexity of items and texts on MET listening items' difficulty in seven MET listening tests. In addition, Coh-Metrix measures which have conventionally been used to measure reading text complexity are also applied in this investigation (Riazi & Knox, 2014). To our knowledge, these methods have not been applied to investigate lexical and syntactic complexity of the listening texts and items. Findings from the study will contribute to the validity argument for the Michigan English Test (MET) and provide additional empirical evidence to assist CaMLA in evaluating the quality of listening test items (see also Goh & Aryadoust, 2010).

Background

Developing a predictive theory of item difficulty allows researchers both to predict the difficulty of test items with reference to salient item- and text-level variables and to manipulate test item difficulty in predictable ways by altering those variables (Daftarifard & Lange, 2009; Perkins, Gupta, & Tammana, 1995). Modelling listening test item difficulty, however, remains a challenge to this day.

To develop predictive theories, researchers have adopted a variety of statistical tools to explore the variables that influence item difficulty in listening comprehension. Notable examples include regression models (Grant & Ginther, 2000), artificial neural networks (Perkins et al., 1995), and classification and regression trees (CART) (Gao, 2006). Application of latent trait models such as the Rasch model to predict the outcomes of test takers' performance on test items has also been criticized as "thin on substantive theory" (Stenner, Stone, & Burdick, 2011, p. 3). Although these studies have informed the field, their methodologies have certain limitations.

The first problem pertains to data analysis tools: some of these studies have relied on linear regression, which assumes linearity and normal distribution. If these assumptions are violated, the model does not fit, likely leading researchers to refute the theory-informed hypotheses. However, the relationships among variables in language and educational assessment may be nonlinear. As a result, the "validity of the studies in which multiple regression is used to predict item difficulty is not high" (Perkins et al., 1995, p. 35). In addition, although CART does not make any assumptions regarding normality and has rigorous algorithms, as Gao's (2006) study of a MELAB reading test showed, tree structures are not stable across samples (Breiman, 2001).

Another problem pertains to the operationalization of independent variables. Researchers have relied on content specialists' evaluation or test takers' verbal protocols to determine the variables predicting item difficulty. However, even though content specialists are highly competent, they may not be able to determine precisely the lower-level comprehension processes used by low-ability test takers just by reading test items (Alderson & Kremmel, 2013; see also Zhang, Goh, & Kunnan, 2014, for the effect of test takers' cognitive and metacognitive strategies). Furthermore, there is a concern that verbal protocols elicited during test-taking may interfere with the cognitive task, inflicting construct-irrelevant factors on the data (Sawaki & Nissan, 2009).

It is therefore necessary to use alternative approaches when determining variables influencing item difficulty. One such method for estimating text difficulty which has been used in reading and writing studies with some reliability is Coh-Metrix (Crossley, Salsbury, & McNamara, 2012), which has not been applied to investigate listening text complexity.

The present study seeks to examine the difficulty of the MET listening test items using two classes of artificial neural networks: a multilayer perceptron neural network and an adaptive neuro-fuzzy inference system (ANFIS)—an artificial neural network model accommodating fuzzy set theory—which may be able to overcome the limitations of methods previously mentioned. Independent variables of the study will be generated through Coh-Metrix and findings of ANFIS modeling will be compared against linear regression and CART to evaluate their effectiveness.

Literature Review

Predictive Data Mining

Predictive modelling is a data mining technique by which various variables are tested for their influence on a future outcome. Data mining itself is a term which is commonly used in computer science and refers to the process of discovering, and summarizing meaningful statistical patterns in large data sets (Geisser, 1971). Some notable predictive models include regression, multivariate adaptive regression splines (MARS), classification and regression tree (CART), neural networks and their extension called adaptive neuro-fuzzy inference systems (ANFIS), as well as Meta-Cognitive learning algorithm for neuro-fuzzy inference system (McFIS) (Aryadoust, 2013a; Subramanian & Suresh, 2012). Most predictive modelling applications involve multiple independent variables or predictors and this can result in multicollinearity—significantly high correlations of independent variables. Therefore, it is important to examine the correlation among independent variables before subjecting data to predictive modelling.

We discuss three primary predictive models including linear regression, CART, and neural networks (including ANFIS) below.

Linear Regression Models

A linear regression analysis models the dependent variable as a function of one or more independent variables. For example, a perfect linear relationship between the values of Y (or dependent/response variable) and X (or independent/predicting variable) can be viewed as follows: $Y = \beta_0 + \beta_1 X$, where β_0 is the intercept parameter and β_1 is the weight parameter of X. The equation can determine the amount of Y given the amount of X, but there is nevertheless some uncertainty regarding the magnitudes of β_0 and β_1, which results in residuals—the differences between the predicted and actual values of Y. Residuals are used to evaluate the fit of the regression model to the data.

Several methods have been proposed to determine the regression models that would best fit data sets. One such method is called Least Squares which minimizes the sum of the squares of the residuals of the equation. If the mapping between dependent and independent variables is nonlinear, the magnitude of residuals increases, thereby rendering the regression model a poor fit to the data. To achieve optimal fit and high precision, nonlinear data analysis techniques such as CART and artificial neural networks have been recommended (Breiman, Friedman, Olshen, & Stone, 1984).

Classification and Regression Trees (CART)

CART modelling is a nonparametric tree-building technique which possesses several important advantages over linear regression: it makes no distributional assumptions about the data; it manages missing data, outliers, multicollinearities, and heteroskedasticity well (e.g., outliers are allocated an independent node and collinear independent variables are used in surrogate splits, thereby having no effect on the model); it can identify interactions among independent variables; and it can reduce high dimensional data (data comprising many independent variables) into a few useful variables.

CART comprises a number of forward growing and backward pruning processes resulting in multiple predictive models or progressively less complicated trees (Breiman et al., 1984). If the dependent variable in the analysis is categorical, CART will give a *classification* tree and if it is continuous, CART will generate a *regression* tree. Classification will help the researcher predict the class of dependent variable data by using the independent variables. That is, CART helps uncover the independent variables responsible for a certain phenomenon or dependent variable. So, the goal is to determine the class in which the dependent variable data would fall.

Classification trees are grown according to the "left-first" rule where CART algorithms initially split data on the left side branch into two nodes and then move to the right side (Steinberg & Colla, 1995). After growing all branches, the largest tree is subsequently pruned in a backward-moving process called "cost-complexity pruning." In this process, the splits which do not optimize the fit of the model (i.e., redundant splits) will be regressively eliminated till the most prudent model is yielded (Steinberg & Colla, 1995). The optimal tree is chosen according to the fit statistics and precision of its parameter estimates as well as the substantive theory available to the researcher (Yohannes & Webb, 1998).

Splitting Nodes and Improvement

CART initially determines a variable and a value in the variable on which basis to split the data set. Once the value or threshold is determined, a rule is made: any data point below or equal to the value will go to the left and data points greater than the value will go to the right node. Splitting the data on the tree disaggregates the sample into two subsamples at each node. Since there might be numerous independent variables, numerous splits would be potentially viable. To rule out the poorly fitting (potential) splits, CART generates and tests all possible disaggregations and chooses the best correlates that split the data optimally (Steinberg & Colla, 1995). For each splitter at each variable, CART estimates a goodness-of-split measure which is called improvement. This processes is repeated for all independent variables, the splitters are ranked in a descending order and the best splitter is chosen. The predicting variables that have close improvement indices at each node are called competitor. If two or more best fitting competitors are identified, the tree is pruned by either deleting the one or more competitors or by forcing the second or other best competitors into the tree as the initial splitter. Finally, if two variables contain highly similar information, one of them is chosen as the primary splitter; the other variable which can be equally important in terms of information is called surrogate variable (Breiman et al., 1984).

Variable Importance Index in CART

As earlier noted, CART is a nonparametric model and accordingly is not based upon the commonly used concepts of statistical significance. To examine the effect of independent variables, CART yields an Importance Index which indicates the contribution of that variable to the dependent variable in a particular tree.

To estimate the Importance Index of each variable, CART uses the improvement index for each variable in its capacity as a primary or a surrogate variable. The magnitudes of these improvement measures at each node are totaled and scaled. The variable with the highest improvement measures is scored 100, and other variables will possess relatively lower Importance Indices. The important index of zero in CART indicates that the input variables never appeared as primary or surrogate variables in the model and therefore make no significant contribution to the tree (Steinberg, Colla, & Martin, 1998). It is also important to note that only the first competitor is awarded credit; the competitors which are below the best competitor will receive no credit or the Importance Index of zero, unless they are surrogate variables.

Another index for assessing the performance of CART models is relative cost or the proportion of misclassifications. The relative cost index ranges between zero and one with values closer to zero indicating better separation of the classes of the dependent variable (Steinberg et al., 1998).

Testing and Cross-Validation

Like linear regression, CART generates R^2 (R-Squared) statistics which is calculated as 1 – relative error. However, this measure often overestimates the goodness of fit. Accordingly, CART analysts have proposed testing and cross validation. CART partitions the data into training (learning) and testing subsamples; estimates the best model for the training data; and fits the yielded model into the testing data to estimate the efficacy of the solution in new data. Training helps estimate the optimum prediction power of independent variables and testing verifies the prediction power (Breiman, 2001).

CART starts by dividing the entire data set into k folds (e.g., k = 10, as proposed by Breiman et al.,

1984). When there is insufficient data for testing, the training data is randomly k-folded and the estimated model is fitted into all folds. The partitions include an even distribution of the dependent variable and have approximately equal size. Subsequently, k models are constructed, each comprising k-1 data partitions for training and only one partition for testing (Breiman, 1994).

CART is a useful method when the researcher wishes to derive a set of simple correlates in a large data set. Some CART computer packages apply "stopping rules" by which the process of model building is terminated at certain points (Schaffer, 1993). Although this would produce simpler models, it could prematurely terminate the algorithm before arriving at an optimal solution. In the present study, the CART algorithm which is used does not use a "stopping rule" so as to preclude premature convergence of the data analysis (Wolpert, 1992).

Despite these advantages, CART is limited by the features of the training data; that is that nonsignificant changes made to the training data can exert significant influences on the model. In addition, CART is a nonprobabilistic model with no confidence interval associated with predicted values derived from CART; therefore, the accuracy of the results yielded is based around the prediction power of the model in other "circumstances" through k-fold cross-validation (Seyoum, Richardson, Webb, Riely, & Yohannes, 1995).

Artificial Neural Networks (ANNs)

ANNs are mathematical nonparametric models comprising an interconnected set of processing units called "neurons," which are adaptive and trainable and contain experiential knowledge. Like the brain, ANNs consist of interconnected units or neurons which are capable of pattern recognition, prediction, classification, and learning. The networks acquire knowledge in data and store the knowledge in a system of neuron connection strengths called synaptic strengths or weights (Barbour, Brunel, Hakim, & Nadal, 2007). Relative to the conventional statistical models of prediction and classification, ANNs have several important advantages: They are highly adaptive and impose no assumption on the relationships between dependent and independent variables such as normality, linearity, homoscedasticity (homogeneity of variance), and error independence which are preconditions of, for example, multiple linear regression. Therefore, if the relationship between variables is linear, ANNs learn the linear structure and approximate linear regression and if the relationship is nonlinear, ANNs would seek the best nonlinear structure fitting the data (IBM, 2012).

Verlinden, Duflou, Collin, Cattrysse (2008, p. 407) stated that, "The biggest advantage of neural networks is the fact that they can approximate functions very well without explaining them. This means that an output is generated based on different input signals and by training those networks, accurate estimates can be generated." Mathematical functions such as multilayer perceptron (MLP) and radial basis function (RBF) are used to predict output or dependent variables in ANNs with minimum error by using input or independent variables. MLP is a simple ANN with three distinct layers: input, hidden, and output, each comprising several neurons with mathematical activation functions such as hyperbolic tangent and logistic functions.

Figure 1 presents an ANN with three inputs notated as $X_{1\text{-}3}$, two neurons in the hidden layer notated as $H_{(1:1)}$ and $H_{(1:2)}$, and two outputs notated as Y_1 and Y_2. Each layer also has an activation function which is a mathematical expression of the amount of output on the basis of the input data (e.g., sigmoid function

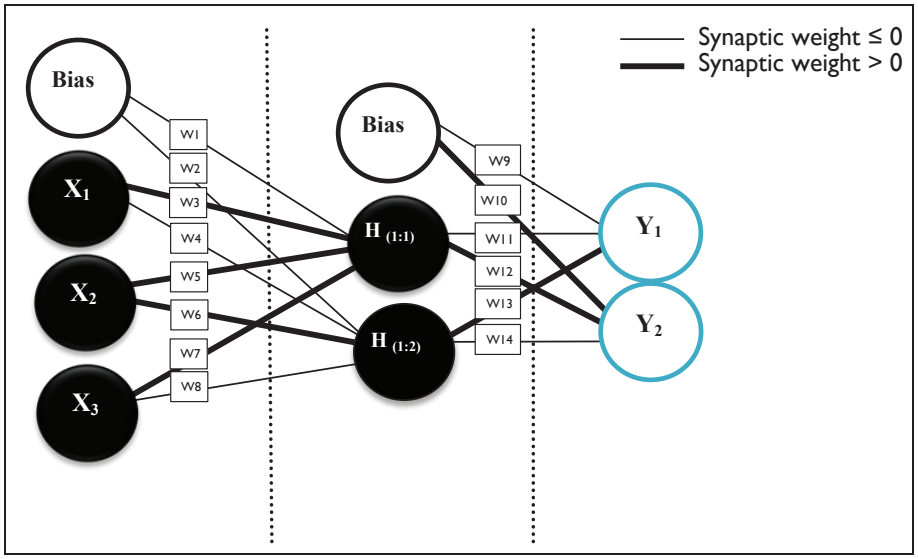


Figure 1: An artificial neural network with three layers, three inputs (X_1, X_2, & X_3), two bias terms, and two outputs (Y_1 & Y_2). The weights are represented as W_i.

mathematically expressed as $f(x)=\frac{1}{(1+e^{-x})}$). The input layer is connected to the hidden layer where the mathematical processing of the data is performed via a network of *weighted* connections. The most important factor determining the type of mathematical functions in the ANNs is the weights.

The training technique we used in this study is called backpropagation which uses various learning rules to learn the patterns in the data. To apply the learning rule, a number of iterations are performed while new data patterns enter the network. When new patterns enter, the network will attempt to randomly estimate the amount of output. The initial estimates are typically imprecise and different from the actual output. Accordingly, the packpropagation technique is used to move backward in the network and reduce the amount of discrepancy by adjusting the weights, presented as $W_{1\text{-}14}$ in Figure 1. The process of adjustment continues till the network reaches the least amount of discrepancy between the estimated and actual outputs. The weights in an ANN are analogous to β coefficients in linear regression models.

The network in Figure 1 also has two bias neurons which help the network to learn the underlying patterns of the data more efficiently and estimate the output accurately. Bias can be viewed as analogous to error of measurement in linear regression modeling.

Adaptive neuro-fuzzy inference system (ANFIS) is an extension of ANNs which integrates them and fuzzy set theory (Landín, Rowe, & York, 2009). Since the ANNs have been previously discussed, fuzzy set theory is explained further below.

Fuzzy set theory (LotfiZadeh, 1965) provides a means of representing relationships which are imprecise or "fuzzy" in ANFIS modelling. Membership in a fuzzy set is determined with a set of conditional statements, including IF-clauses and THEN-clauses. Magnitude of error is estimated by goodness-of-fit indices such as coefficient of efficiency and root mean squared error.

Figure 2 presents a fuzzy inference system with one input, and two trapezoidal "low" and "high" membership functions created by a neuro-fuzzy model. The first step in neuro-fuzzy modeling is "fuzzification," in which the input, the continuous variable X_1, enters the system and associates with two subsets, A (low) and B (high). X_1 (value = 14) is fuzzified as 0.80 and 0.40: that is, $\mu_1(X_1 = A_{[low]}) = 0.80$; and $\mu_2(X_1 = B_{[high]}) = 0.40$ (Lotfi Zadeh, 1965). The second step assesses the defined fuzzy rules by applying the fuzzified input to the rules' antecedents. The rules take the following formats:

$$\text{Rule}_1\text{: IF } X_1 = A_{[\text{low}]}\text{, THEN } Y = 2. \quad (1)$$

$$\text{Rule}_2\text{: IF } X_1 = B_{[\text{high}]}\text{, THEN } Y = 5. \quad (2)$$

(Y is the output and could take any value depending on the range of the data to be predicted). Suppose the model has two inputs, $X_1 = 14$ and $X_2 = 18$. X_2 would also have (at least) two membership functions—low and

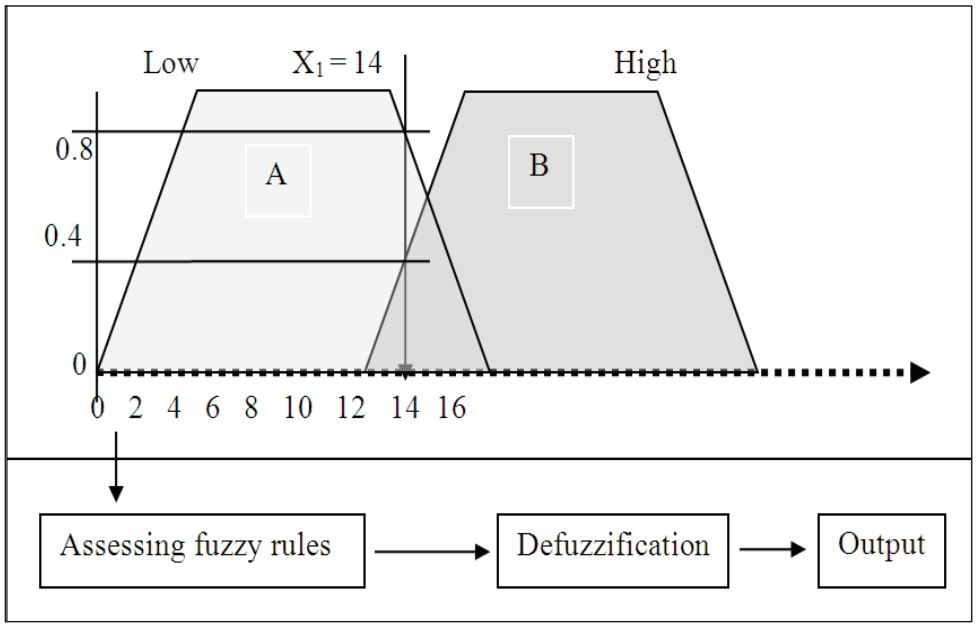


Figure 2: Illustration of a simple neuro-fuzzy system

high—and the rules of the joint functions of the two inputs might be rewritten as:

$$\text{Rule}_1\text{: IF } X_1 = A_{[low]} \text{ AND } X_2 = A_{[low]}, \text{ THEN } Y_1 = 2. \quad (3)$$

$$\text{Rule}_3\text{: IF } X_1 = A_{[low]}, \text{ AND } X_2 = B_{[high]}, \text{ THEN } Y_2 = 3. \quad (4)$$

$$\text{Rule}_2\text{: IF } X_1 = B_{[high]}, \text{ AND } X_2 = A_{[high]}, \text{ THEN } Y_3 = 4. \quad (5)$$

$$\text{Rule}_2\text{: IF } X_1 = B_{[high]}, \text{ AND } X_2 = B_{[high]}, \text{ THEN } Y_4 = 5. \quad (6)$$

Suppose that on the fuzzy functions, the low and high values of X_2 take the values 0.3 and 0.6. The rules are then evaluated as follows: Rule_1: $\mu_1 = 0.80 \times 0.30 = 0.024$, THEN $Y_1 = 2$; and so forth. The rules are then defuzzified and the value of Y is estimated as follows: $Y = (\mu_1 \times Y_1) + (\mu_2 \times Y_2) + (\mu_3 \times Y_3) + (\mu_4 \times Y_4)$ [31].

Listening and Reading Studies

Linear Regression in TOEFL Research

Language assessment research has largely focused on reading test item difficulty and listening studies have relied heavily on reading research (Aryadoust, 2013a; Buck, 2001). Research into listening and reading item difficulty with linear regression models might be categorized into two major eras: pre- and post-1990s.[1] In the former period, several researchers such as Davey (1988), Green (1984), and Drum, Calfee, and Cook (1981) modeled reading test item difficulty as a function of, for example, surface structure features, length of the passage, along with several item-related variables such as the number of multisyllabic words and stem and option length. Embretson and Wetzel (1987) found that connective propositions alongside some of the variables in Drum et al.'s study predicted the difficulty of reading test items.

In the 1990s, English Testing Service (ETS) researchers started to research the variance in reading and listening item difficulty of the TOEFL and reading sections of Scholastic Aptitude test (SAT) and Graduate Record Exam (GRE). They found that item difficulty is primarily attributed to text features (for example, lexical density, number of referentials, sentence length, and negation), item features (for example, stem and distractors' length), and their interactions represented (Freedle & Kostin, 1991, 1992, 1993a, 1993b, 1996; Nissan, DeVincenzi, & Tang, 1996).

In a series of important publications, Freedle and Kostin studied reading item difficulty in SAT, GRE, and the TOEFL. Using stepwise and hierarchical regression models, Freedle and Kostin (1991) examined SAT reading item difficulty and found that test takers' ability to identify main ideas, to make inferences, and to understand explicit statements would account for a part of the variance of SAT reading item difficulty. They further found that text features, item attributes, and their interactions were influential variables, thereby providing evidence that multiple-choice questions tap a construct similar to reading under nontest conditions. Freedle and Kostin's (1992) study of GRE reading item difficulty yielded similar results.

In two consecutive studies, Freedle and Kostin (1993a, 1993b) found that multiple variables accounted for some portion of the variance of the TOEFL reading section. For example, Freedle and Kostin (1993a) reported that variables such as the lexical overlap between the correct option and passage, the location of information for items demanding inference-making, and subject matter predicted approximately 32% of the variance in TOEFL reading item difficulty. They also found that "rhetorical organization, sentence length, location of relevant information, and lexical overlap" (p. 21) were important influential variables which emerged in a nested regression model.

For the TOEFL listening test, Freedle and Kostin's (1996, 1999) studies suggest that 12 sentential and discourse level variables such as lexical density, negation, and stem length accounted for 33% of the variance of item difficulty. This is in line with early research conducted by Carpenter and Just (1975) and Grimes (1975) who showed that negations and rhetorical structure would impact prose recall accuracy (see also Meyer & Freedle, 1984) and text comprehension accuracy (see also Hare, Rabinowitz, & Schieble, 1989).

Finally, Nissan et al. (1996) and Kostin (2004) found that three major variables significantly influenced the difficulty of the mini-talks of the TOEFL listening test: (a) negation in the dialogue (at least two negative words); (b) the cognitive process demanded by the item, specifically inference-making beyond explicitly articulated information; and (c) "the pattern of utterances in the dialogue" (Kostin, 2004, p. 27). Drawing on Adams, Carson, and Cureton, (1993), Kostin argued that item difficulty can be adjusted by manipulating these variables.

Linear regression models have achieved varying degrees of success in the aforementioned studies, though the highest amount of variance explained by item difficulty has hardly exceeded 50%. This might be due

1 The latter period is marked primarily by ETS (English Testing Service) research.

to the judgment of experts who were tasked to identify the influential variables (see Alderson & Kremmel, 2013, for the effect of expert judgment) or the linear model which is affected by nonlinear patterns in the data. To ascertain the reliability of regression model output and control for the effect of rater judgments, regression-based Coh-Metrix research has recently emerged where linear regression modelling is used innovatively. This is reviewed in the next section followed by reviews of other research studies.

Regression-Based Coh-Metrix Research

Coh-Metrix is a free analytic tool for measuring psycholinguistic features of texts. Coh-Metrix has been used in several reading, first and second language (L1 & L2) writing, and speaking studies and achieved relative success (Crossley & McNamara, 2010, 2011; Crossley & Salsbury, 2010; Crossley, Salsbury, McNamara, & Jarvis, 2011). In this section, we survey the Coh-Metrix reading research, as it is more closely related to listening research, both being comprehension skills.

Crossley, Greenfield, and McNamara (2008) examined the association of L2 text readability with lexical frequency, syntactic similarity, and content word overlap which are related to three L2 reading comprehension processes: decoding, syntactic parsing, and meaning construction, respectively. Multiple regression modelling gave an R2 value of 0.86, meaning that 86% of variance in L2 text readability was explained by these variables. Relative to the traditional readability formulas such as Flesch Reading Ease and Flesch-Kincaid Grade Level, Coh-Metrix readability index achieved more accuracy in predicting reading difficulty.

Similarly, Crossley, Allen, and McNamara (2012) compared the precision of traditional readability and Coh-Metrix readability indices in classifying the texts into beginner, intermediate, and advanced levels. The Coh-Metrix readability index outperformed the other two indices in classifying texts into those three levels. The researchers argued that the Coh-Metrix index took into account factors related to text comprehensibility including cohesion and meaning construction, along with cognitive-processing indices such as decoding and syntactic parsing.

In a different study, Crossley, Louwerse, McCarthy, and McNamara (2007) compared the syntactic, rhetorical, and lexical features of simplified and authentic texts. They found that authentic texts were marked by diversity in parts of speech, larger numbers of logical operators (for example, if, then, & but), and causality, whereas simplified texts were lexically less diverse but syntactically more complex. Both groups of texts contained equal levels of abstractness, although simplified texts contained more referential cohesion (i.e., when a pronoun such as *it* or *he* refers to another word which has been mentioned in the text) and common connectives (for example, and & or).

Crossley and McNamara (2008) replicated Crossley et al.'s (2007) study by using a larger corpus of simplified and authentic reading texts. Their findings were consistent with Crossley et al.'s (2007) results, indicating that Coh-Metrix would be an efficient alternative to traditional readability measures.

The aforementioned studies have applied regression models in an innovative way. They split the data into training and testing samples, estimate a regression model for the training sample, and use its intercept and beta coefficients to predict the dependent variable (which is writing or reading scores) in the testing sample. However, the researchers often deleted highly correlated predictors, arguing that they would cause multicollinearity (i.e., high linear association between predictors or independent variables), which can alter the structure of the postulated models.

Cognitive Diagnostic Assessment

To address the limitation of linear regression, another group of researchers have used cognitive diagnostic assessment models (Aryadoust, 2012; Lee & Sawaki, 2009). In one of the earliest studies, Buck, Tatsuoka, and Kostin (1997) used rule-space methodology (Tatsuoka, 1983, 2009) to determine the influential variables in a reading test. They found that item features such as syntactic relations, discourse structure, and vocabulary difficulty were significant features influencing students' performance on reading tests.

Buck and Tatsuoka's (1998) application of rule-space methodology to a listening test identified 15 primary attributes (for example, the ability to identify the task, to process medium/low amounts of information, and to use background knowledge) and 14 interactions which could classify 96% of students successfully. Although they used no vocabulary and syntactic complexity measures in the study, they suggested that these measures be used in future research specifically "some general index of complexity, which would somewhat take into account the cumulative effects of all the most important characteristics which make up syntactic [and lexical] complexity" (Buck & Tatsuoka, 1998, p. 141). Buck

and Tatsuoka's speculation concerning the impact of syntactic and lexical features of the listening test items was supported in Aryadoust's (2012) application of the fusion model to section four of the IELTS listening test. Aryadoust found that the linguistic features of items such as grammar and vocabulary alongside the ability to make paraphrases, understand specific information, and "integrate listening and reading in short-term memory" influenced test performance.

Using the fusion model, Sawaki, Kim, and Gentile (2009) examined the attributes tested by the listening and reading sections of the Internet-Based TOEFL (TOEFL iBT). They found that four major attributes classified reading test takers successfully, including the ability to: (a) understand "word meaning"; (b) understand "specific" and key information; (c) connect information; and (d) synthesize and organize information (Sawaki et al., 2009, p. 199). Similarly, the listening was influenced by the ability to: (a) understand "general information"; (b) understand details; (c) understand "text structure" and intention of the speaker; and (d) link ideas (p. 203). Lee and Sawaki (2009) also examined the listening and reading sections of the TOEFL iBT, comparing latent class analysis, general diagnostic model, and the fusion model. The three methods performed equally well, yielding similar results to Sawaki et al.'s study.

In another fusion model study, Jang (2009) reported two primary groups of attributes influencing reading test takers' performance text- and test-related. These comprised nine major attributes including context-dependent and context-independent words; semantic and syntactic connections, negation, understanding textually explicit and/or explicit information, inference-making, summarizing, and "mapping contrasting ideas into framework" (Jang, 2009, p. 231). Jang stated that "A reading comprehension assessment that is designed to elicit such process-oriented skills can provide more authentic accounts of readers' competencies in L2 reading comprehension" (p. 232).

Application of CART in Language Assessment

CART or similar tree-based methods have achieved relative success in language and educational assessment. Gao and Rogers (2011) tested the validity of a cognitive model for reading assessment and identified multiple variables influencing reading test item difficulty. They reported that the "plausibility of distractors" was the most significant variable predicting item difficulty, rendering the test-taking process a "problem-solving process" influenced significantly by "verbal reasoning abilities" (Gao & Rogers, 2011, p. 97). Whereas the number of plausible distractors can alter the difficulty level of test items, it might also invite the execution of cognitive processes which are different from real-life reading comprehension processes, thereby affecting the cognitive validity of the test (Field, 2009). Gao and Rogers summarized other findings of their study as follows:

> An item bearing the following features would likely be an easy item: it does not have plausible distractors and requires basic syntactic knowledge. In contrast, an item bearing the following features would likely be more difficult: it has more than one plausible distractor, requires recognizing the meaning of unknown words using context clues, and requires information located in the entire passage. (Gao & Rogers, 2011, p. 99)

Gao and Rogers's (2011) study has two primary limitations. First, although they used two test sample tests for analysis, they did not partition the data into training and testing samples nor did they apply cross-validation, yielding overfitting solutions (Breiman et al., 1984)—a limitation which was acknowledged by the authors. (Overfitting occurs when the CART or other statistical tools model measurement errors in lieu of the underlying structure of the data). This issue also resulted in different trees in the two tests. Second, the sample of test items chosen was fairly small, which would inflate the fit of the model.

In another study, Sheehan and Ginther (2001) applied a tree-based regression approach to explore the variables that affect the reading test item difficulty of the Test of English as a Foreign Language (TOEFL). The study revealed that the cognitive processes engaged by the items such as understanding the main idea predicted 87% of the observed variance in item difficulty indices. They operationalized main idea as (a) correspondence between the passage and correct options; (b) location of important information in the passage; and (c) the length of the passage that test takers should process to respond to the item. This finding is partially in line with Sheehan's (1997) tree-based regression study where several variables accounted for reading test item difficulty, including the cognitive processes engaged by items such as understanding implicitly or explicitly articulated information, vocabulary, and linguistic features of the items' options. It is also consistent with

Table 1: Demographic Information of the Listening Tests

	No. of Items	Age Mean Score	Sample Size	Gender Distribution	
				M	*F*
Form 1	46	22.09	963	457	506
Form 2	46	27.71	612	302	310
Form 3	46	26.85	564	235	329
Form 4	46	22.96	758	336	422
Form 5	46	23.50	608	253	355
Form 6	46	20.87	708	347	361
Form 7	46	23.62	826	348	487
Total	322	23.94[a]	5039	2278	2770

Note. [a] average age

Huff's (2003) study of reading and listening sections of the TOEFL. Huff reported that features of and the interactions between items' stems and (reading and oral) passages accounted for 56% and 48% of the variance in reading and listening item difficulty, respectively.

Finally, Rupp et al. (2001) applied both linear and tree-based regression models to examine the effect of the required cognitive processes induced by passage feature such as information density or length of sentences, item features such as the length of the options, and item by passage features on difficulty of reading and listening comprehension items. Their linear regression model showed that passage and interaction features accounted for item difficulty and the tree-based regression model revealed more details about these effects. However, they speculated that the study might have been affected by the modalities of the data since reading and listening items were examined jointly.

Application of Neural Networks in Language Assessment

The application of neural network in language and reading studies is critically underresearched. To our knowledge, only two studies have used these models. The first study was conducted by Perkins et al. (1995, p. 34) who used "a three-layer backpropagation" neural network to predict item difficulty in 29 TOEFL reading comprehension items. Perkins and colleagues split the sample into training (15) and testing (14 items) subsamples and tested a reading model containing "text structure, propositional analysis of passages and stems, and cognitive demand" (p. 39) by using a sigmoid function (a mathematical function having an S-curve). They achieved significantly high correlation (> 0.90) between actual item difficulty and the attributes after optimizing the relationship between items and attributes.

More recently, Aryadoust (2013a) applied a neuro-fuzzy inference system (ANFIS), an extension of neural networks, to a 40-item listening test. He found that "word frequency," item and information type, density of prepositional phrases, modal verbs, and propositional density of oral texts and items ("the number of independent units conveying discrete messages within each text" [p. 45]) predicted item difficulty. Compared with ANFIS, the path model yielded a less accurate model, though it accommodated the interaction between independent variables. He concluded that "The results of the ANFIS model seem to be more intuitive and theory-informed, which is a significant advantage" (p. 48).

Despite their contribution to listening and reading studies, Aryadoust's (2013a) and Perkins et al.'s (1995) studies used small samples. The present study aims to use a larger sample of listening test items to examine the effect of psycholinguistic features on listening test item difficulty.

Methodology

Data Source and Materials

The data and materials required for this study were provided by CaMLA and include the item-level data of students performing on seven independent MET listening tests. The test takers were from South American countries including Columbia, Costa Rica, Peru, Brazil, and Chile. The demographic information of the test takers is presented in Table 1. Each test form consists of 46 test items, giving a sample of 322 multiple choice test items (7 tests × 46 items), which will yield stable solutions in data mining analyses (Hair, Black, Babin, Anderson, & Tatham, 2010). Form 1 has the largest sample size (n = 963) and Form 3 the smallest sample

size (n = 564). Overall, the number of test takers was 5039, which will yield stable item Rasch difficulty parameters (Bond & Fox, 2007).

CaMLA also provided the test materials including test items and audio materials, which were transcribed and subjected to Coh-Metrix analysis. Each test comprises three sections, as follows:

(a) Part one comprises 17 short conversations between a man and a woman. Each conversation is followed by a test item with four options.

(b) Part two comprises four lengthy conversations between a man and a woman. Each conversation is followed by three or four comprehension test items with four options.

(c) Part three comprises three mini-talks on academic topics, each followed by three or four comprehension test items with four options.

Generating Independent Variables

To create independent variables, we used measures of semantic, lexical, and syntactic complexity for each test items as computed by Coh-Metrix. We also attempted to estimate traditional statistics such as the average length of t-units—the shortest grammatically accurate sentence. However, analyzing several test items, we found that the number of t-units were equal to the number of sentences counted by Coh-Metrix in most cases. Therefore, we took the number of sentences as a proxy for t-units.

Next, we identified the "necessary information" (NI) to answer the listening test items (Buck & Tatsuoka, 1998). For each item, we merged the NI, item stem, and distractor texts and performed Coh-Metrix analysis. Although it would have been desirable to examine each component (NI, item stem, and distractor texts) independently, the short length of the components would preclude us from estimating reliable Coh-Metrix statistics. The advantage of this combination lies in the length of the yielded text and the reliability of Coh-Metrix indices. However, one limitation of this approach would be the mix of written and oral modalities, likely affecting the precision of the results. This limitation has also affected previous research such as Rupp et al.'s (2001) study of reading and listening tests.

After estimating the Coh-Metrix statistics for all test items, we chose the best input variables (explanatory correlates) among the statistics. Because too little is known about the use of Coh-Metrix in listening comprehension studies, choosing the optimal variables would be extremely difficult. We chose the Coh-Metrix variables similar to the influential variables emerging from the previous research such as lexical diversity, situation model, syntactic pattern density, syntactic complexity, text easability, and word information. Following previous researchers (see Crossley & Salsbury, 2010), we correlated the chosen variables with item difficulty parameters, choosing 12 variables with significant correlations with the output variable ($p < 0.05$), including word count, text easability (PC temporality), content word overlap, given-new sentences' average, type-token ratio (content word lemmas or labels), logical connectives, causal verbs and particles, left embeddedness (words before main verb), preposition phrase density, verb incidence, hypernymy for nouns and verbs, and Flesch-Kincaid grade level. These variables are presented in Table 2. For example, the lexical diversity category is measured by the type-token ratio or the ratio of unique words (types) to the total number of words (tokens). A low type-token ratio would suggest low cohesion or short length of the text (McNamara, Louwerse, Cai, & Graesser, 2013).

Finally, the independent variables were discretized using the software *Discretize.exe*, since the Coh-Metrix variables have a wide spread, likely yielding less accurate results (Lui, Hussain, Chew, & Dash, 2002). The discretized variables comprised three to five levels.

Data Analysis

This study uses multiple primary data analysis techniques: (a) initially, Rasch item difficulty indices were estimated. The measures were discretized by using a median split technique where items were categorized into low- and high-difficulty items; (b) next, as noted earlier, independent variables were generated by estimating the semantic and syntactic complexity measures of test items and texts through Coh-Metrix methods; (c) theoretical correlates of the item difficulty measures were determined; and (d) finally, the data were subjected to logit (logistic) regression model, CART, ANN (including perceptron and ANFIS) to determine the predictive independent variables.

The Rasch Model

Test items were subjected to the Rasch model and their difficulty measures—which constitute the dependent variable of the study—were estimated.

Fit statistics. Rasch model infit and outfit mean square (MNSQ) statistics were also estimated. (Bond

Table 2: Coh-Metrix Theoretical Correlates of Item Difficulty Parameters

Variable	Category	Remarks
Word count	Descriptive	Number of words in the text. Lengthy texts can tax listeners' cognitive resources and working memory and be difficult to process.
Temporality	Text easability principal component scores	Research shows that temporal features such as tense can facilitate comprehension (Duran, McCarthy, Graesser & McNamara, 2007). In writing research, Crossley and McNamara (2010, p. 17) found that "writers judged as highly proficient provide readers with less temporal cohesion and word overlap."
Content word overlap (adjacent sentences)	Referential cohesion	It measures the overlap of content words in two adjacent sentences. The overlap tends to facilitate the text and comprehension (Kintsch & Van Dijk, 1978).
Given-new sentences' average	Latent semantic analysis (LSA)	This is a measure of semantic overlap between sentences and was chosen because the item stems, options, and necessary information had some semantic overlap in the present study (see Hempelmann, Dufty, McCarthy, Graesser, Cai, & McNamara, 2005; McCarthy, Dufty, Hempelman, Cai, Graesser, & McNamara, 2012).
Type-token ratio (content word lemmas)	Lexical diversity	Crossley, Allen, and McNamara (2012) found that higher type-token ratios tend to increase the difficulty of the test item.
Logical connectives incidence	Connectives	Connectives create coherence in texts and facilitate comprehension (McNamara et al., 2010).
Incidence of causal verbs and causal particles	Situation model	Causal verbs and particles convey agenthood and cause-effect relationships (e.g., affect & because) (McNamara, Ozuru, Graesser, & Louwerse, 2006).
Words before main verb mean (left embeddedness)	Syntactic complexity	Crossley et al. (2012), Graesser, Cai, Louwerse, and Daniel (2006), and Just and Carpenter (1992) found that texts with a large number of words before the main verb are more difficult. More recently, Aryadoust, Mehraban, and Alizadeh (2014) used an MLP ANN and verified the influence of words before the main verb.
Preposition phrase density	Syntactic pattern density	The number of phrases starting with a preposition indicates the syntactic density of the text. Syntactically dense texts tend to be more difficult to parse and comprehend (Crossley et al., 2012).
Verb incidence	Word information	Verbs convey important information and parsing them successfully helps listeners achieve comprehension.
Noun and verb hypernymy	Word information	Hypernymy is calculated based on Miller, Beckwith, Fellbaum, Gross, and Miller's (1990) WordNet and indicates texts' lexical sophistication and word specificity. It had weak but statistically significant correlation with essay grades in Crossley and McNamara's (2010) study, but failed to predict the grades in the linear regression model. By contrast, Crossley, Salsbury, & McNamara (2009) reported significantly high prediction power for hypernymy.
Flesch-Kincaid grade level	Readability	Crossley, Allen, and McNamara (2012) found that Coh-Metrix readability index would outperform Flesch-Kincaid grade level in explaining the difficulty level of texts, although the latter also achieved significant accuracy.

& Fox, 2007). Infit MNSQ is an index sensitive to the perturbations of inliers or the responses targeted on the test takers and outfit is a weighted fit index sensitive to the data patterns far from test takers' ability. According to Bond and Fox (2007), an item is underfit if its fit MNSQ index is greater than 1.4 and overfit if fit MNSQ values are below 0.6. Aryadoust, Goh, and Lee (2011) proposed a more stringent fit criterion for multiple choice questions that treats items with fit statistics falling outside of range between 0.8 and 1.20 as misfit. In this study, we apply Aryadoust et al.'s criterion in order to improve the precision of the measurement (Baghaei & Amrahi, 2011).

Item and person reliability. We estimated Rasch model item and person reliability (true variance / observed variance) and separation coefficients. The person reliability index ranges between zero and 1.00 and indicates the sensitivity of the test to distinguish among high- and low-ability test takers, hence the precision of the measurement for the test takers. The item reliability index also ranges between zero and 1.00 and indicates the sufficiency of the sample size. Reliability is also expressed as separation (true standard deviation / root mean square error of measurement) which is the number of statistically distinct levels of test item difficulty or test taker ability. For example, item separation index of three indicates three statistically distinguishable strata of items.

Z-scores and Discretization. We initially prepared the data for linear regression analysis by calculating the z-scores and identifying the outliers which violated the normality assumption (Hair et al., 2010). In this study, we deleted the items whose z-scores would fall outside of |-2 – +2|, retaining 241 test items for analysis.

We then converted the continuous item difficulty measures into a categorical variable. We tested two approaches: initially we developed a three-level categorical variable where the test items were coded as high, medium, and low difficulty. Then, we estimated the median of the item and made two item difficulty levels: values below and equal to the median were put in the "low" difficulty level and values above it were labeled "high." To test the precision of the two variables, we correlated them with the original difficulty measures and found that the median split variable had a higher correlation (0.81) than the three-level variable (0.71). Accordingly, we chose the two-level variable as the dependent variable of the study.

Regression Analysis

We applied a multiple regression model on SPSS, Version 21, with 12 independent variables generated by Coh-Metrix and the discrete item difficulty variable, as follows: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{12} X_{12} + e$, where β_0 is the intercept, $\beta_{1\text{-}12}$ are the coefficients that describe the effect size of the independent variables on item difficulty, $X_{1\text{-}12}$ are the independent variables or the theoretical correlates generated by Coh-Metrix, and e is the error of measurement.[2]

To explore the precision of different regression model-testing methods, we tested four regression methods: (a) Enter: where all independent variables are simultaneously entered into the equation, (b) Remove: where all independent variables are simultaneously removed from a block of variables, (c) Stepwise: where independent variables are chosen based on their p values, and (d) Backward: where all independent variables enter the system and then removed one by one to reach the optimal combination of the variables. For each independent variable, we estimated the p value which indicates how confident we are that each independent variable is associated with the dependent variable (Tabachnick & Fidell, 2013). We further estimated the R^2 index, which is the amount of the variation in the dependent variable explained by (or predicted by) the independent variables. Finally, we estimated the adjusted R^2 index which adjusts the magnitude of the R2 index for the sample size.

To test for multicollinearity, we estimated the variance inflation factors (VIF) for each model, which determines the increase in the variance of the estimated coefficients if the independent variables have no correlations (O'Brien, 2007; Stevens, 2009). If the independent variables are not correlated, the VIF indices will be equal to unity. VIF statistics equal to or below five are treated as the indicator of noncollinearity (Hair et al., 2010). We also used the tolerance index of the independent variables which is computed as one minus the squared multiple correlation of the variable with other independent variables. Small tolerance indices (> 0.2) indicate variable redundancy or a lack of variable's influence on the dependent variable (O'Brien, 2007).

Finally, we computed the d statistic (Durbin–Watson statistic) for each model, which ranges between zero and four and indicates the relationship between the residuals.

2 We recognize that a logistic loglinear regression model could also be a suitable analysis technique (see Ruczinski, Kooperberg, & LeBlanc, 2003). It would certainly be worth testing that model in future studies.

The *d* statistics below two suggest that the independent variables' residuals are positively correlated and the *d* statistics close to four suggest that the independent variables' residuals are negatively correlated. It is desirable that the *d* statistic not significantly deviate from two.

Classification and Regression Trees (CART) Analysis

We performed the CART analysis on the Salford Predictive Modeler® (SPM) software,[3] Version 7.0. We used the 12 independent variables generated by Coh-Metrix and the discrete item difficulty as the dependent variable. We attempted to divide 20% testing and 80% training subsamples, but the program forced a partition of training (n = 202; 83.82%) and testing (n = 39; 16.18%) subsamples. (Conventionally, the majority of the sample is used for training the CART algorithm in order to represent every subgroup in the data.)

Table 3: CART's Data Set Information

Class	Sample	Number	Percentage
1 (*low difficulty*)	Learn	80	39.60%
	Test	16	41.03%
	Total	96	39.83%
2 (*high difficulty*)	Learn	122	60.40%
	Test	23	58.97%
	Total	145	60.17%

To choose the optimal tree, we used a number of quality control indices. We computed the *sensitivity* and *specificity* of the CART models to examine the performance of the models (Jensen, Muller, & Schafer, 2000). Sensitivity or true positive proportion refers to the proportion of true positives which are accurately classified as having that condition—that is,

$$\frac{\text{No of true negatives}}{\text{No of true negatives + No of false positives}}$$

For example, in this study sensitivity is the proportion of high difficulty items which have been correctly classified. Specificity is the proportion of true negatives which have been accurately identified—that is,

$$\frac{\text{No of true positives}}{\text{No of true positives + No of false negatives}}$$

A perfect classification would achieve 100% of sensitivity and specificity, but in reality there is a tradeoff between the two indices, which is represented graphically as a *receiver operating characteristic* (ROC) curve (see Swets, 1996).

ROC curves plot the proportion of true positives (TP) against false positives (FP). TP represents sensitivity, and FP (1 – specificity) represents true negatives. By examining the area under the ROC curve, which ranges from zero to unity, we collected further evidence as to which model was optimal and should be chosen over other models (Zhou, & Qin, 2005). In this study, we adopted the criteria proposed by the Department of Math of the University of Utah (n.d.) to interpret the area under the ROC curve

(A) 0.90 – 1 = excellent
(B) 0.80 – 0.90 = good
(C) 0.70 – 0.80 = fair
(D) 0.60 – 0.70 = poor
(E) 0.50 – 0.60 = fail

We further estimated misclassified 1 (low difficulty) and 2 (high difficulty) cases in the training and testing samples as well as the overall classification correct percentage. Finally, we estimated a normalized Importance Index for each independent variable, which indicates the weight of each independent variable in predicting the dependent variable and ranges between 0.00% and 100%. Higher indices indicate that the variable has a higher contribution to predicting or classifying the dependent variable.

Neural Networks

We tested two classes of neural networks: Multilayer Perceptron (MLP) Artificial Neural Network (ANN) on IMB SPSS Neural Network computer package, Version 21, and Adaptive Neuro-Fuzzy Inference System (ANFIS) on MATLAB, Version 2012b. As previously discussed, ANFIS integrates fuzzy set theory and ANNs to emulate the data. By contrast, the MLP ANN does not impose the fuzzy sets, rendering a more exploratory approach than ANFIS. Both ANFIS and MLP ANN analyses consisted of 12 independent variables training and validation stages. Similar to the CART modelling, we partitioned the data into training (n = 192; 79.70%) and testing (n = 49; 20.30%) subsamples for both ANFIS and perceptron neural networks. (As stated, a larger portion of the sample was used for training the algorithm in order to represent every subgroup in the data). We estimated specificity, sensitivity, the area under

3 We also tested IBM SPSS CART application which resulted in a relatively less accurate model.

Table 4: Item and Person Reliability and Separation Alongside Infit and Outfit MNSQ Indices of the Seven Tests

	Item Reliability	Item Separation	Person Reliability	Person Separation	Average Item Infit MNSQ (SD)	Average Item Outfit MNSQ (SD)
Form 1	0.99	8.45	0.90	3.08	1.00 (0.09)a	0.99 (0.23)
Form 2	0.98	8.09	0.90	3.08	1.00 (0.11)	1.00 (0.23)
Form 3	0.96	5.16	0.90	3.08	1.00 (0.08)	0.99 (0.22)
Form 4	0.98	7.83	0.90	2.38	1.00 (0.11)	0.90 (0.20)
Form 5	0.99	9.05	0.90	3.03	0.99 (0.11)	1.05 (0.30)
Form 6	0.98	6.57	0.89	2.87	1.00 (0.11)	1.00 (0.17)
Form 7	0.98	7.45	0.90	3.07	1.00 (0.13)	1.01 (0.22)

Note: [a] standard deviation of the fit statistics

the ROC curve, Variable Importance Index, and the proportion of accurately classified items for the MLP ANN.

In addition, a number of ANFIS models were initially generated, incorporating between one and 12 hypothesized independent variables. In models comprising between two and 12 independent variables, numerous submodels consisting of all possible variable combinations were assessed to arrive at maximal solutions. To our knowledge, Importance Index or the area under the ROC curve has hardly been reported in ANFIS studies. Since ANFIS modelling is an extremely time-consuming analysis and the computer must run for long hours, we decided to initially identify the optimal ANFIS model based on the fit statistics that ANFIS researchers have proposed and subsequently estimate the area under the ROC curve only if the model fits well. Following Aryadoust (2013a, p. 46), we computed three goodness-of-fit indices for each ANFIS model, as follows:

(a) Squared correlation coefficient (*R2*): A goodness-of-fit index used to explore the fit of the model to the measured data. It ranges between zero and one, with values near one indicating good fit.

(b) Root mean squared error (RMSE): An error measure. Lower RMSE values indicate smaller error terms. RMSE values tend to decrease with larger datasets.

(c) Mean Absolute Error (MAE): A measure of error that computes all deviations from original data, regardless of their signs. (Aryadoust, 2013a, p. 46)

We further computed three fit statistics which have been used in other ANFIS studies:

(d) Correlation between the expected and modeled (predicted) output (R): Values closer to unity indicate good fit.

(e) Normalized mean square error (NMSE): A goodness-of-fit index which shows the difference between the fit of different models. Low NMSE values indicate that the model is performing well.

(f) Mean absolute error (MAE): A measure estimating how close the predicted are to actual output values. Values closer to zero are desirable.

We examined various combinations of mathematical membership functions to determine the optimal solutions including gbellmf (generalized bell-shaped membership function), gaussmf (Gaussian curve membership function), gauss2mf (Gaussian combination membership function), dsigmf (difference between two sigmoidal functions membership function), psigmf (product of two sigmoidal membership functions), and pimf (∏-shaped membership function).

Next, we deleted the fuzzy set functions from the ANFIS and tested a perceptron neural network model. Much of the statistics computed in the CART analysis was also computed for this model, including Importance Index, ROC curve and area under the curve, sensitivity, and specificity. Appendices A1 and A2 present the syntax input for IBM SPSS MLP ANN application and MATLAB ANFIS toolkit, respectively.

Results

The Rasch Model

We subjected the test data to the Rasch model; estimated item difficulty and person ability measures; computed item/person reliability and separation indices as well as infit and outfit MNSQ values.

Table 5: Tolerance, VIF, *t* Values, Significance Level, and β Coefficients for the Enter Regression Model

	Tolerance	VIF	t value	Significance	βCoefficient
(Constant)	NA	NA	**4.138**	0.000	NA
(a) Word count	0.455	2.200	-0.039	0.969	-0.003
(b) Temporality	0.361	2.767	1.808	0.072	0.176
(c) Content word overlap (adjacent sentences)	0.506	1.975	0.624	0.533	0.051
(d) Average givenness	0.276	3.627	-1.559	0.120	-0.174
(e) Type-token ratio	0.699	1.431	-1.471	0.143	-0.103
(f) Logical connectives incidence	0.846	1.182	0.174	0.862	0.011
(g) Incidence of causal verbs	0.900	1.111	-0.068	0.946	-0.004
(h) Words before the main verb	0.821	1.218	2.334	0.020	0.151
(i) Prepositional phrase density	0.834	1.198	1.412	0.159	0.090
(j) Verb incidence	0.803	1.246	1.267	0.206	0.083
(k) Noun and verb hypernymy	0.702	1.424	-2.847	0.005	-0.199
(l) Flesch-Kincaid grade level	0.619	1.617	-0.267	0.790	-0.020
Predicted item difficulty	0.476	2.101	2.609	0.010	0.221

Note: VIF = variance inflation factor

Overall, all items and persons fitted the Rasch model and the test forms all had significantly high item and person reliability. For example, Table 4 shows that Form 5 had the highest item separation index (9.05; reliability = 0.99), and a relatively high person separation index (3.03; reliability = 0.90), hence nine and three statistically distinguishable item difficulty and person ability strata, respectively. Form 5 items fit the model well, with an average item infit and outfit MNSQ of 0.99 and 1.05, respectively. Appendix A1 presents the item person maps of the seven test forms.

Linear Regression Model

Table 5 gives the VIF, Tolerance, *t* values along with their significance level, and the *β* coefficient for the Enter regression model. The variables influencing the dependent variable are words before the main verb and noun and verb hypernymy. The VIF indices of these variables are below five (1.218 and 1.424 < 5) and Tolerance indices are close to unity (0.821 and 0.702 > 0.2), indicating the lack of multicollinearity. The *d* statistic of the model was 2.094, also supporting the lack of multicollinearity.

Of the two influential independent variables, the hypernymy index has a greater impact on item difficulty, as indicated by its *β* coefficient which is -0.199. The negative sign indicates an inverse relationship where the higher the noun and verb hypernymy, the lower the difficulty level. The rest of the independent variables would make no contribution to the item difficulty. The R, R2 and adjusted R2 values of this models were 0.476, 0.227, and 0.183, respectively. That is, the two independent variables can predict 18.3% of the variance in the data.

The regression models built by using the Remove, Stepwise, and Backward methods were identical to the Enter method (as expected), indicating that the optimal linear regression model for this data would include two independent variables regardless of the model: words before the main verb and noun and verb hypernymy.

CART Model

We applied CART to determine the influential independent variables affecting test item difficulty. The optimal model has an R2 index of 0.240, indicating that the independent variables in the model accounted for 24% of the variance in the dependent variables.

The splitting rules of the CART model (training data) are presnted in Figure 3. The topmost node includes temporality, which splits the data into two child nodes: the left node includes 54 test items (class 1 = 36; class 2 = 16) with a temporality index equal to or below 0.50 and the right node includes 148 items (class 42 = 36; class 2 = 106) with a temporality greater than 0.50. The data in these nodes are further partitioned downward by the other independent variables.

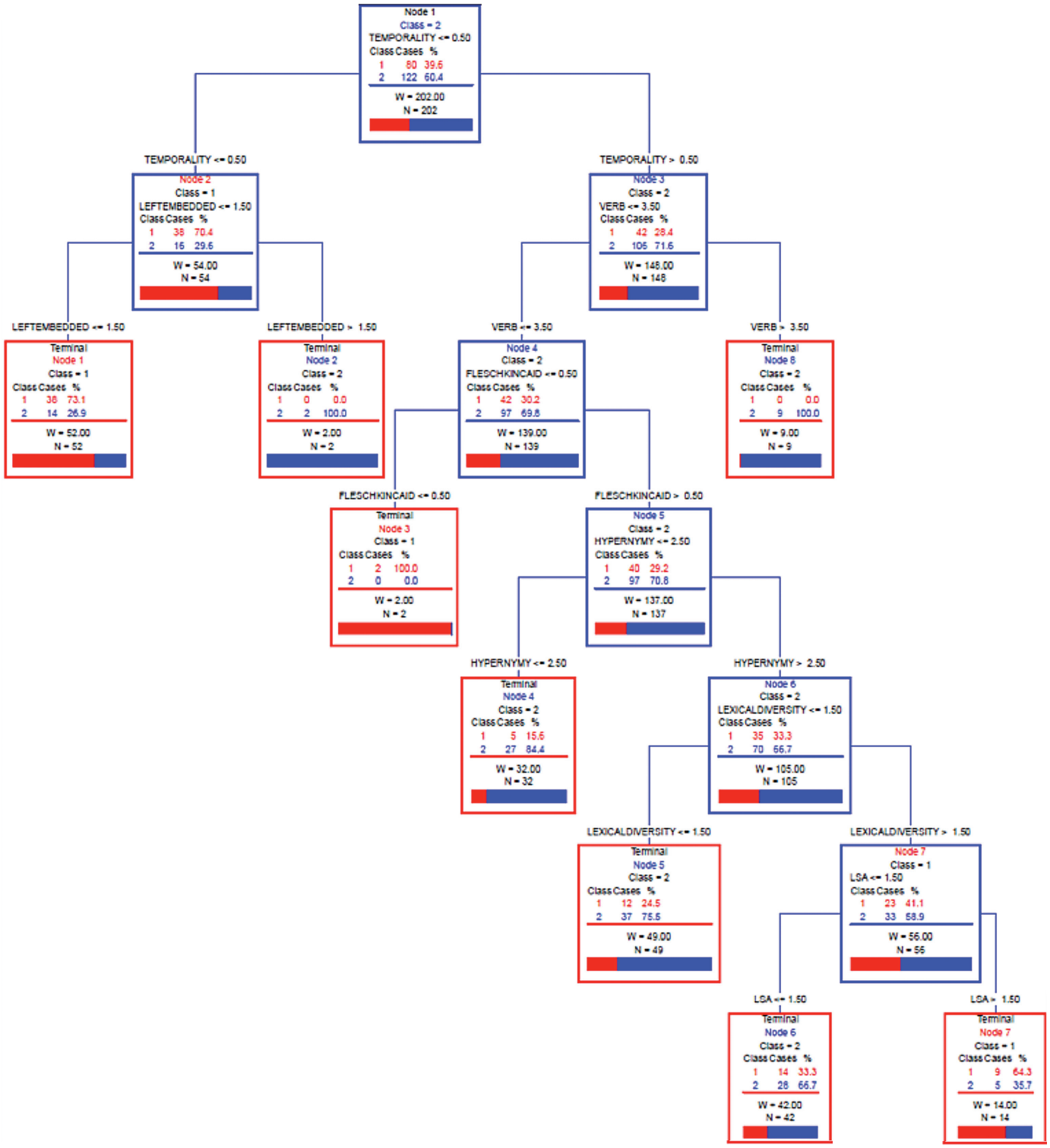


Figure 3: Classification and regression tree (CART) model of the training data ($R^2 = 240$).

Legend: Verb = Verb incidence. LEFTEBMEDDED = Words before the main verb.

Table 6: Classification Accuracy, Specificity, Sensitivity, and Relative Cost Estimated by the CART Model

	Class	No. of Cases	Accurately Classified Cases	Percentage	Relative Cost	Specificity	Sensitivity	ROC
Learning	1[a]	80	63	78.75%	0.155			
	2[b]	122	75	61.48%	0.387			
	Overall	202	138	68.32%		78.75%	61.48%	0.78
Testing	1	16	8	50%	0.50			
	2	23	17	73.91%	0.26			
	Overall	39	25	64.10%		50.00%	73.91%	0.60

Note: [a] low-difficulty items
[b] high-difficulty items

Table 6 presents the classification results in the learning and test subsamples. Overall, the accuracy of classification of class 1 was higher for the training subsample (n = 63 out of 80 or 78.75%) than the testing subsample (n = 8 out of 16 or 50%). By contrast, the accuracy of classification of class 2 was higher for the testing subsample (n = 17 out of 23 or 78.75%) than the training subsample (n = 75 out of 122 or 50%).

The higher the proportion of the misclassified cases, the higher their corresponding relative cost indices. For example, the relative cost of class 1 in the testing subsample is 0.50, which is the highest, and the relative cost of class 2 in the learning subsample is 0.155, which is the lowest. Table 6 also gives specificity and sensitivity statistics. Specificity of the learning sample is higher than that of the testing sample (78.75% > 50.00%), but the sensitivity of the testing sample is greater than the learning sample (73.91% > 61.48%). The area under the ROC curve was 0.78 in the learning subsample, which is regarded as fair (closer to good) and 0.60 in the testing sample which is poor.

Table 7 presents the nine variables which influenced item difficulty to varying degrees and three variables with no importance. Temporality had the highest influence as indicated by its Normalized Importance Index (100), followed by

(a) latent semantic analysis (the average givenness of each sentence) with the Normalized Importance Index of 94.83,

(b) word count (Normalized Importance Index = 53.46),

(c) content word overlap (Normalized Importance Index = 47.22),

(d) Flesch-Kincaid grade level (Normalized Importance Index = 42.07),

Table 7: The CART-Estimated Importance of Coh-Metrix Variables in Classifying Item Difficulty

Variable	Importance
(a) Temporality (a text easability index)	100.00
(b) Average givenness of each sentence (a latent semantic analysis or LSA index)	94.83
(c) Word count	47.22
(d) Flesch-Kincaid grade level	42.07
(e) Words before the main verb (an index of syntactic complexity)	17.50
(f) Verb incidence	15.80
(g) Noun and verb hypernymy	14.28
(h) Type-token ratio	12.14
(i) Prepositional phrase density	6.97
(j) Incidence of causal verbs	0.00
(k) Logical connectives incidence	0.00
(l) Content word overlap	0.00

(e) syntactic complexity measured by the words before the main verb (Normalized Importance Index = 17.50),

(f) verb incidence (Normalized Importance Index = 15.80),

(g) noun and verb hypernymy (Normalized Importance Index = 14.28),

(h) type-token ratio (Normalized Importance Index = 12.14), and

(i) prepositional phrase density (Normalized Importance Index = 6.97).

However, incidence of causal verbs, content word overlap, and logical connectives incidence made no contribution to test item difficulty.

Neural Networks

We analyzed the data using ANFIS and Perceptron Neural Network models. The results are reported in the following sections.

MLP Neural Network

The Perceptron Neural Network diagram comprised 12 input variables, four neurons in the hidden layer, and two output levels (low and high item difficulty levels). The input variables comprised word count, temporality, content word overlap, LSA (Latent Semantic Analysis), type-token ratio, logical connectives, causal verbs and prepositions, left embeddedness of the sentence, prepositional phrase, verb incidence, hypernymy, and Flesch Kincaid grade level.

Of the 241 items, 192 (79.7%) were used to train the network and 49 (20.3%) were used to test the network. Overall, the network had a fairly high accuracy, evidenced by the percentage of incorrect classifications in the training and testing stages which was 14% and 16%, respectively.

Table 8 gives the weight indices of the input and output variables. Unlike the β coefficients of the regression models, the ANN weight statistics have intra-variable variation. For example, the weights of word count level 1 across four neurons in the hidden layer, notated as $H_{(1:1\text{-}4)}$, is 0.472, -0.378, 0.118, and 0.069, indicating some degree of nonlinearity in the data. There is relatively high intra-variable variance in some inputs such as Flesch Kincaid grade level 2 (-1.081, -0.692, -0.939, and -0.008), indicating high nonlinearity between this variable and item difficulty. By contrast, intra-variable variance is relatively low in, for example, Word count level 3 (0.280, 0.086, 0.081, and 0.216), indicating some degree of linearity between this variable and item difficulty. Similarly, the weight of the connection between the hidden and output layers has a relatively large range, indicating high nonlinearity. For example, neuron $H_{(1:1)}$ gives 0.935 to low difficulty items (Difficulty = 1) and -0.967 to high-difficulty items (Difficulty = 2).

Table 8 also presents bias statistics for hidden and output layers. As previously noted, bias helps the network to learn the underlying patterns of the data more efficiently. The bias coefficients for the neurons in the hidden layer are 0.262, -0.630, -0.181, and -0.716, showing some degree of variation. The bias coefficients for the neurons in the output layer are -0.270 and -0.241, which suggests relatively less variation.

Table 9 presents the ANN classification results for the learning and testing subsamples. Unlike the CART model, the accuracy of classification of class 2 was higher than the accuracy of class 1 for both the training (n = 105 out of 118 or 89.00%) and testing subsamples (n = 23 out of 27 or 89.00%). The accuracy of classification of class 1 for the training subsample was 81.10% (n = 60 out of 74) which is close to the accuracy level of the testing subsample (n = 18 out of 22 or 81.80%). The overall accuracy of classification in the learning and testing samples was 85.90% and 83.70%, which is significantly higher than the CART modelling.

Table 9 also gives specificity and sensitivity statistics. Specificity of both the learning and testing subsamples is significantly high (81.10% and 81.80%, respectively), and so is the sensitivity statistics of both the learning and testing samples (89.00% & 85.20%). This provides further evidence that the ANN model outperformed the CART analysis significantly. The area under the ROC in the learning and testing subsamples curve is 0.900, which is considered excellent (Department of Math of the University of Utah, n.d.).

Finally, we estimated the Normalized Importance Index for the input variables. Table 10 presents the 12 variables which influenced item difficulty to varying degrees. Noun and verb hypernymy had the highest influence as indicated by its Normalized Importance Index (100), followed by Flesch-Kincaid grade level (Normalized Importance Index = 99.40). The Normalized Importance Index of the remainder of the input variables gradually decreases but never reaches zero, which runs counter to the CART modelling results: average givenness of each sentence (Normalized Importance Index = 77.00), prepositional phrase density

Table 8: Weight Indices of the Input and Output Variables of the Neural Network

		Predicted					
		Hidden Layer				Output Layer	
	Predictors in the input layer	$H_{(1:1)}$	$H_{(1:2)}$	$H_{(1:3)}$	$H_{(1:4)}$	Difficulty=1	Difficulty=2
Input Layer	(Bias)	0.262	-0.630	-0.181	-0.716		
	[Word count=1]	0.472	-0.378	0.118	0.069		
	[Word count=2]	-0.313	-0.312	-0.813	-0.325		
	[Word count=3]	0.280	0.086	0.081	0.216		
	[Temporality=0]	0.628	-0.315	-0.428	-0.080		
	[Temporality=1]	-0.530	0.531	0.569	0.106		
	[Temporality=2]	-0.121	-0.852	-0.815	-0.385		
	[Content word overlap=0]	0.364	0.495	-0.079	0.399		
	[Content word overlap=1]	0.349	-0.652	-0.050	-0.930		
	[Given-new sentences=0]	0.876	0.166	-0.869	0.178		
	[Given-new sentences=1]	-0.384	0.781	0.139	0.171		
	[Given-new sentences=2]	-0.314	-0.670	0.890	-0.455		
	[type-token ratio=1]	-0.031	0.027	-0.554	0.050		
	[type-token ratio=2]	0.275	-0.611	0.113	-0.138		
	[Logical connectives=0]	1.105	0.194	-0.731	-0.890		
	[Logical connectives=1]	-0.690	-0.528	0.032	0.490		
	[Causal verbs and particles=0]	0.546	0.564	-0.320	0.112		
	[Causal verbs and particles=1]	-0.358	-0.452	-0.128	-0.281		
	[Left embeddedness=0]	0.418	-0.492	-0.737	-0.056		
	[Left embeddedness=1]	-0.866	0.495	0.113	-0.536		
	[Left embeddedness=2]	-0.072	0.016	-0.497	-0.175		
	[Preposition phrase density=0]	0.189	-0.767	-0.207	0.004		
	[Preposition phrase density=1]	-0.522	-0.560	0.362	-0.020		
	[Preposition phrase density=2]	0.377	0.492	0.485	0.332		
	[Preposition phrase density=3]	-0.839	0.258	-0.082	-0.825		
	[Verb incidence=1]	0.923	0.096	0.210	0.592		
	[Verb incidence=2]	0.202	0.342	0.624	-0.111		
	[Verb incidence=3]	0.271	0.296	-0.092	0.123		
	[Verb incidence=4]	-0.845	-0.366	-0.328	-0.698		
	[Hypernymy=2]	-1.698	0.086	-0.236	-1.219		
	[Hypernymy=3]	0.995	0.230	-0.714	0.167		
	[Hypernymy=4]	0.341	-0.942	0.402	0.154		
	[Hypernymy=5]	0.119	-0.138	-0.241	-0.144		
	[Flesch Kincaid grade level=0]	1.086	0.132	-0.489	-0.570		
	[Flesch Kincaid grade level=1]	-0.016	0.989	0.532	-1.060		
	[Flesch Kincaid grade level=2]	-1.081	-0.692	-0.939	-0.008		
	[Flesch Kincaid grade level=3]	0.005	-0.212	0.515	0.224		
	[Flesch Kincaid grade level=4]	0.705	-0.574	-0.108	-0.196		
	[Flesch Kincaid grade level=5]	-0.472	0.334	-0.171	0.257		
Hidden Layer	(Bias)					-0.270	-0.241
	H(1:1)					0.935	-0.967
	H(1:2)					-1.477	1.450
	H(1:3)					0.965	-1.539
	H(1:4)					0.420	-0.650

Table 9: Classification Accuracy, Specificity, and Sensitivity Estimated by the ANN Model

	Class	No. of Cases	Accurately Classified Cases	Percentage	Relative Cost	Specificity	Sensitivity	ROC
Learning (*training*)	1[a]	74	60	81.10%				
	2[b]	118	105	89.00%				
	Overall	192	165	85.90%	NA	81.10%	89.00%	0.900
Testing	1	22	18	81.80%				
	2	27	23	85.20%				
	Overall	49	41	83.70%	NA	81.80%	85.20%	0.900

Note: [a] low-difficulty items
[b] high-difficulty items

Table 10: The ANN-Estimated Importance of Coh-Metrix Variables in Classifying Item Difficulty

Variable	Importance
(a) Noun and verb hypernymy	100.00
(b) Flesch-Kincaid grade level	94.83
(c) Average givenness of each sentence (a latent semantic analysis or LSA)	47.22
(d) Prepositional phrase density	42.07
(e) Verb incidence	17.50
(f) Temporality (a text easability index)	15.80
(g) Words before the main verb (an index of syntactic complexity)	14.28
(h) Word count	12.14
(i) Logical connectives incidence	6.97
(j) Type-token ratio	0.00
(k) Incidence of causal verbs and particles	0.00
(l) Content word overlap in adjacent sentences	0.00

(Normalized Importance Index = 62.30), verb incidence (Normalized Importance Index = 61.20), temporality (a text easability index) (Normalized Importance Index = 59.40), words before the main verb (an index of syntactic complexity) (Normalized Importance Index = 59.20), word count (Normalized Importance Index = 48.30), logical connectives incidence (Normalized Importance Index = 46.70), type-token ratio (Normalized Importance Index = 39.50), incidence of causal verbs and particles (Normalized Importance Index = 36.80), and content word overlap adjacent sentences (Normalized Importance Index = 33.60). Overall, the ANN modelling yielded different Importance Indices, higher accuracy, higher sensitivity, and higher specificity than the CART modeling and outperformed the linear regression model.

ANFIS Modelling

We combined the fuzzy membership functions with the ANN model to test whether the misclassified cases could be improved. Using various membership functions in the fuzzification phase, we initially trained a large number of ANFIS networks containing between one and 12 independent variables. The models were then tested on a test subsample to examine their precision. Table 11 present the best two-variable ANFIS models. Model 1 has the best fit to the data in both learning and testing stages: it fits the learning data moderately well (R = 0.801; R2 = 0.642; NMSE = 0.357; RMSE = 0.426; MAE = 0.350), but its fit to the testing data drops significantly (R = 0.427; R2 = 0.179; NMSE = 0.820; RMSE = 0.437; MAE = 0.392). Variable 1 (word count) appeared in the top five models in the learning and

Table 11: Best Two-Variable ANFIS Models Generated in the Learning Stage and Validated in the Testing Stage

Model	Input Variables	R	R Squared	NMSE	RMSE	MAE	Membership Function
Learning							
1*	1, 12	0.801	0.642	0.357	0.426	0.350	gbellmf
2	1, 2	0.796	0.634	0.365	0.431	0.372	gbellmf
3	1, 11	0.790	0.628	0.371	0.435	0.365	gbellmf
4	1, 10	0.778	0.605	0.394	0.447	0.389	gbellmf
5	1, 4	0.775	0.601	0.398	0.450	0.392	gbellmf
6	2, 11	0.790	0.630	0.360	0.429	0.359	gbellmf
7	2, 5	0.790	0.629	0.370	0.430	0.370	gbellmf
8	3, 11	0.776	0.603	0.396	0.449	0.376	gbellmf
9	4, 11	0.786	0.618	0.381	0.440	0.365	gbellmf
10	4, 5	0.781	0.611	0.388	0.444	0.382	gbellmf
11	5, 11	0.790	0.630	0.369	0.433	0.363	gbellmf
12	5, 12	0.786	0.618	0.381	0.440	0.375	gbellmf
13	6, 11	0.787	0.620	0.370	0.439	0.370	gbellmf
14	7, 11	0.787	0.620	0.379	0.430	0.370	gbellmf
15	8, 11	0.783	0.614	0.385	0.443	0.367	gbellmf
16	9, 11	0.775	0.601	0.398	0.450	0.378	gbellmf
17	10, 11	0.784	0.615	0.384	0.442	0.367	gbellmf
18	11, 12	0.786	0.618	0.381	0.440	0.364	gbellmf
Testing							
1*	1, 12	0.427	0.179	0.820	0.437	0.392	gbellmf
2	1, 2	0.386	0.137	0.862	0.448	0.415	gbellmf
5	1, 4	0.320	0.102	0.897	0.457	0.426	gbellmf
3	1, 11	0.429	0.177	0.822	0.437	0.395	gbellmf
12	5, 12	0.308	0.094	0.905	0.459	0.426	gbellmf

Note. * Best model. R = correlation between estimated and actual output; NMSE = normalized mean square error; RMSE = root-mean-square error; MAE = mean absolute error. gbellmf = Generalized bell curve membership function.

Variables: 1 = Word count; 2 = Temporality; 3 = Content word overlap (adjacent sentences); 4 = Average givenness; 5 = Type-token ratio; 6 = Logical connectives incidence; 7 = Incidence of causal verbs; 8 = Words before the main verb; 9 = Prepositional phrase density; 10 = Verb incidence; 11 = Noun and verb hypernymy; 12 = Flesch-Kincaid grade level.

testing stages, indicating its significance in determining the amount of output. Similarly, Variable 12 (Flesch-Kincaid grade level) had a significant influence over the output, since it emerged in Model 1 (the best fitting model) as well as Models 5 and 18. We tested other possible combinations in an attempt to determine the best fitting model. Because the remainder of the models performed poorly in the testing stage, they are not reported in Table 11. It is important to note that the best fitting models presented in the table used generalized bell curve membership function (gbellmf) in the fuzzification stage.

Table 12 presents the best three-variable ANFIS models. Although Models 2, 3, and 4 yielded larger R2 values during learning, due to their relatively poorer fit in the testing stage, we opted for Model 1 which fits the learning data moderately well (R = 0.807; R2 = 0.651; NMSE = 0.348; RMSE = 0.421; MAE = 0.341) but its fit to the testing data drops significantly (R = 0.430; R2 = 0.184; NMSE = 0.815; RMSE = 0.435; MAE = 0.383). As in the best two-variable models, Variables 1 (word count) and 12 (Flesch-Kincaid grade level) appeared in the three-variable models in the learning and testing stages, indicating their significance in determining the amount of output. The remainder of the models performed poorly in both the learning and testing stage and will not be reported here. Overall, the fit statistics of the three-variable models indicated that they did not classify the low and high-difficulty

Table 12: Best Three-Variable ANFIS Models Generated in the Learning Stage and Validated in the Testing Stage

Model	Input Variables	R	R Squared	NMSE	RMSE	MAE	Membership Function
Learning							
1*	1, 5, 12	0.807	0.651	0.348	0.421	0.341	gbellmf
2	1, 2, 12	0.825	0.682	0.317	0.402	0.324	gbellmf
3	1, 10, 12	0.827	0.685	0.314	0.400	0.309	gbellmf
4	1, 2, 11	0.826	0.68	0.316	0.401	0.324	gbellmf
Testing							
1*	1, 5, 12	0.430	0.184	0.815	0.435	0.383	gbellmf
2	1, 2, 12	0.392	0.150	0.849	0.444	0.393	gbellmf
3	1, 10, 12	0.310	0.09	1.560	0.603	0.445	gbellmf
4	1, 2, 11	0.023	0.00	3.320	0.879	0.472	gbellmf

Note. * Best model. R = correlation between estimated and actual output; NMSE = normalized mean square error; RMSE = root-mean-square error; MAE = mean absolute error. gbellmf = Generalized bell curve membership function.

Variables: 1 = Word count; 2 = Temporality; 3 = Content word overlap (adjacent sentences); 4 = Average givenness; 5 = Type-token ratio; 6 = Logical connectives incidence; 7 = Incidence of causal verbs; 8 = Words before the main verb; 9 = Prepositional phrase density; 10 = Verb incidence; 11 = Noun and verb hypernymy; 12 = Flesch-Kincaid grade level.

items with high accuracy and, as a result, we tested other combinations.

We postulated a number of four-variable ANFIS models, the best-fitting of which are presented in Table 13. We examined all models to identify the model that had the best fit in both learning and training stages. Several models had reasonably good fit to the learning data, but they fitted the test data poorly. We opted for Model 6 which fits the learning data moderately well (R = 0.823; R2 = 0.678; NMSE = 0.321; RMSE = 0.404; MAE = 0.316) even though its fit to the testing data drops significantly (R = 0.318; R2 = 0.163;

Table 13: Best Four-Variable ANFIS Models Generated in the Learning Stage and Validated in the Testing Stage

Model	Input Variables	R	R Squared	NMSE	RMSE	MAE	Membership Function
Learning							
1	3, 4, 5, 12	0.818	0.669	0.330	0.409	0.325	gbellmf
2	2, 5, 7, 11	0.835	0.698	0.301	0.390	0.310	gbellmf
3	2, 4, 5, 12	0.839	0.704	0.295	0.387	0.302	gbellmf
4	2, 3, 5, 8	0.810	0.657	0.342	0.417	0.349	gbellmf
5	2, 3, 5, 7	0.809	0.654	0.345	0.410	0.351	gbellmf
6*	1, 5, 9, 12	0.823	0.678	0.321	0.404	0.316	gbellmf
7	1, 5, 7, 12	0.824	0.670	0.320	0.400	0.313	gbellmf
8	1, 2, 3, 5	0.815	0.665	0.334	0.412	0.340	gbellmf
Testing							
1	3, 4, 5, 12	0.340	0.107	0.892	0.455	0.393	gbellmf
2	2, 5, 7, 11	0.381	0.113	0.886	0.454	0.375	gbellmf
3	2, 4, 5, 12	0.379	0.132	0.867	0.449	0.381	gbellmf
4	2, 3, 5, 8	0.362	0.131	0.868	0.449	0.416	gbellmf
5	2, 3, 5, 7	0.318	0.100	0.899	0.457	0.422	gbellmf
6*	1, 5, 9, 12	0.439	0.163	0.836	0.441	0.362	gbellmf
7	1, 5, 7, 12	0.355	0.100	0.899	0.457	0.379	gbellmf
8	1, 2, 3, 5	0.349	0.121	0.878	0.452	0.405	gbellmf

Note. * Best model. R = correlation between estimated and actual output; NMSE = normalized mean square error; RMSE = root-mean-square error; MAE = mean absolute error. gbellmf = Generalized bell curve membership function.

Variables: 1 = Word count; 2 = Temporality; 3 = Content word overlap (adjacent sentences); 4 = Average givenness; 5 = Type-token ratio; 6 = Logical connectives incidence; 7 = Incidence of causal verbs; 8 = Words before the main verb; 9 = Prepositional phrase density; 10 = Verb incidence; 11 = Noun and verb hypernymy; 12 = Flesch-Kincaid grade level.

Table 14: Best Five-Variable ANFIS Models Generated in the Learning Stage and Validated in the Testing Stage

Model	Input Variables	R	R Squared	NMSE	RMSE	MAE	Membership Function
Learning							
1*	1, 3, 6, 10, 12	0.657	0.432	0.567	0.369	0.273	gbellmf
2	1, 3, 6, 7, 12	0.557	0.308	0.691	0.408	0.333	gbellmf
3	1, 3, 5, 7, 12	0.530	0.281	0.718	0.416	0.346	gbellmf
4	1, 3, 5, 10, 12	0.599	0.359	0.640	0.393	0.309	gbellmf
5	1, 5, 7, 10, 12	0.611	0.373	0.626	0.388	0.302	gbellmf
6	1, 3, 7, 10, 12	0.598	0.356	0.643	0.393	0.311	gbellmf
7	1, 3, 4, 10, 12	0.630	0.397	0.602	0.381	0.290	gbellmf
8	3, 5, 6, 7, 12	0.495	0.245	0.754	0.426	0.363	gbellmf
9	1, 3, 4, 6, 12	0.579	0.335	0.664	0.400	0.320	gbellmf
10	1, 5, 6, 10, 12	0.633	0.400	0.599	0.380	0.290	gbellmf
Testing							
1*	1, 3, 6, 10, 12	0.696	0.457	0.542	0.354	0.283	gbellmf
2	1, 3, 6, 7, 12	0.621	0.379	0.620	0.379	0.318	gbellmf
3	1, 3, 5, 7, 12	0.652	0.379	0.620	0.379	0.333	gbellmf
4	1, 3, 5, 10, 12	0.614	0.355	0.644	0.386	0.301	gbellmf
5	1, 5, 7, 10, 12	0.568	0.308	0.691	0.399	0.318	gbellmf
6	1, 3, 7, 10, 12	0.563	0.307	0.692	0.400	0.303	gbellmf
7	1, 3, 4, 10, 12	0.553	0.304	0.695	0.401	0.326	gbellmf
8	3, 5, 6, 7, 12	0.549	0.297	0.702	0.403	0.357	gbellmf
9	1, 3, 4, 6, 12	0.543	0.286	0.713	0.406	0.342	gbellmf
10	1, 5, 6, 10, 12	0.540	0.283	0.716	0.407	0.307	gbellmf

Note. * Best model. R = correlation between estimated and actual output; NMSE = normalized mean square error; RMSE = root-mean-square error; MAE = mean absolute error. gbellmf = Generalized bell curve membership function.

Variables: 1 = Word count; 2 = Temporality; 3 = Content word overlap (adjacent sentences); 4 = Average givenness; 5 = Type-token ratio; 6 = Logical connectives incidence; 7 = Incidence of causal verbs; 8 = Words before the main verb; 9 = Prepositional phrase density; 10 = Verb incidence; 11 = Noun and verb hypernymy; 12 = Flesch-Kincaid grade level.

NMSE = 0.836; RMSE = 0.441; MAE = 0.362). As in the best two-variable models, Variable 1 (word count), Variable 2 (Temporality), Variable 4 (Average givenness), Variable 5 (Type-token ratio), and Variable 12 (Flesch-Kincaid grade level) surfaced in quite a few four-variable models in both the learning and testing stages, indicating their significance in determining the amount of output. The best four-variable model is the one which includes Variables 2, 3, 5, and 8.

Regardless of the type of membership functions, the four-variable ANFIS models did not yield promising results. Accordingly, we examined five-variable models to identify the model that might give the best fit in both learning and training stages. Table 14 presents 10 five-variable ANFIS models with reasonably good fit. Specifically, Model 1 fitted the test data moderately in the learning stage (R = 0.657; R2 = 0.432; NMSE = 0.567; RMSE = 0.369; MAE = 0.273) as well as the testing stage (R = 0.696; R2 = 0.457; NMSE = 0.542; RMSE = 0.354; MAE = 0.283). As in the best aforementioned models, Variables 1 (word count) and 12 (Flesch-Kincaid grade level) surfaced in quite a few four-variable models in both learning and testing stages. In addition, Variable 3 (Content word overlap) contributed to all models in the learning and testing stages.

Finally, we tested six- through 12-variable ANFIS models, but we found poor fit in a great proportion of the models. Indeed, as the complexity of the models increases, ANFIS's fit statistics decreased. Table 15 presents the best fitting six-variable models generated in the learning stage and validated in the testing stage.

Table 16 summarizes the best-fitting ANFIS models. It should be noted that even though they are entitled "best-fitting," their fit is relatively far from perfect. The five-variable model including Variables 1 (Word count), 3 (Content word overlap), 6 (Logical connectives incidence), 10 (Verb incidence), and 12 (Flesch-Kincaid

Table 15: Best Six-Variable ANFIS Models Generated in the Learning Stage and Validated in the Testing Stage

Model	Input Variables	R	R Squared	NMSE	RMSE	MAE	Membership Function
Learning							
1*	1, 4, 5, 8, 9, 10	0.729	0.532	0.467	0.335	0.225	gbellmf
2	1, 5, 7, 8, 9, 10	0.608	0.370	0.629	0.389	0.303	gbellmf
3	1, 4, 5, 7, 8, 9	0.702	0.493	0.506	0.349	0.244	gbellmf
Testing							
1*	1, 4, 5, 8, 9, 10	0.486	0.127	0.872	0.449	0.333	gbellmf
2	1, 5, 7, 8, 9, 10	0.355	0.086	0.913	0.459	0.393	gbellmf
3	1, 4, 5, 7, 8, 9	0.387	0.007	0.992	0.479	0.351	gbellmf

Note. * Best model. R = correlation between estimated and actual output; NMSE = normalized mean square error; RMSE = root-mean-square error; MAE = mean absolute error. gbellmf = Generalized bell curve membership function.

Variables: 1 = Word count; 2 = Temporality; 3 = Content word overlap (adjacent sentences); 4 = Average givenness; 5 = Type-token ratio; 6 = Logical connectives incidence; 7 = Incidence of causal verbs; 8 = Words before the main verb; 9 = Prepositional phrase density; 10 = Verb incidence; 11 = Noun and verb hypernymy; 12 = Flesch-Kincaid grade level.

Table 16: The Best Fitting ANFIS Models

Model	Input Variables	R	R Squared	NMSE	RMSE	MAE	Membership Function
Learning							
Two-variable	1, 12	0.801	0.642	0.357	0.426	0.350	gbellmf
Three-variable	1, 5, 12	0.807	0.651	0.348	0.421	0.341	gbellmf
Four-variable	1, 5, 9, 12	0.823	0.678	0.321	0.404	0.316	gbellmf
Five-variable*	1, 3, 6, 10, 12	0.657	0.432	0.567	0.369	0.273	gbellmf
Six-variable	1, 4, 5, 8, 9, 10	0.729	0.532	0.467	0.335	0.225	gbellmf
Testing							
Two-variable	1, 12	0.427	0.179	0.820	0.437	0.392	gbellmf
Three-variable	1, 5, 12	0.430	0.184	0.815	0.435	0.383	gbellmf
Four-variable	1, 5, 9, 12	0.439	0.163	0.836	0.441	0.362	gbellmf
Five-variable*	1, 3, 6, 10, 12	0.696	0.457	0.542	0.354	0.283	gbellmf
Six-variable	1, 4, 5, 8, 9, 10	0.486	0.127	0.872	0.449	0.333	gbellmf

Note. * Best model. R = correlation between estimated and actual output; NMSE = normalized mean square error; RMSE = root-mean-square error; MAE = mean absolute error. gbellmf = Generalized bell curve membership function.

Variables: 1 = Word count; 2 = Temporality; 3 = Content word overlap (adjacent sentences); 4 = Average givenness; 5 = Type-token ratio; 6 = Logical connectives incidence; 7 = Incidence of causal verbs; 8 = Words before the main verb; 9 = Prepositional phrase density; 10 = Verb incidence; 11 = Noun and verb hypernymy; 12 = Flesch-Kincaid grade level.

grade level) was the best fitting ANFIS model developed and tested in this study. Appendix B1 presents the fuzzy set rules of the five-variable model.

Figure 4 presents the five-variable ANFIS model tested in the study. The difference between the ANFIS and ANN models lies in the membership functions assigned to the input. In this figure, the sceond column from the left side (white circles) represents the membership functions. Every input will take either a low or a high membership function and the fuzzification rules, which are represented by the blue circles, will be applied to them. The output will be defuzzified and the predicted values will be estimated.

Discussion

This study was designed to examine the performance of three data mining methods: linear regression, classification and regression trees (CART) and two classes of artificial neural networks (ANNs): multilayer perceptron ANN and adaptive neuro-fuzzy inference system (ANFIS). The data came from the performance

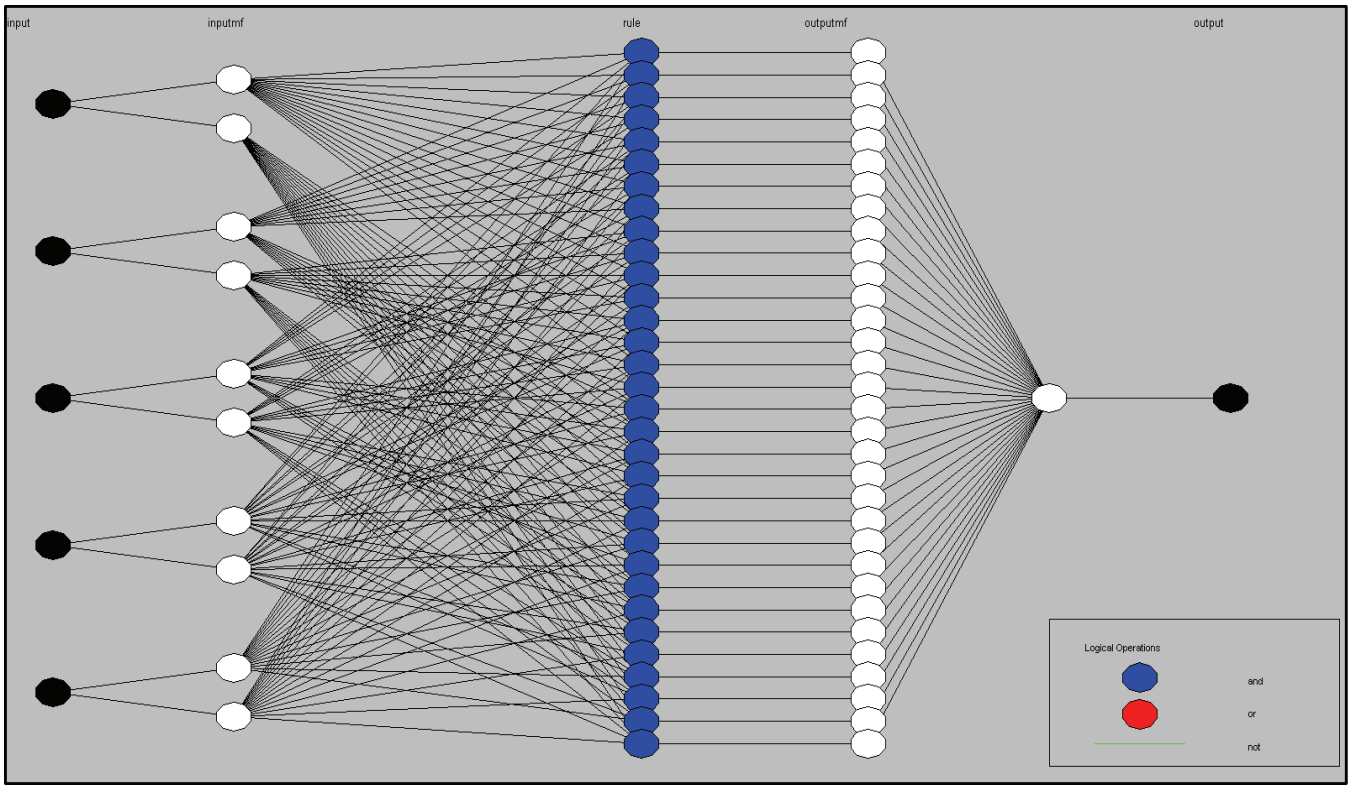


Figure 4: Structure of the Five-Variable ANFIS model

of 5,039 test takers on seven forms of the MET listening test. After screening the data, we used 242 items alongside their linguistic features estimated by Coh-Metrix.

Linear Regression Modelling

Initially, the assumptions of regression modelling including normality, lack of multicollinearity, and homoscedasticity were tested. Next, we estimated a regression model where 12 input variables were used to predict listening test item difficulty. We tested Enter, Remove, Stepwise, and Backward model selection procedures to identify the best fitting model. The methods provided highly similar solutions, the best model comprising two input variable (i.e., words before the main verb and noun and verb hypernymy) which explained 18.3% of the variance in the data.

The relatively low prediction power of the regression model in the present study is in the agreement with some of the previous Coh-Metrix studies such as McNamara, Crossley, McCarthy's study (2010) where the input variables accounted for 20% of the variance in the output variable, but inconsistent with, for example, Guo, Crossley, and McNamara's (2013) recent study where the Coh-Metrix variables account for more than half of the variance of the output. One reason for the low prediction power of the regression analysis in the present study might be the limitation of the texts analyzed by Coh-Metrix. Although every attempt was made to generate sufficiently lengthy texts, the texts related to each of the test items can be viewed as of medium lengths (approximately between 50 to 100 words), which is shorter than the texts used in previous Coh-Metrix analyses. Acknowledging that this could have affected the results of the analysis, we presume linear regression would be less efficient when the data is less than ideal. Therefore, it might be said that the output of linear regression models might not necessarily point to problems in the postulated theory tested; rather, it might suggest the inefficiency and imprecision of the statistical method and the requirement of using a more rigorous data mining technique that does not fall prey to the inherent features of the data.

CART Modelling

The precision of the CART modelling analysis was examined through specificity and sensitivity indices alongside the area under the ROC curve and

the proportion of accurately and inaccurately classified cases. The accuracy of the CART model was around 68% and 64% for the learning and testing subsamples, respectively. Although the area under the ROC curve was relatively small, the classification accuracy in the study is greater than, for example, Rupp et al.'s (2001) study.

One of the advantages of this study is the validation method of the accuracy of the model. We partitioned the data into learning and testing subsamples, thereby precluding the model from overfitting. In their CART study of two reading comprehension texts, Gao and Roger (2011) reported a high accuracy and prediction power, but did not test the estimated model on a testing sample. The highly accurate results can be attributed to overfitting which occurs when the estimated model is not validated across a testing sample (Steinberg et al., 1998). It is suggested that future researchers either use a testing sample or perform k-fold validation on their data so as to prevent overfitting (Breiman et al., 1984).

CART's outputs were relatively more consistent with the theoretical postulations than the results of the linear regression model. The precision of CART in classifying low- and high-difficulty test items is in line with previous CART research conducted by, for example, Brodley, and Utgoff (1995), Frank, Wang, Inglis, Holmes, and Witten (1998), and Kim and Loh (2001).

A point concerning the choice of CART computer packages is in place. The first split of CART models has a significant impact on the tree structure. If the input variable has no effect on the output variable, the CART tree will be less precise. It is important to choose an algorithm which can identify the best splitter for the topmost node (Hsieh, Hsiao, Chang, Wang, & Fann, 2011). We found that the SPM algorithm would provide a more precise solution than, for example, the SPSS CART algorithm. This, however, does not indicate our endorsement of SPM products or underestimating the SPSS algorithm. We suggest that future researchers consider the powerful CART algorithm if achieving high precision is a priority in their research.

Artificial Neural Networks (ANNs)

We initially tested an MLP ANN with one hidden layer, which proved to be the most precise and substantively reliable model among the data mining models assessed. It achieved excellent classification accuracy and precision in both the learning and testing subsamples. The MLP ANN analysis showed that all posited input variables would contribute to differentiating test items. That is, if we have the 12 linguistic features of test items, we can predict with a high degree of accuracy whether the test item will be difficult or easy. ANNs have achieved significantly high precision over regression models in previous research (Crone, Guajardo, & Weber, 2006). The precision is attributable to the flexible structure of the models and their powerful algorithms. However, ANN has hardly been used in language assessment, although it offers significant advantages over conventional approaches such as discriminant analysis.

By contrast, ANFIS modelling did not achieve the expected precision. The best model comprised five variables and displayed moderate fit to both learning and testing subsamples, which runs counter to the study conducted by Aryadoust (2013a). It might be that ANFIS can perform well in data with small size but poorly in large data sets. The poor performance on ANFIS might be due to the fuzzy set components, because the ANN with no fuzzy set functions proved to have a high precision.

In all, we found MLP ANN had high precision, followed by CART, ANFIS, and linear regression, respectively. The study shows that the type of data analysis techniques can exert a significant impact on the results and claims. It is plausible to presume that less precise statistical models would have the potential to refute the otherwise well-established hypotheses and/or theories, thereby convincing the researcher that the postulated models are inaccurate or less useful than expected. Caution should be exercised when choosing prediction and classification models.

Further Predictive / Classification Models

This study set out to compare the three previously discussed data mining approaches. Future research can assess the predictive power of logistic regression, generalized linear models (see, for example, Cheong, 2006), hierarchical linear models (see, for example, Barkaoui, 2013), generalized linear mixed method, genetic-programing symbolic regression (see, for example, Aryadoust, 2014), and automatic linear modelling (see, for example, Aryadoust, 2013b).

Table 17: Variable Importance Indices Estimated by the MLP ANN and CART Models

Variable	ANN's Normalized Importance	CART's Normalized Importance
(a) Noun and verb hypernymy	100.00	14.28
(b) Flesch-Kincaid grade level	99.40	42.07
(c) Average givenness of each sentence (a latent semantic analysis or LSA)	77.00	94.83
(d) Prepositional phrase density	62.30	6.97
(e) Verb incidence	61.20	15.80
(f) Temporality (a text easability index)	59.40	100.00
(g) Words before the main verb (an index of syntactic complexity)	59.20	17.50
(h) Word count	48.30	47.22
(i) Logical connectives incidence	46.70	0.00
(j) Type-token ratio	39.50	12.14
(k) Incidence of causal verbs and particles	36.80	0.00
(l) Content word overlap in adjacent sentences	33.60	0.00

Implications of the Findings for Listening Comprehension Assessment

As earlier noted, choosing the right statistical model for data mining is of paramount importance. Choosing linear regression over other models would result in a significantly different and less accurate theory. In this section, we examine the results of the MLP ANN alongside the CART model as presented in Table 17, since they provided the most plausible and intuitive solutions.

The MLP ANN modelling showed that all postulated input variables would exert a sizeable impact on the difficulty of the test items. The CART model also yielded similar results, although it estimated the Importance Index of zero for logical connectives, causal verbs and prepositions, and content word overlap, which had the lowest Importance Index in the MLP ANN analysis. The input variables recognized as influential in both models had nevertheless different Importance Indices. For example, the most important variable in the MLP ANN analysis was noun and verb hypernymy, whereas hypernymy had one of the lowest Importance Indices in the CART modelling. By contrast, the most important variable in the CART analysis was temporality—a text easability index—which had a medium Importance Index in the MLP ANN analysis.

Previous research has yielded inconclusive results concerning the effect of hypernymy on texts. For example, Crossley and McNamara (2010) argued that lower hypernymy indices indicate the use of more general words, whereas higher hypernymy values indicate the use of fewer general words. Hypernymy has been shown to improve alongside vocabulary knowledge over time, suggesting that high-ability English learners would have a fairly extended hypernymy knowledge (Crossley, Salsbury, & McNamara, 2009). This finding coupled with the findings of the present study would indicate that hypernymy can help distinguish low- and high-ability listening test takers.

Although previous research showed that the Coh-Metrix reading difficulty index can estimate the level of text difficulty more accurately than the Flesch-Kincaid grade level index (Crossley & McNamara, 2010; Crossley et al., 2011), in this study we found that the latter index would discriminate low- and high-difficulty listening items. We presume that the Coh-Metrix reading difficulty index might be less sensitive in short texts, whereas Flesch-Kincaid grade level index might be sensitive to the difficulty of short texts. Since Flesch-Kincaid grade level index has rarely been used for estimating the difficulty of listening texts, it is highly desirable that this index alongside its counterpart Coh-Metrix index be studied in future listening research.

Average givenness of each sentence, which measures the semantic overlap between sentences, was chosen because it had a high correlation with the item difficulty indices and because the item stems, options, and necessary information had some semantic overlap in

the present study (see Hempelmann, Dufty, McCarthy, Graesser, Cai, & McNamara, 2005; McCarthy, Dufty, Hempelmann, Cai, Graesser, & McNamara, 2012). We anticipated that this index would be able to discriminate low- and high-difficulty items because the semantic overlap between the necessary information (NI) and item stem or options can determine item difficulty (Buck & Tatsuoka, 1998). This LSA feature is, therefore, a useful index in listening assessment studies and test development, since it shows the effect of the repetition of NI in the listening text and stimuli on item difficulty. Relatedly, higher values of logical connectives incidence, which shows coherence in texts, can facilitate comprehension and decrease item test difficulty (see McNamara et al., 2010).

Prepositional phrase density measures the syntactic density of texts; dense texts tend to be more difficult to parse and comprehend (Crossley et al., 2012; Biber, Gray, & Poonpon, 2011). Together with verb incidence, incidence of causal verbs and particles and word count—which measure surface features of texts—prepositional phrase density can help item developers predict the difficulty of their test items. For example, it is expected that lengthier listening texts tax cognitive resources and working memory of listeners. Similarly, verb and prepositional density can probably increase the information density of the listening texts and likely render the test items more difficult. Another related index is type-token ratio, which as Crossley et al. (2012) argued, tends to alter the difficulty of test items.

Both MLP ANN and CART identified temporality as an influential variable in discriminating between low- and high-difficulty items. Previous research shows that texts with higher and more consistent tense and aspect (i.e., higher temporality) are relatively easier to comprehend. Relatedly, temporal cohesion would contribute to the comprehenders' situation model and accordingly their comprehension of the events in the oral/written message (McNamara, Louwerse, Cai, & Graesser, 2013; McNamara, Graesser, McCarthy, & Cai, in press). The temporality values would help item developers to control the difficulty of test items. For example, repeating tense consistently can create cohesion in texts and facilitate the activation of schemata or information in the working memory (Duran, McCarthy, Graesser & McNamara, 2007). It can also help maintain the construct validity of the listening test items / tasks. Natural oral and written discourse includes consistent repetitions of tense and aspect. If any of these two elements shifts, the comprehender would need signal words that reflect the shift (e.g., *the next day* or *tomorrow*). If these words are deleted during text simplification (a process that occurs in preparing listening tests), test takers will encounter problems due to the unnaturalness of the listening text, thereby undermining the construct validity of the test (Buck, 2001).

Another way listening test items are a challenge is when their syntactic complexity is increased by, for example, increasing the verbal density before the main verb. Texts with a large number of words before the main verb are more difficult to parse (Crossley et al., 2012; Graesser, Cai, Louwerse, & Daniel. 2006; Just & Carpenter (1992). In listening research, Aryadoust, Mehraban, and Alizadeh (2014) also verified the influence of words before the main verb. Nevertheless, caution should be exercised in adjusting the syntactic complexity of test items and NI. Overloading the NI's syntax might render the listening text unnatural and jeopardize the cognitive validity of the test (Field, 2009, 2012). Relatedly, content word overlap in adjacent sentences would facilitate text comprehension because it provides clues as to what concepts or ideas are important and should be attended to carefully to reach local and global comprehension.

Having determined variables affecting item difficulty, it is desirable to explore the interrelations of the variables. For example, it is important to determine whether the effect of syntactic density or temporality is moderated through, for example, word count. That is, would lengthier texts with higher syntactic density be less difficult than texts with lower syntactic density? It is important that future researchers address such questions so as to move toward developing a theory of item difficulty.

Conclusion and Implications

This study has methodological and substantive implications for both CaMLA and the wider field. Methodologically, although previous research has examined the effect of item- and text-related variables on item difficulty, most studies apply traditional methods such as regression models which assume linearity of data. CART, and to a greater extent MLP ANN provided an alternative evaluation of the underlying patterns of the listening tests. We believe MLP ANN would hold promise in the investigation of language assessment and this study is one of the first of its kind to do so.

Using Coh-Metrix in listening in the present study is innovative. Unlike previous studies, the application of these measures to listening tests which essentially involve spoken texts is yet to be explored. The findings of the study will inform the usefulness of such methods for studying the complexity in spoken language.

The present study has further contributed to the validity argument for the MET listening test and specifically supports *the explanation inference* of the MET. The explanation inference is invoked when there is evidence that the test data are related to the latent trait measured by the test and the variables influencing item difficulty are construct-relevant (Aryadoust, 2013). The present study has yielded evidence supporting the relevance of the influential lexical features of MET items to a general listening construct.

First, hypernymy knowledge is intimately related to learners' vocabulary knowledge, and its significant role in determining item difficulty suggests that the items demanding a greater depth of vocabulary knowledge would be more challenging for test takers. Put differently, test takers who have a deeper lexical level knowledge are expected to process highly hypernymic relations more accurately than the low-ability test takers who typically have a shallow lexical level knowledge. Although this topic has hardly been researched in listening studies, vocabulary acquisition research provides supporting evidence for the determining role of hypernymy in the depth of vocabulary knowledge (Haastrup & Henriksen, 2000).

Another vocabulary-related index is Flesch-Kincaid grade level which distinguished low- and high-difficulty items. This finding also yields warrants for the validity argument of the MET, as difficult oral texts would tax cognitive resources of test takers and it will be more difficult to score highly on the items (Buck, 2001). Similarly, variations in the text features related to the listening construct such as prepositional phrase density, word count, and connectives would influence the difficulty level of listening messages. This finding resonates with previous listening research (e.g., Buck & Tatsuoka, 1998; Geranpayeh & Taylor, 2013) and offers support for the validity argument of the MET.

The significant role of average givenness of each sentence and content word overlap would indicate that overlaps between the item stems, options, and NI would determine item difficulty. This finding can be taken as evidence supporting the explanation inference of the MET, as it is consistent with previous research where the role of overlaps—as an important listening construct-relevant factor—in determining cognitive demands of test items has been shown (Buck & Tatsuoka, 1998).

The main implication of these findings for the writing of listening test items for the MET is that manipulating textual features would affect item difficulty. For example, oral texts and the NI (necessary information) demanding a higher level of hypernymy knowledge and/or with higher Flesch-Kincaid grade levels can be used to increase the cognitive demand and difficulty level of the test; conversely, low-level hypernymy texts or NI, or texts with lower Flesch-Kincaid grade levels can be more suitable for low-ability test takers. However, further experimental research is required to show that these variables would indeed help item writers develop high-quality listening items, as other text features can also influence test difficulty. Accordingly, we propose three potential topics for future research: (a) the investigation of the effect of prosodic features such as stress, rhythm and intonation, rate of speech delivery, and text topics on listening test difficulty; (b) test takers' cognitive processes when answering the test items; and (c) the combined effect of *a* and *b* with the important lexical features identified in the present study. It is hoped that future listening test researchers will examine the effects of these variables in the MET and other listening tests.

Acknowledgements

We are grateful to Varun Subramanian at the National University of Singapore for helping in the ANFIS analysis, Catherine Evashuk for proofreading the manuscript, and Kavishna Ranmali and Desmond Wong for transcribing the listening texts and performing the Coh-Metrix analysis.

References

Adams, R., Carson, J., & Cureton, K. (1993). *Item difficulty adjustment study: GRE verbal discretes.* (ETS RR-92-79). Princeton, NJ: ETS.

Alderson, C. J., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im) possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535–556. doi: 10.1177/0265532209340188

Aryadoust, V. (2012). Using cognitive diagnostic assessment to model the structure of the lecture comprehension section of the IELTS listening test: A sub-skill-based approach. *Asian EFL Journal, 13*(4), 81–106.

Aryadoust, V. (2013a). Predicting item difficulty in a language test with an Adaptive Neuro Fuzzy Inference System. *IEEE Workshop on Hybrid Intelligent Models and Applications, 2013*, 43–55. doi: 10.1109/HIMA.2013.6615021

Aryadoust, V. (2013b, October). *Can computer-generated linguistic features predict second language students' writing scores across time?* Proceedings of the Technology-enhanced Learning (TEL2013) Conference, National University of Singapore.

Aryadoust, V. (2014). Application of genetic algorithm-based symbolic regression in ESL writing research. In V. Aryadoust & J. Fox (Eds.) *Current trends in language testing in the Pacific Rim and the Middle East: Policies, analysis, and diagnoses*. Newcastle: Cambridge Scholars Publishing.

Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361–385. doi: 10.1080/15434303.2011.628632

Aryadoust, V., Mehraban, P., & Alizadeh, P. (2014). Using a perceptron neural network to classify reading test items of Iranian Entrance Exam for engineering graduate students. In V. Aryadoust & J. Fox (Eds.) *Current trends in language testing in the Pacific Rim and the Middle East: Policies, analysis, and diagnoses*. Newcastle: Cambridge Scholars Publishing.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*(2), 192–211.

Barbour, B., Brunel, N., V. Hakim, & Nadal, J. P. (2007). What can we learn from synaptic weight distributions? *TRENDS in Neurosciences,* 28, *387–394.* doi:10.1016/j.tins.2007.09.005

Barkaoui, K. (2013). Using multilevel modeling in language assessment research: A conceptual introduction. *Language Assessment Quarterly*, *10*(3), 241–273. doi: 10.1080/15434303.2013.769546

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45(1),* 5–35. doi: 10.5054/tq.2011.244483

Bond, T. G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J.: Lawrence Erlbaum. doi: 10.1111/j.1745–3984.2003.tb01103.x

Breiman, L. (1994). *Bagging predictors.* Technical report, Department of Statistics, University of California, Berkeley, CA. doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brodley, C. E., & Utgoff, P. E. (1995). Multivariate decision trees. *Machine Learning, 19*, 45–77. doi: 10.1007/BF00994660

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157. doi: 10.1177/026553229801500201

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423–466. doi: 10.1111/0023-8333.00016

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*(1), 45–73. doi: 10.1037/h0076248

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*, 57–79. doi: 10.1207/s15327574ijt0601_4

Crone, S. F., Guajardo, J., & Weber, R. (2006). A study on the ability of Support Vector Regression and Neural Networks to Forecast Basic Time Series Patterns. *Artificial Intelligence in Theory and Practice, 217,* 149–158. doi: 10.1007/978-0-387-34747-9_16

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, *16*(1), 89–108. doi: 10.1177/1362168811423456

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(2) 15-30. doi: 0026-7902/07/15–30

Crossley, S. A., & McNamara, D. S. (2008). Assessing Second Language Reading Texts at the Intermediate Level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). *Language Teaching, 41*(3), 409–229. doi:10.1017/S0261444808005077

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In R. Catrambone, & S. Ohlsson, (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp.984–989). Cognitive Science Society, Austin, TX.

Crossley, S. A., & Salsbury, T. (2010). Using lexical indices to predict produced and not produced words in second language learners. *The Mental Lexicon, 5* (1), 115–147. doi: 10.1075/ml.5.1.05cro

Crossley, S. A, Salsbury, T., & McNamara, D. S. (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307–334. doi: 10.1111/j.1467-9922.2009.00508.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 240–260. doi: 10.1177/0265532211419331

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing, 28*(4), 561-580. doi: 10.1177/0265532210378031

Daftarifard P., & Lange R. (2009). Theoretical complexity vs. Rasch item difficulty in reading tests. *Rasch Measurement Transactions, 23*(2), 1212–1213.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Journal of Experimental Education*, *56*(2), 67–76.

Department of Math of the University of Utah (n.d.). *Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model*. Retrieved from http://www.math.utah.edu/~gamez/files/ROC-Curves.pdf

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*(4), 486–514. doi: 10.2307/747313

Duran, N.D., McCarthy, P.M., Graesser, A.C. & McNamara, D.S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods, 29*, 212–223. doi: 10.3758/BF03193150

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175–193. doi: 10.1177/014662168701100207

Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9). Canberra: IELTS Australia, Pty Ltd & British Council.

Field, J. (2012). Into the mind of the academic listener. *Journal of English for Academic Purposes, 10*(2), 102–112. doi: 10.1016/j.jeap.2011.04.002

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. (1998). Using model trees for classification. *Machine Learning, 32*, 63–82. doi: 10.1023/A:1007421302149

Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 162-192). Norwood, NJ: Ablex.

Freedle, R., & Kostin, I. (1991). *The prediction of SAT reading comprehension item difficulty for expository prose passages* (ETS Research Report No. RR-91-29). Princeton, NJ: Educational Testing Service.

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 explicit statements* (ETS Research Report No. RR-91-59). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1993a). *The prediction of TOEFL reading comprehension item difficulty for the expository prose passages for three item types: Main idea, inference, and supporting idea items* (ETS Research Report No. RR-93-13). Princeton, NJ: Educational Testing Service. doi: 10.1177/026553229301000203

Freedle, R., & Kostin, I. (1993b). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*(2), 133–170.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for the expository prose passages for minitalk passages: Implications for construct validity* (ETS Research Report No. RR-96-29). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*(1), 2–32. doi: 10.1177/026553229901600102

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 4*, 1–40.

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing,* 28(1), 77–104. doi: 10.1177/0265532210364380

Geisser, S. (1971). The inferential use of predictive distributions. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference* (pp. 456–469). Holt, Rinehart, and Winston, Toronto.

Goh, C., & Aryadoust, V. (2010). Investigating the construct validity of MELAB listening test through the Rasch analysis and correlated uniqueness modeling. *Spaan Fellowship Working Papers in Second of Foreign Language Assessment, 8*, 31–68.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3–22.

Grant, L. & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*, 123–145. doi:10.1093/poq/nfj012

Geranpayeh, A., & Taylor, L. (Eds.) (2013). *Examining listening: Research and practice in assessing second language listening, Studies in language testing volume 35*. Cambridge: UCLES/Cambridge University Press.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement*, *44*(3), 551–561. doi: 10.1177/0013164484443002

Grimes, J. (1975). *The thread of discourse*. The Hague: Mouton.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218–238.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R.L. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Hare, V., Rabinowitz, M., & Schieble, K. (1989). Text effects on main idea comprehension. *Reading Research Quarterly*, *24*(1), 72–88. doi: 10.2307/748011

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10*(2), 221–240.

Hempelmann, C. F., Dufty, D., McCarthy, P., Graesser, A. C., Cai, Z., & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In B. Bara (Ed.*), Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 941–946). Mahwah, NJ: Erlbaum.

Hsieh, A. R., Hsiao, C. L., Chang, S. W., Wang, H. M., Fann, C. S. J. (2011). On the use of multifactor dimensionality reduction (MDR) and classification and regression tree (CART) to identify haplotype–haplotype interactions in genetic studies. *Genomics, 97*, 77–85. doi: 10.1016/j.ygeno.2010.11.003

Huff, K. (2003). *An item modeling approach to providing descriptive score reports.* Unpublished doctoral dissertation, University of Massachusetts Amherst.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

Jensen, K., Muller, H.H., & Schafer, H. (2000). Regional confidence bands for ROC curves. *Statistics in Medicine, 19*, 493–509. doi: 10.1002/(SICI)1097-0258(20000229)1

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension. *Psychological Review, 99,* 122–149.

Kim, H., & Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association, 96*, 589–604.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Kostin, I. (2004). *Exploring item characteristics that are related to difficulty of TOEFL dialogue items*. (Research Report No. 79). Princeton, NJ: Educational Testing Service.

Landín, M., Rowe, R. C., & York, P. (2009). Advantages of neurofuzzy logic against conventional experimental design and statistical analysis in studying and developing direct compression formulations. *European Journal of Pharmaceutical Sciences, 38*(4), 325–331.

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172–189. doi: 10.1080/15434300902985108

LotfiZadeh, A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353.

Lui, H., Hussain, F., Chew, L. T., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery, 6,* 393–423.

McCarthy, P. M., Dufty, D., Hempelmann, C., Cai, Z., Graesser, A. C., & McNamara, D. S. (2012). Newness and givenness of information: Automated identification in written discourse. In P.M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 457–478). Hershey, PA: IGI Global.

McNamara, D. S., Crossley, S. A., McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27, 57–86. doi: 10.1177/0741088309351547

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0*. Retrieved from http://cohmetrix.com

McNamara, D. S., Ozuru, Y., Graesser, A. C., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual conference of the cognitive science society* (pp. 573–578). Austin, TX: Cognitive Science Society.

Meyer, B., & Freedle, R. (1984). The effects of different discourse types on recall. *American Educational Research Journal*, *21*(1), 121–143.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). *Five papers on WordNet.* Cognitive Science Laboratory, Princeton University, No. 43.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension.* (TOEFL Research Rep. No. 51). Princeton, NJ: ETS.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity, 41*(5), 673–690. doi: 10.1007/s11135-006-9018-6

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing, 12*, 34–53. doi: 10.1177/026553229501200103

Riazi, A. M., & Knox, J. S. (2014). An investigation of the relations between test takers' first language and the discourse of written performance on the IELTS academic writing test, task 2. *IELTS Research Reports*. London: British Council/IDP Australia.

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics, 12*, 475–511. doi: 10.1198/1061860032238

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to under¬stand difficulty in second language reading and listening comprehension tests. *International Journal of Testing, 1*, 185–216. doi: 10.1080/15305058.2001.9669470

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL® iBT listening section* (TOEFL iBT Research Report No.TOEFLiBT-08). Princeton, NJ: Educational Testing Service.

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190–209. doi: 10.1080/15434300902801917

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, *10*, 153–178.

Seyoum, S., Richardson, E., Webb, P., Riely, F., & Yohannes. Y. (1995). *Analyzing and mapping food insecurity: An exploratory CART methodology applied to Ethiopia.* Final report to the United States Agency for International Development. International Food Policy Research Institute, Washington, D.C.

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

Sheehan, K., & Ginther, A. (2001). *What do multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on a standardized test of reading comprehension skill.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Steinberg, D., Colla, P. (1995). *CART: Tree-structured non-parametric data analysis*. San Diego, Calif., U.S.A.: Salford Systems.

Steinberg, D., Colla, P., & Martin, K. (1998). CART—*Classification and regression trees: Supplementary manual for Windows.* San Diego, Calif., U.S.A.: Salford Systems.

Stenner, J. A., Stone, M., & Burdick, D. (2011). *How to model and test for the mechanisms that make measurement systems tick.* Proceedings of the Joint International IMEKO TC, TC7, and TC13Symposium, Jena, Germany.

Stevens, J. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Mahwah, NJ: Lawrence Erlbaum.

Subramanian, K., & Suresh, S. (2012). *Human action recognition using Meta-Cognitive Neuro-Fuzzy Inference System.* Paper presented at the 2012 International Joint Conference on Neural Networks (IJCNN).

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Lawrence Erlbaum Associates, Mahwah, NJ.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–354.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the Rule Space Method.* New York, NY: Routledge.

Verlinden, B., Duflou, J., Collin, P., Cattrysse, D. (2008). Cost estimation for Sheet Metal Parts using multiple regression and artificial neural networks: A case study. *International Journal of Production Economics, 111*(2), 484–492. doi:10.1016/j.ijpe.2007.02.004

Wolpert, D. H. (1992). *On overfitting avoidance as bias.* Technical Report SFI TR 92-03-5001, The Santa Fe Institute.

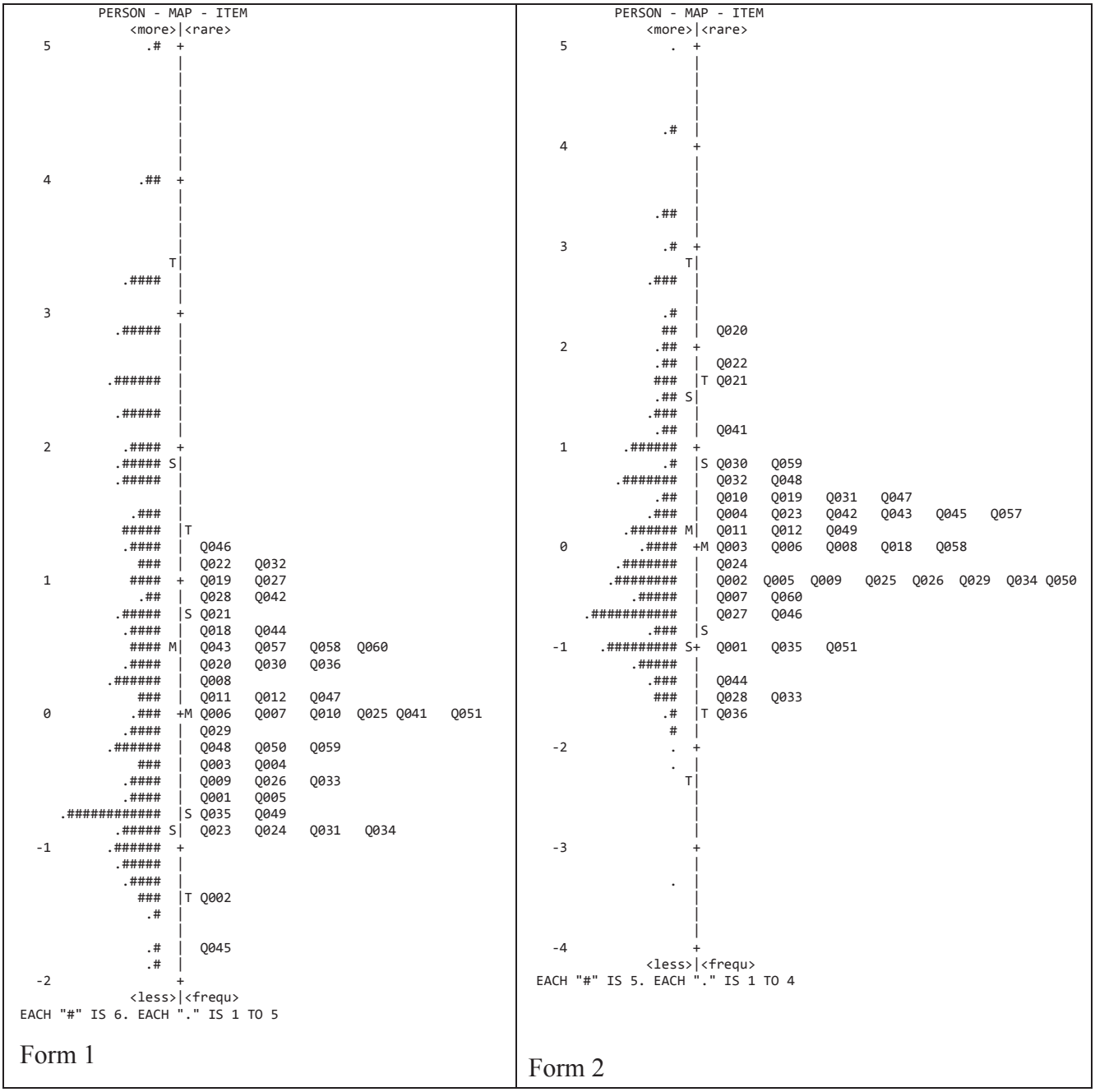
Yohannes, Y., & Webb. P. (1998). *Classification and regression trees: A user manual for identifying indicators of vulnerability to famine and chronic food insecurity.* International Food Policy Research Institute, Washington, D.C. Mimeo.

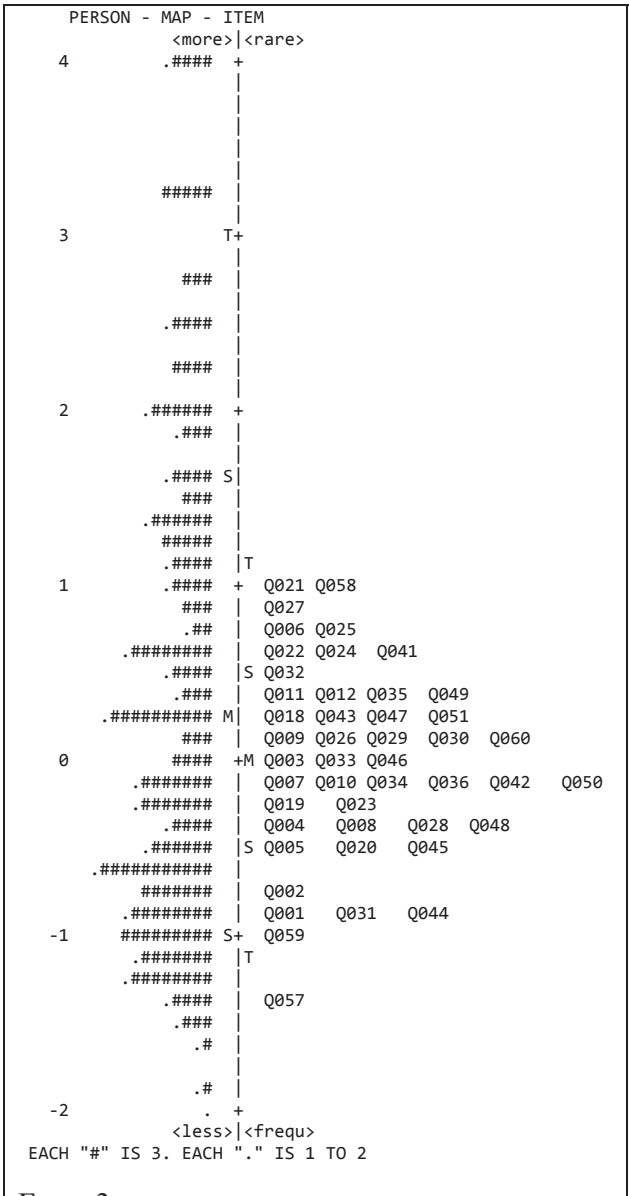
Zhang, L-M., Goh, C., & Kunnan, A. (2014). Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach. *Language Assessment Quarterly.*

Zhou X. H., & Qin, G. (2005). Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Statistics in Medicine, 24,* 465–477. doi: 10.1002/sim.1563

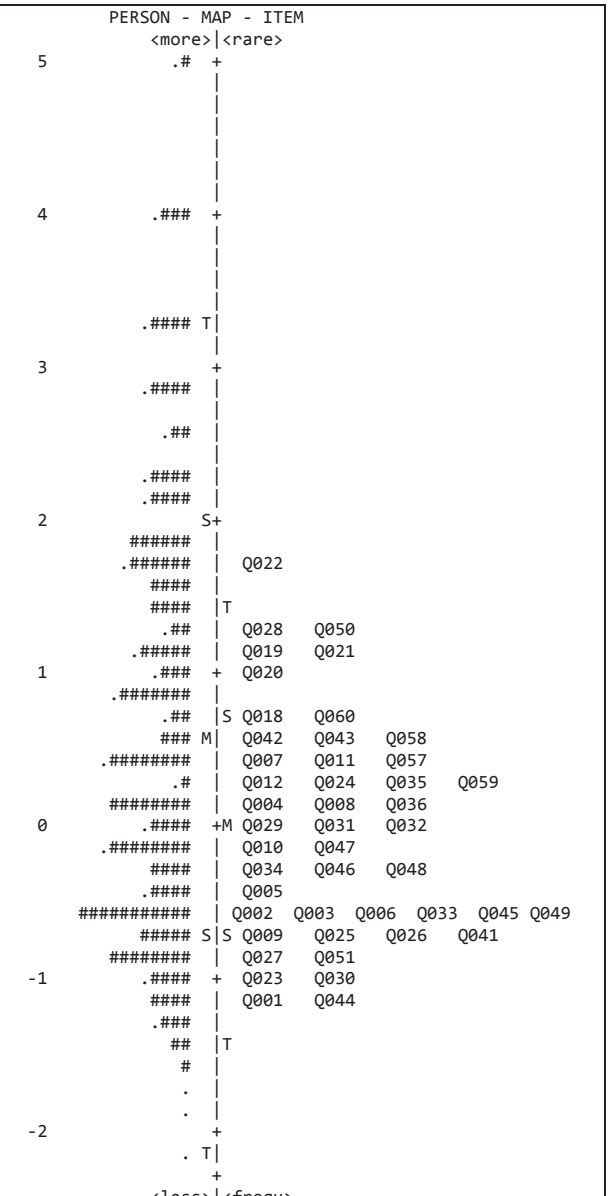
Appendices

Appendix A1: Item Person Map of the Tests

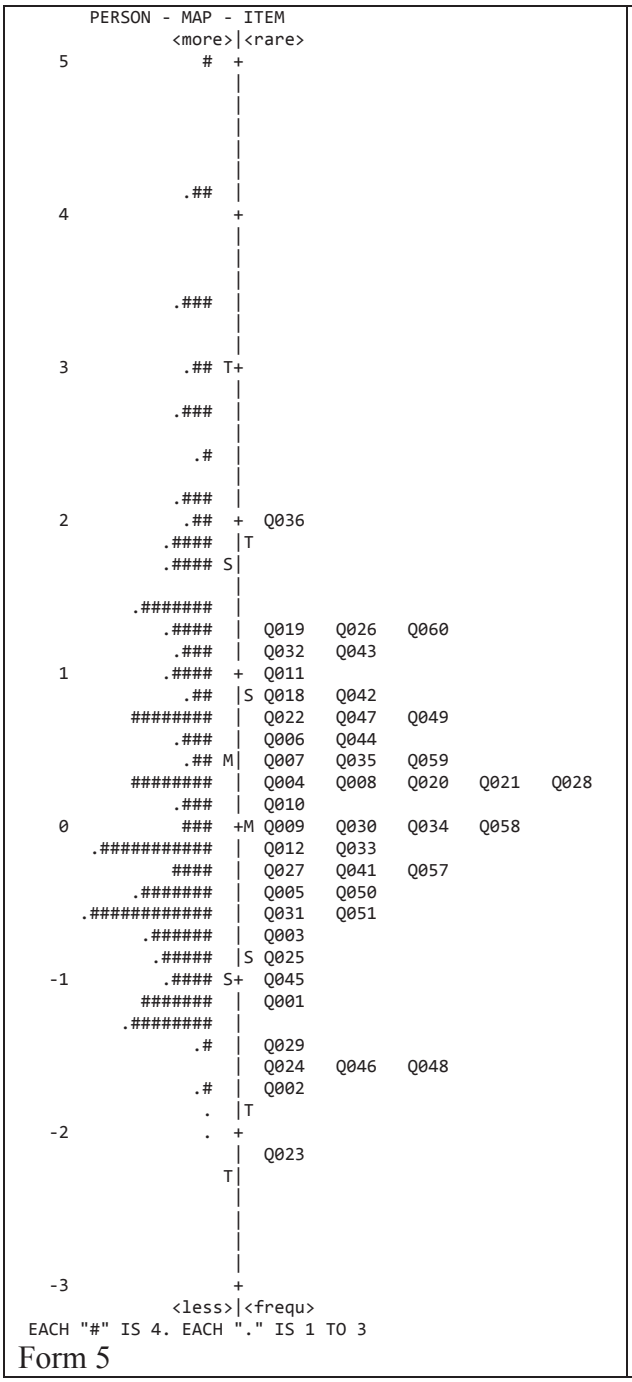


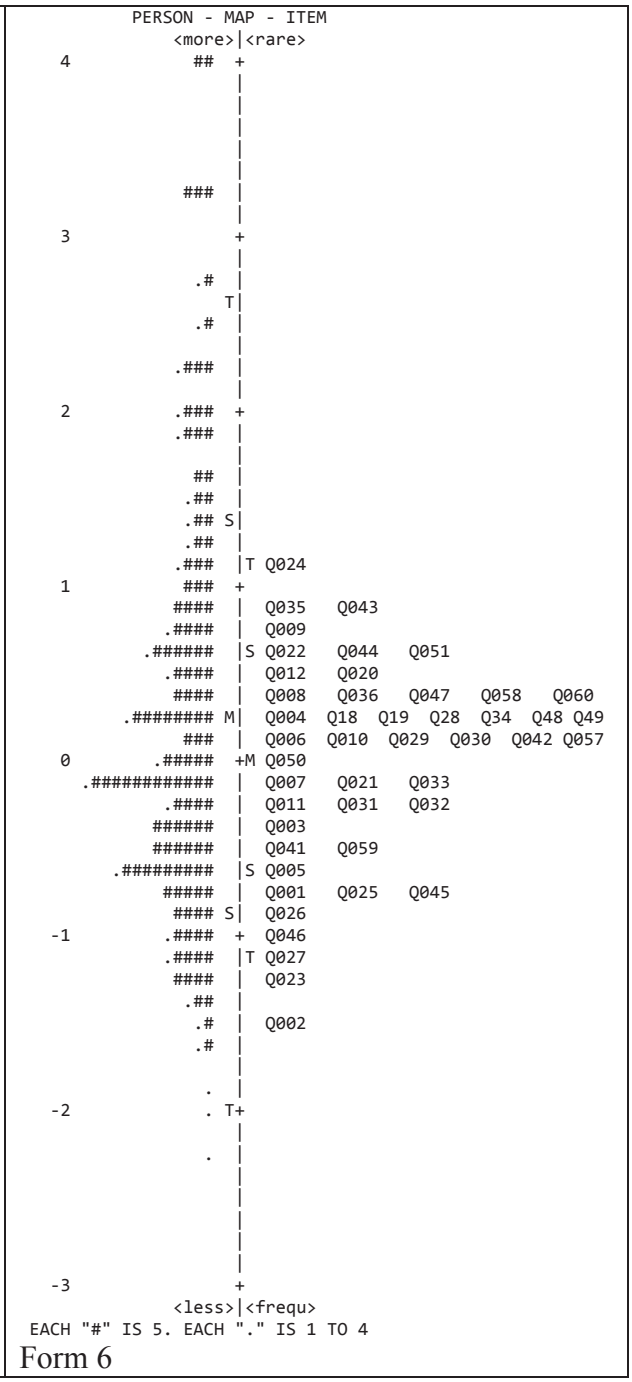

PERSON - MAP - ITEM
<more>|<rare>
4 .#### +
.#####
3 T+
###
.####
####
2 .###### +
.###
.#### S|
###
.######
#####
.#### |T
1 .#### + Q021 Q058
| Q027
.## | Q006 Q025
.######## | Q022 Q024 Q041
.#### |S Q032
.### | Q011 Q012 Q035 Q049
.########## M| Q018 Q043 Q047 Q051
| Q009 Q026 Q029 Q030 Q060
0 #### +M Q003 Q033 Q046
.######## | Q007 Q010 Q034 Q036 Q042 Q050
.######## | Q019 Q023
.#### | Q004 Q008 Q028 Q048
.###### |S Q005 Q020 Q045
.########### |
####### | Q002
.######## | Q001 Q031 Q044
-1 ######### S+ Q059
.######## |T
.######## |
.#### | Q057
.### |
.# |
.# |
-2 . +
<less>|<frequ>
EACH "#" IS 3. EACH "." IS 1 TO 2

Form 3


PERSON - MAP - ITEM
<more>|<rare>
5 .# +
4 .### +
.#### T|
3 +
.####
.##
.####
.####
2 S+
######
.###### | Q022
|
|T
.## | Q028 Q050
.##### | Q019 Q021
1 .### + Q020
.######## |
.## |S Q018 Q060
M| Q042 Q043 Q058
.######## | Q007 Q011 Q057
.# | Q012 Q024 Q035 Q059
######## | Q004 Q008 Q036
0 .#### +M Q029 Q031 Q032
.######## | Q010 Q047
| Q034 Q046 Q048
.#### | Q005
########## | Q002 Q003 Q006 Q033 Q045 Q049
S|S Q009 Q025 Q026 Q041
######## | Q027 Q051
-1 .#### + Q023 Q030
| Q001 Q044
.### |
|T
|
. |
. |
-2 +
. T|
+
<less>|<frequ>
EACH "#" IS 5. EACH "." IS 1 TO 4

Form 4


PERSON - MAP - ITEM
<more>|<rare>
5 # +
.## |
4 +
.### |
3 .## T+
.### |
.# |
.### |
2 .## + Q036
.#### |T
.#### S|
.######## |
.#### | Q019 Q026 Q060
.### | Q032 Q043
1 .#### + Q011
.## |S Q018 Q042
######## | Q022 Q047 Q049
.### | Q006 Q044
.## M| Q007 Q035 Q059
######## | Q004 Q008 Q020 Q021 Q028
.### | Q010
0 ### +M Q009 Q030 Q034 Q058
.############ | Q012 Q033
| Q027 Q041 Q057
.######## | Q005 Q050
.############ | Q031 Q051
.###### | Q003
.##### |S Q025
-1 .#### S+ Q045
####### | Q001
.######## |
.# | Q029
| Q024 Q046 Q048
.# | Q002
. |T
-2 . +
| Q023
T|
-3 +
<less>|<frequ>
EACH "#" IS 4. EACH "." IS 1 TO 3
Form 5


PERSON - MAP - ITEM
<more>|<rare>
4 ## +
|
3 +
.# |
T|
.# |
.### |
2 .### +
.### |
|
.## |
.## S|
.## |
.### |T Q024
1 ### +
| Q035 Q043
.#### | Q009
.###### |S Q022 Q044 Q051
.#### | Q012 Q020
| Q008 Q036 Q047 Q058 Q060
.######## M| Q004 Q18 Q19 Q28 Q34 Q48 Q49
| Q006 Q010 Q029 Q030 Q042 Q057
0 .###### +M Q050
.############ | Q007 Q021 Q033
.#### | Q011 Q031 Q032
| Q003
| Q041 Q059
.########## |S Q005
| Q001 Q025 Q045
S| Q026
-1 .#### + Q046
.#### |T Q027
| Q023
.## |
.# | Q002
.# |
. |
-2 . T+
|
. |
-3 +
<less>|<frequ>
EACH "#" IS 5. EACH "." IS 1 TO 4
Form 6

```
          PERSON - MAP - ITEM
                <more>|<rare>
    5            .##  +
                      |
                      |
                      |
                      |
                      |
                      |
    4         .#####  +
                      |
                      |
                      |
                     T|
              .####   |
                      |
    3                 +
              .####   |
                      |
             .#####   |
                      |
               .###   |
                      |
    2          .#### S+
               .###   |
              #####   |
             .#####   |
               .###   |
              .####   |T Q019   Q020
               .###   |
    1           ####  +  Q042   Q046
         .#########   |
                ###   |S Q022   Q029   Q033   Q043   Q058
                .##   |  Q018   Q021   Q059
             .##### M|  Q028   Q041   Q045   Q047   Q049   Q060
               .###   |  Q012
            ########  |  Q024
    0           ###  +M Q002   Q057
          .########  |  Q003   Q009   Q050
               .###  |  Q001   Q011   Q030   Q034
           #########  |  Q007   Q008   Q010   Q025   Q032   Q035   Q051
              .#####  |  Q027   Q048
              .#####  |S Q005   Q006   Q023   Q026   Q044
      .#############  |  Q004
   -1         .##### S+
             .######  |  Q031
               .###  |T
               .###  |
               ####  |
                 .#  |
                  .  |  Q036
   -2             .  +
                     |
                  .  |
                    T|
                     |
                     |
                     |
   -3                +
              <less>|<frequ>
 EACH "#" IS 5. EACH "." IS 1 TO 4
```

Form 7

Appendix B1: Fuzzy Set Rules Generated in the Five-Variable ANFIS Model

#	Fuzzification rule for the input	Output
1	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf1) (1)
2	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf2) (1)
3	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf3) (1)
4	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf4) (1)
5	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf5) (1)
6	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf6) (1)
7	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf7) (1)
8	If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf8) (1)
9	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf9) (1)
10	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf10) (1)
11	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf11) (1)
12	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf12) (1)
13	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf13) (1)
14	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf14) (1)
15	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf15) (1)
16	If (input1 is in1mf1) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf16) (1)
17	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf17) (1)
18	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf18) (1)
19	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf19) (1)
20	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf20) (1)
21	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf21) (1)
22	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf22) (1)
23	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf23) (1)
24	If (input1 is in1mf2) & (input3 is in2mf1) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf24) (1)
25	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf25) (1)
26	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf26) (1)
27	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf27) (1)
28	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf28) (1)
29	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf1) & (input10 is in4mf2) & (input12 is in5mf2)	then (output is out1mf29) (1)
30	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf1)	then (output is out1mf30) (1)
31	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf1) & (input12 is in5mf2)	then (output is out1mf31) (1)
32	If (input1 is in1mf2) & (input3 is in2mf2) & (input6 is in3mf2) & (input10 is in4mf2) & (input12 is in5mf1)	then (output is out1mf32) (1)

Note. Input: 1 = Word count; 3 = Content word overlap (adjacent sentences); 6 = Logical connectives incidence; 10 = Verb incidence; 12 = Flesch-Kincaid grade level.

The rules will be activated when the right condition is met. For example, Rule #1 is stated as "If (input1 is in1mf1) & (input3 is in2mf1) & (input6 is in3mf1) & (input10 is in4mf1) & (input12 is in5mf1) then (output is out1mf1) (1). This indicates that if the magnitude of any input datum is low (mf1 stands for the low membership function), the predicted amount of output will be equal to membership function one. These rules are applied in the hidden layer of the ANFIS which is presented in in Figure 4.