Psychological Test and Assessment Modeling, Volume 57, 2015 (3), 301-337

Application of evolutionary algorithm-based symbolic regression to language assessment: Toward nonlinear modeling

Vahid Aryadoust[1]

Abstract

This study applies evolutionary algorithm-based (EA-based) symbolic regression to assess the ability of metacognitive strategy use tested by the metacognitive awareness listening questionnaire (MALQ) and lexico-grammatical knowledge to predict listening comprehension proficiency among English learners. Initially, the psychometric validity of the MALQ subscales, the lexico-grammatical test, and the listening test was examined using the logistic Rasch model and the Rasch-Andrich rating scale model. Next, linear regression found both sets of predictors to have weak or inconclusive effects on listening comprehension; however, the results of EA-based symbolic regression suggested that both lexico-grammatical knowledge and two of the five metacognitive strategies tested predicted strongly and nonlinearly listening proficiency ($R^2 = .64$). Constraining prediction modeling to linear relationships is argued to jeopardize the validity of language assessment studies, potentially leading these studies to inaccurately contradict otherwise well-established language assessment hypotheses and theories.

Keywords: evolutionary algorithm-based symbolic regression; lexico-grammatical knowledge; listening comprehension; metacognitive awareness; regression

[1] *Correspondence concerning this article should be addressed to:* Vahid Aryadoust, PhD, National University of Singapore, Centre for English Language Communication, 10 Architecture Drive, Singapore 117511; email: vahidaryadoust@gmail.com

Nomenclature			
EA	Evolutionary algorithm		
√x	Square root	sin	Sine
BEC	Business English certificate	sma	Simple moving average
cos	Cosine	tan	Tangent
EFL	English as a Foreign Language	tanh	Hyperbolic tangent
exp	Exponents	wma	Weighted moving average
log	Natural logarithm	δ	Vocabulary knowledge
MAE	Mean absolute error	ζ	Grammar knowledge
MALQ	Metacognitive awareness listening questionnaire	η/DA	Directed attention
MC	Multiple choice	θ_L	Academic listening proficiency
MSE	Mean squared error	κ/MT	Mental translation
n!	Factorial	λ/PK	Person knowledge
R	Correlation coefficient	M/PE	Planning & evaluation
R^2	R Square	ξ/PS	Problem solving

Linear regression is one family of linear mathematical functions that has been widely used in predictive modelling and achieved some degree of success in language assessment. Linear regression seeks to create mathematical solutions by which to predict output values from input values. A simple linear regression model can be mathematically expressed as follows:

$$\gamma = \alpha + \beta \chi_1 \quad (1)$$

, where

γ = output value or dependent variable,

χ = input value or independent variable;

β = slope, and

α = intercept.

The χ value is chosen to predict γ with as high accuracy as possible. In practice, however, some data points often fall far from the linear regression line. These data points are known as "outliers," and the presence of multiple outliers can affect the linearity of data and consequently worsen the model's fit and predictive power (Keith, 2006). As such, outliers are generally pruned in expert-informed predictive models (Hair, Black, Babin, & Anderson, 2010); otherwise, the yielded equation usually provides a relatively imprecise profile of the relationship between input and output data. This process destroys (valid) data, and although nonlinear data distant from the linear regression line may appear to be chaotic, random, or "useless," in reality it reflects the influence of networks of interrelated variables likely with meaningful interactions, which remain unexplored by linear models (Alamir, 1999).

Furthermore, destroying outlying data is typically not enough to render linear regression models highly accurate. A quick survey of the available literature shows that the average

precision of regression models, as indicated by their correlation coefficients, is approximately 0.40, suggesting an inherently nonlinear (or less linear) relationship between the variables examined. Linear models may be able to predict part of the data near the regression line, but will estimate the large proportion of data lying distant from the line with significant imprecision (Keith, 2006).

As a final issue, most studies applying linear regression do not attempt to test their postulated models with new data sets to examine whether their findings can be replicated (Keith, 2006). While this is not an intrinsic problem of regression modelling, lack of validation samples can question the credibility of the models yielded in linear regression analysis.

This set of methodological problems suggests that many of the conclusions drawn from linear regression studies in language assessment research may be oversimplified and imprecise. Rather than defining imprecise linear models or omitting data points that cause "error," researchers can use a flexible data analysis technique to pinpoint the structure of both the linear and nonlinear elements of the data and test it across an unseen sample (Koza, 2010). Evolutionary algorithm-based (EA-based) symbolic regression, also called symbolic function identification, seeks to identify influential independent variables by discovering the mathematical functions that fit the data (Fogel & Corne, 1998). Symbolic regression builds models using the symbolic functional operators selected by the researcher, and then by applying a genetic programing algorithm which results in the selection of input variables and a final set of models (Koza, 2010). To choose the best model from this set, the researcher can use a number of fit and importance statistics (Schmidt & Lipson, 2009).

To demonstrate the application of EA-based symbolic regression, this study uses data from a listening test, a vocabulary test, a grammar test, and the metacognitive awareness listening questionnaire (MALQ) (Vandergrift & Goh, 2012). Although both lexicogrammatical knowledge and metacognitive strategies have been posited to predict performance in listening comprehension tests (e.g., Buck, 2001; Goh, 2000; Vandergrift & Goh, 2012), neither set of variables has received enough empirical examination. Metacognition dimensions have been modelled as sole predictors of listening proficiency (Goh & Hu, 2014), but the exclusion of other important variables such as lexicogrammatical knowledge and the imposition of linear relationships between listening proficiency and metacognition has led to a low level of precision in prediction. Lexicogrammatical knowledge remains similarly underevaluated as a predictor of listening comprehension.

Evolutionary algorithm-based symbolic regression

> Evolution in natural systems gives rise to behaviors that are optimized […] In an analogy to natural systems, the evolutionary process itself can be modelled on a computer and applied to problems where heuristics are not available or generally lead to unsatisfactory results […] The advantages of this approach include its conceptual simplicity, broad applicability, ability to outperform classical optimization proce-

> dures on real-world problems, and ease of hybridization with existing methods and data structures [....]. (Fogel & Corne, 1998, pp. 19-20)

Unlike the linear and quantile regression methods used in predictive modelling, which assume and impose a priori mathematical structures on the data and use the data to estimate the parameters of models reflecting those structures, EA-based symbolic regression seeks to "breed" and "evolve" the most fitting mathematical solutions (Schmidt & Lipson, 2009) by searching among a population of potentially optimal solutions (Fogel & Corne, 1998). Both predictive modelling and symbolic regression involve a selection process to disqualify inappropriate mathematical solutions; however, the key advantage of symbolic regression over classical predictive models is that the selection and refinement of best solutions take place at the level of individual data points rather than the entire dataset. Using fit information from each data point, symbolic regression generates and evolves an ensemble of mathematical solutions, among which the researcher chooses the relatively better-fitting model by comparing the models' simplicity, precision, and fit (Schmidt & Lipson, 2010).

Data analysis in symbolic regression

Symbolic regression determines influential variables and estimates parameters in several stages, including operator selection, model solving, and model selection.

Operators

Symbolic regression initially yields a set of mathematical functions from the building blocks specified by the researcher (McRee, 2010). These building blocks consist of a number of mathematical functions, or *operators*, which are categorized into three groups on the basis of their complexity:

1. *Level 1 operators:* simple functions, such as summation (+), negation (-), and multiplication (*);
2. *Level 2 and level 3 operators:* more complex functions such as division (÷) and trigonometric and circular functions such as sine, cosine, and tangent functions; and
3. *Level 4 and level 5 operators:* highly complex functions such as exponential functions (e.g., natural logarithm and square root) and inverse trigonometric functions such as arcsine, arccosine, and arctangent (Schmidt & Lipson, 2010).

Model solving

Following operator selection, symbolic regression uses evolutionary algorithms (EAs) which imitate some of the mechanisms of Darwin's theory of evolution, such as mating, reproduction, and selection (Fogel, 1999; Schmidt & Lipson, 2008). EA-based symbolic regression begins with an *initialization* stage, in which the first generation of mathematical functions potentially fitting the data ("solutions") is randomly generated and varied; next, the fit of each of these potential solutions is evaluated with reference to one or more fit indices (Gwiazda, 2006). Based on the results of these fit indices, the fittest of

the solutions are chosen as "parents," and reproduce new solutions ("offspring") for use in successive iterations ("generations").

EA-based symbolic regression must maintain diversity in successive generations of solutions; otherwise the process will end in "premature convergence" (Schmidt & Lipson, 2010, 2011), stagnation based on the early convergence of solutions around a suboptimal solution point called a *local optimum*. This early end precludes significant improvements to the fit and precision of the solutions. Multiple methods have emerged to preclude EAs from premature convergence, among which Eureqa (Nutonian, n.d.a), the software package used in the present study, uses three: crossover, mutation, and age-fitness Pareto optimization.

Crossover is a process by which two or more parent solutions are taken to reproduce one or more child solutions, or "offspring" (Holland, 1975). In this study, Eureqa applies a one-point crossover: a random point on each parent solution is taken, and each parent function is divided into sections before and after the crossover point. These four sections are combined into two child solutions, one with a beginning from the first parent and an end from the second, and one with a beginning from the second parent and an end from the first. These two child solutions replace misfitting solutions from the previous generation (Koza, 2010; Schmidt & Lipson, 2008).

Mutation is another mechanism by which a single solution is partially or entirely altered to generate a potentially better solution (Gwiazda, 2006; Holland, 1975). Mutation is designed to be relatively rare: 50% of functions go through crossover, but only 1% go through mutation (Michalski, 2000). Mutation helps prevent premature convergence by increasing the diversity among solutions.

Better-fitting operators have a higher chance of being selected for both crossover and mutation (Schmidt & Lipson, 2010). Individual operators' probability of selection is a function of their "average progress" (Nicoară, 2009), which is estimated from the outset of the evolution process using the following formula:

$$\text{Progress}(x) = \frac{\sum_{i=1}^{\|x\|} \prod_i(x)}{\|x\|} \tag{2}$$

, where

$\Pi_i(x)$ = operator x's progress at the i^{th} application, and

$\|x\|$ = the frequency by which x is applied. (Nicoară, 2009, p. 91)

Following the application of every operator, the EA algorithm updates its probability of selection using the following formula:

$$x = \frac{\text{Progress}(x)}{\sum_{i=1}^{n} Progress(OP_i)} * (1 - n * \delta_x) + \delta_x \tag{3}$$

, where

OP_i = an operator in the x operator's class, and

$\delta_x \in (0,1)$ = the minimum index of "selection probability" for the operator. (Nicoară, 2009, p. 91)

All operators are initially assigned the same probability, and this probability is continually updated according to the progress of the operator.

A third mechanism used in Eureqa to avoid premature convergence is age-fitness Pareto optimization, in which evolving populations of solutions are selected based on two dimensions: age – how long the solution has been present in the population – and fit to the data (Aryadoust, 2015a; Schmidt & Lipson, 2010). Age-fitness Pareto optimization chooses the youngest and fittest solutions, and, over generations, attempts to minimize error, improve fit, and maximize the models' predictive power (Schmidt & Lipson, 2011). Schmidt and Lipson (2010) showed that this optimization technique outperforms other available techniques such as deterministic crowding and age-layered optimization. Figure 1 presents a schematic representation of this optimization technique, plotting the error of measurement (the inverse of fit) against age.

When multiple computers are connected to maximize the processing speed of model estimations, some computer packages, including Eureqa, use an *island model* for parallelization, in which large numbers of potential solutions are partitioned into several independent islands evolving independently, each worked on by an independent computer. At particular intervals, some of the solutions on each island randomly migrate to other islands on other computers to preclude premature convergence. This process often yields multiple optimal solutions, the best of which is chosen on the basis of its simplicity, fit in the validation sample, sensitivity of parameters, and consistency with the postulated theoretical frameworks (Guyon & Elisseeff, 2003).

Unlike traditional algorithms, in which only one best solution is maintained, EAs maintain an entire population of the fittest solutions (Koza, 2010). This feature confers an important advantage over the traditional approaches: it helps EAs avoid local optimum traps. Local optima are the solutions that are optimal among a subpopulation of neighboring solutions, but which are not optimal among all possible solutions in the population (Koza, Keane, & Streeter, 2003). When algorithms become trapped in local optima, they likely miss better solutions available among other subpopulations.

Figure 2 summarizes the aforementioned mechanisms of EA-based symbolic regression, as applied in the Eureqa software package.

Model selection

Choosing the optimal model in symbolic regression entails examining existing models based on three primary criteria: the fit metrics of the cross-validation subsample or subsamples, sensitivity statistics (see Methodology), and the size and complexity of the model. This section discusses some of these criteria with reference to the Eureqa software package, although the discussion can be generalized to other software.

To assess whether the best-fitting solutions are generalizable to untested data, EA-based symbolic regression divides the sample into training and cross-validation subsamples.

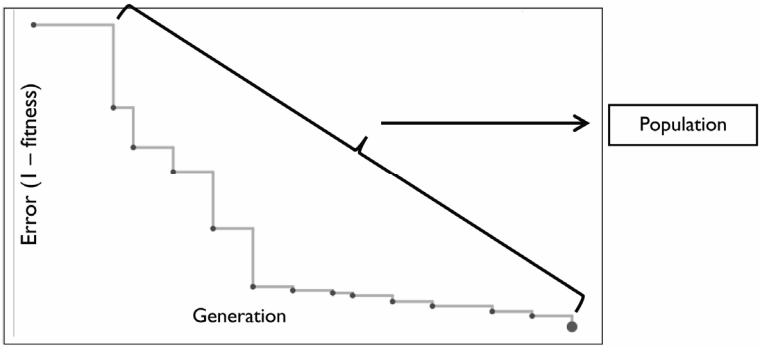

Figure 1:
Illustration of the age-fitness Pareto optimization (modified from Schmidt & Lipson, 2010). As new generations are created, the error of measurement reduces and fit improves.

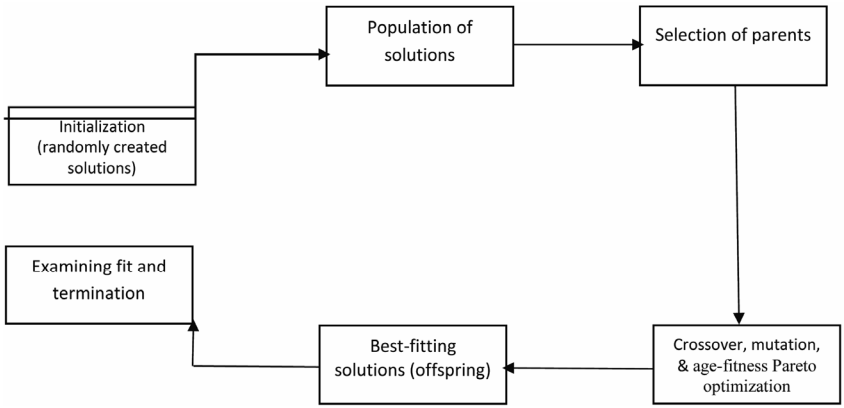

Figure 2:
Schematic representation of EA-based symbolic regression as applied in Eureqa software package. As the search progresses, Eureqa attempts to find a simpler solution with better fit (per the selected error metric) and discard more complicated and less accurate solutions.

The first subsample is used to generate accurate solutions, and the second subsample is used to assess the performance of the chosen optimal solution. Cross-validation helps prevent overfitting, which occurs when a model describes error of measurement instead of the constituent structure of the data (Leinweber, 2007). Since the training stage algorithm attempts to maximize the fit of the solutions to the training sample data, the yielded solutions can only be evaluated by verifying them across an unknown sample (Leinweber, 2007).

Eureqa further evaluates the solution verified in cross-validation against a second held-out sample. Assessing the precision of the solution across a second untested sample

provides further evidence supporting the cross-validated solution, allowing researchers to posit stronger data-driven theories. In addition, this validation process can also help researchers choose among competing solutions which fit the test data equally well. This supplementary cross-validation process has very rarely been applied in language assessment research, and predictive studies replicating previous research have often neglected to verify the parameters in their own models (Zhang, Goh, & Kunnan, 2014).

Fit statistics. Multiple fit statistics are used to assess the efficacy of the cross-validated solutions, as follows:

1) Mean absolute error (MAE): assuming that error of measurement follows a double-exponential distribution, MAE estimates the difference between predicted and observed values. The closer the MAE to zero, the higher the precision.
2) Mean squared error (MSE): like MAE, MSE estimates the difference between the predicted and observed values. Unlike MAE, however, MSE assumes that error of measurement is normally distributed. Lowe MSE indices indicate higher precision.
3) Correlation coefficient (R): ranging between 0 and 1, the R index quantifies the correlation between observed and predicted values. Values above 0.7 indicate significant correlation between model-estimated and actual output.
4) R^2 goodness of fit: The R^2 index indicates the percentage of output (values of the dependent variable) that can be explained by the input or independent variables. It ranges between 0 and 1, with values closer to 1 indicating higher predictive power.

Eureqa assigns each operator in the solutions a numerical value indicating its complexity. For example, addition and negation have a value of 1, whereas logistic and step functions have a value of 4. The total complexity of each solution is the sum of the complexity values of the operators used in that solution. As less complex models with low errors of measurement are desirable, sometimes the researcher has to make a trade-off between complexity and fit by choosing less complex models over more complex models with slightly better fit statistics (Schmidt & Lipson, 2010). The researcher can also base this decision on the sensitivity of each input variable, estimated according to its frequency of occurrence across all solutions, as well as the sensitivity or relative impact of each input variable on the output within each model.

Rationale for using EA-based symbolic regression in language assessment

EA-based models are particularly well-suited for language assessment applications for several important reasons. First, like linear regression, EA-based models generate a quantitative estimate of the relationship between input and output variables and are therefore appropriate for parameter optimization (Koza, Keane, & Streeter, 2003). However, unlike linear regression models, EA-based techniques can yield more precise models without deleting data points falling farther away from the linear regression line, allowing these models to better fit the full range of data. Nonlinearity and certain types of deviation from normality represent archetypical aspects of language test performance data, and considering these qualities in data modelling allows for a more accurate and

nuanced understanding of language test performance. In addition, unlike linear regression, EA-based techniques allow a vast range of mathematical functions that cannot be provided in traditional methods (Koza et al., 2003).

Listening comprehension and statistical prediction

Studies show that listening comprehension involves multiple sub-processes, and that listeners use both bottom-up and top-down comprehension strategies to make sense of perceived words (Baghaei & Aryadoust, 2015; Dunkel, Henning, & Chaudron, 1993; Vandergrift & Goh, 2012). Bottom-up comprehension entails constructing the smaller units of aural stimuli, such as sounds, into larger units, such as words, and then into the grammatical relations between words (Aryadoust, 2015b). By contrast, in top-down processing listeners use their knowledge to construct interpretations that are complete and meaningful (Wagner, 2004). The result of the application of these knowledge sources to an aural input is a set of mental representations of the intended message, called propositions. Different endowments of knowledge contribute to differences in the propositions generated by listeners. When accurate propositions are formed and comprehension is achieved, listeners are equipped to respond accurately, in the form of written or spoken responses to the oral stimuli (Bejar et al., 2000).

In both top-down and bottom-up listening, mental lexicon and grammatical knowledge are key elements of successful comprehension (Baghaei & Carstensen, 2013). Deficiencies in vocabulary repertoire hamper word recognition and hence comprehension, and incomplete grammatical knowledge hinders listeners' attempts to juxtapose and make sense of the string of words in their mind (Dunkel et al., 1993). This suggests that listeners' lexico-grammatical resources can predict their listening performance (Buck, 2001). However, although most researchers agree on the effect of lexico-grammatical resources on language comprehension (Kintsch, 1998), some early attempts at explaining comprehension downplayed the role of these resources (McKeown, Beck, Omanson, & Perfetti, 1983); in addition, virtually no research has investigated the effect of lexico-grammatical resources in *listening* comprehension specifically, and most existing hypotheses are drawn from reading studies (Buck, 2001).

Metacognitive strategies

In an assessment context, listeners also use metacognitive listening strategies to aid in comprehension (Goh, 2000). Metacognition refers to learners' awareness of their own cognitive strategy use, and monitoring and adjusting these strategies to achieve a certain objective (Flavell, Miller, & Miller, 1993). Metacognition consists of knowledge of the task (learners' knowledge concerning the requirements of learning tasks as well as the factors that determine their difficulty), oneself (learners' awareness of their own confidence, anxiety, and reactions to learning requirements), and strategy (learners' knowledge of the learning techniques they would use to achieve their learning objectives) (Goh & Hu, 2014).

Vandergrift, Goh, Mareschal, and Tafaghodtari (2006) operationalized metacognitive strategies in listening comprehension as a multidimensional construct comprising 21 items loading on to five factors: directed attention (DA), mental translation (MT), planning and evaluation (PE), problem solving (PS), and person knowledge (PK). They used this model to develop an instrument called the "metacognitive awareness listening questionnaire" (MALQ).

Vandergrift and Goh (2012) categorized the first four dimensions of the construct as representing learners' attempts to regulate the comprehension process, and the fifth, PK, as representing learners' knowledge of themselves. Similarly, Goh and Hu (2014) detailed the five dimensions as follows:

Directed attention strategies are needed to focus attention on the task; mental translation strategies help learners translate what they hear into their first language; planning and evaluation strategies assist listeners to plan and prepare for listening, as well as evaluating their performance after listening; and problem-solving strategies enable learners to make inferences when they are unable to hear or understand a certain word. Person knowledge reveals what learners know about themselves as L2 listeners, particularly in terms of their confidence. (Goh & Hu, 2014, p. 259)

Despite its defined theoretical structure, MALQ has yielded varying degrees of correlation with listening test performance in different studies. For example, total MALQ scores explained 20% ($R^2 = 0.20$) of the variance observed in learners' test scores in Goh and Hu's (2014) research, 13% of the variance ($R^2 = 0.13$) in Vandergrift et al.'s (2006) study, and 15% of the variance ($R^2 = 0.15$) in Zeng's (2012) research. The present study seeks to provide more empirical research into the predictive power of MALQ in EFL contexts.

The literature reviewed suggests that listening proficiency is associated with lexico-grammatical knowledge and metacognitive strategies. This relationship can be mathematically expressed as follows:

$$\text{Listening} = f(\text{DA, MT, PE, PS, PK, vocabulary, grammar}) \quad (5)$$

The present study seeks to find the fittest function that can optimally map listening onto its theoretical correlates: DA, MT, PE, PS, PK, vocabulary, and grammar.

METHODS

Data source and instruments

This study uses item-level data from the administration of multiple psychometric instruments to 250 first- and second-year Chinese English as a foreign language (EFL) college students aged between 17 and 23 ($M = 19.73$; $SD = 0.90$). All students consented to participate in the study, and after their participation, each student received a personalized test performance report. The assessment and report were partly designed to help participants prepare for their English exams in college, which contain lexico-grammatical and listening sections.

The study uses data on participant performance on a vocabulary knowledge test, a grammatical knowledge test, and the metacognitive awareness listening questionnaire (MALQ) (Vandergrift et al., 2006).

The listening test chosen for the study is a sample business English certificate (BEC) listening paper which is developed by Cambridge English. BEC is widely taken by Chinese students seeking employment in industries that require English language proficiency. The sample test contained 30 test items: items 1 to 12 were fill-in-the-gap items assessing test takers' ability to understand phone conversations or phone messages; items 13 to 22 were fill-in-the-gap items based on five short recordings describing a problem that occurred; and items 23 to 30 were multiple choice (MC) items based on a recorded interview with a manager of a restaurant. Test takers were given some time to read the questions before listening to the text. Consistent with the BEC listening test requirements, each audio recording was played twice.

The vocabulary knowledge test consisted of 30 MC items. Each item included an underlined target word whose synonym was to be chosen among the available options. The vocabulary chosen ranged from easy to difficult to discriminate among participants. The grammatical knowledge test comprised 15 MC items chosen from sample paper-based TOEFL (test of English as a foreign language) tests. The test items measured participants' familiarity with a range of grammatical structures, such as independent clauses, infinitives, and prepositions. The results of the vocabulary and knowledge tests were provided to the students as diagnostic feedback.

MALQ is a psychometric instrument that measures learners' awareness of their own use of cognitive strategies during listening comprehension. It includes five subscales: directed attention (four items), mental translation (three items), person knowledge (four items), planning and evaluation (five items), and problem solving (five items). MALQ uses a six-point Likert scale ranging from *strongly disagree* to *strongly agree.*

Preliminary psychometric validation

Participants' performance on the tests was initially subjected to Rasch measurement for psychometric validation, and to estimate participants' linearized measures on each subscale (Kubinger, Rasch, & Yanagida, 2011; Rasch, 1960/1980). *WINSTEPS* computer package, Version 3.80 (Linacre, 2015a), was used to perform the analyses. Data from the grammar, vocabulary, and listening tests were subjected to dichotomous Rasch model analysis, and each MALQ subscale was independently analysed by the Rasch-Andrich rating scale model[2] (Andrich, 1978). The fit of the data to the models was estimated using infit and outfit mean square (MNSQ) statistics. The former index is an inlier sensitive index suitable for capturing erratic response patterns of the items near test takers'

[2] The model is expressed as:

Pr $\{X_{ni} = x\} = \exp\sum_{k=0}^{x}(B_n - (\delta_n - \tau_k)) \Big/ \sum_{j=0}^{m}\exp\sum_{k=0}^{j}(B_n - (\delta_n - \tau_k))$, where δ_i is item difficulty for item *i*, and τ is the threshold of the scale common between all test items.

ability level; the latter index, on the other hand, is outlier-sensitive. Generally, MNSQ values between 0.6 and 1.4 would indicate good model fit (Bond & Fox, 2007). Participant and item fit statistics were subjected to Rasch model analysis, and participant ability measures in logits (log-odds units) were read to IBM SPSS and Eureqa for linear regression and EA-based symbolic regression analysis, respectively. I used scale scores (person parameters from the Rasch measurement) in the regression analysis.

EA-based symbolic regression analysis

Eureqa Software Package

This study uses Eureqa Version 0.99 beta (Schmidt & Lipson, 2013). Developed by Hod Lipson at the Computational Synthesis Lab at Cornell University and nicknamed "the robot scientist," Eureqa is one of the few software packages to estimate symbolic regression (Lin, 2009). Eureqa has achieved success in multiple scientific disciplines, such as psychology (e.g., Slater et al., 2013), neurology (e.g., Pardoe, Abbott, & Jackson, 2012), and physics (e.g., Potomkin, Gyrya, Aranson, & Berlyand, 2013).

Target expression and building blocks

To perform EA-based symbolic regression, I initially indicated target expressions or the type of model to explore in the analysis. I modeled listening proficiency as a function of vocabulary knowledge, grammar knowledge, and five primary metacognitive strategies, as follows:

$$\theta_L = f(\delta, \zeta, \kappa, \mu, \lambda, \xi, \eta) \quad (6)$$

where

θ_L (theta) = academic listening proficiency,

δ (delta) = vocabulary knowledge,

ζ (zeta) = grammar knowledge,

η (eta) = directed attention,

κ (kappa) = mental translation,

λ (lambda) = person knowledge,

μ (mu) = planning and evaluation, and

ξ (xi) = problem solving.

Next, I chose the mathematical "building blocks" (operators such as addition, negation, and multiplication) to use in the search. As previously discussed, building blocks may be simple, moderately complex, or highly complex. To maintain the simplicity of the final solutions, I opted for simple and moderately complex operators (Schmidt & Lipson, 2009, 2010); since the data appeared to be nonlinear in most (but not all) areas, I used

trigonometric, exponential, and linear building blocks to capture the complexity and dynamics of the data. Table 1 presents the building blocks chosen for the analysis, comprising: (a) basic operators with complexity coefficients of 1 or 2, such as constant, addition, subtraction, multiplication, and division; (b) trigonometric operators with complexity coefficients of 3, such as sine and cosine; and (c) exponential operators with complexity coefficients of 4, such as exponents and natural logarithms.

Two rounds of EA-based symbolic regression analysis were conducted. To conduct the initial analysis, Eureqa was integrated with Amazon EC2 Secure Cloud to accelerate the search for well-fitting solutions, thereby using 126 additional computer cores. The best fitting solution generated in the first phase was chosen for use in the second round of analysis. In the second round, Eureqa initiated search by using the chosen solution, which improved the fit of the solutions significantly.

Variable preparation

Figure 3 presents the distribution of the input and output data. The data includes participants' ability measures on the vocabulary test, grammar test, and MALQ subscales as estimated by Rasch measurement, plotted against participants' linearized listening ability measures. Each figure includes a trendline and R^2 indices. Directed attention (η) has the largest R^2 index (0.270), and planning and evaluation (μ) has the smallest index (.00006). It should be noted that the R^2 indices are computed using the correlation coefficients and, as such, they are likely to change when their joint effect is estimated in regression analysis.

To rescale the data into a common metric with the same offset, they were normalized using the following normalization formula:

$$\text{Normalized variable} = (\text{variable} - \text{offset}) / \text{scale} \quad (7)$$

(Nutonian, n.d.b)

Normalization improves the performance of Eureqa and the fit of the solutions (Schmidt & Lipson, 2010). Following Rasch measurement, the data were normalized and divided into a training batch (70%; $n = 175$) and a validation batch (30%; $n = 50$).

Progress and performance of the solutions

Two statistics were used to assess the progress and performance of the list of solutions: a *stability index* describing for how many generations the fittest solutions have not improved, and a *maturity index* describing how many generations ago the fittest solutions improved. Both statistics vary between 0 and 1; when both stability and maturity are near 1, the algorithm is unlikely to improve further. At this juncture, the fittest solution was chosen and *seeded* (used as the "base and prior solution" for a new search). Seeding the new search with prior optimal solutions significantly accelerates the new analysis and renders it substantially more precise (Schmidt & Lipson, 2010).

Following the selection of generated solutions, multiple fit statistics were used to examine the performance of these solutions, including R, R^2, MAE, and MSE.

Table 1:
Mathematical Building Blocks Used

Building block (operator)	Formula	Sign	Complexity
Basic			
Constant	c	c	1
Addition	+	+	1
Input variable	x	x	1
Subtraction	-	-	1
Multiplication	×	×	1
Division	÷	÷	2
Trigonometric			
Sine	$\frac{\text{Opposite side of triangles}}{\text{Hypotenus side of triangles}} = \frac{a}{h}$	sin	3
Cosine	$\frac{\text{Adjacent side of triangles}}{\text{Hypotenus side of triangles}} = \frac{b}{h}$	cos	3
Tangent	$\frac{\text{Adjacent side of triangles}}{\text{Opposite side of triangles}} = \frac{a}{b}$	tan	4
Squashing			
Hyperbolic tangent	$\frac{1-e^{-2x}}{1+e^{-2x}}$	tanh	4
Exponential			
Exponents	$y = a^x$	exp	4
Natural logarithm	$\log_e x$	log	4
Factorial	$n!$	factorial	4
Power	$f(x) = a^x$	^	4
Square root	$f(x) = \sqrt{x}$	sqrt	4
History			
Simple moving average	$sma = \frac{V_n + V_{n-1} + \ldots V_{n-(n-1)}}{n}$ *	sma	4
Weighted moving average	$wma = \frac{mV_n + (m-1)V_{n-1} + \ldots + 2V_{n-(m+2)} + V_{(n-m+1)}}{n}$	wma	4

Note: Complexity1 = simple; complexity 2 & 3 = medium complexity; complexity 4 & 5 = complex.

* V is the amount of the variable in time n and later.

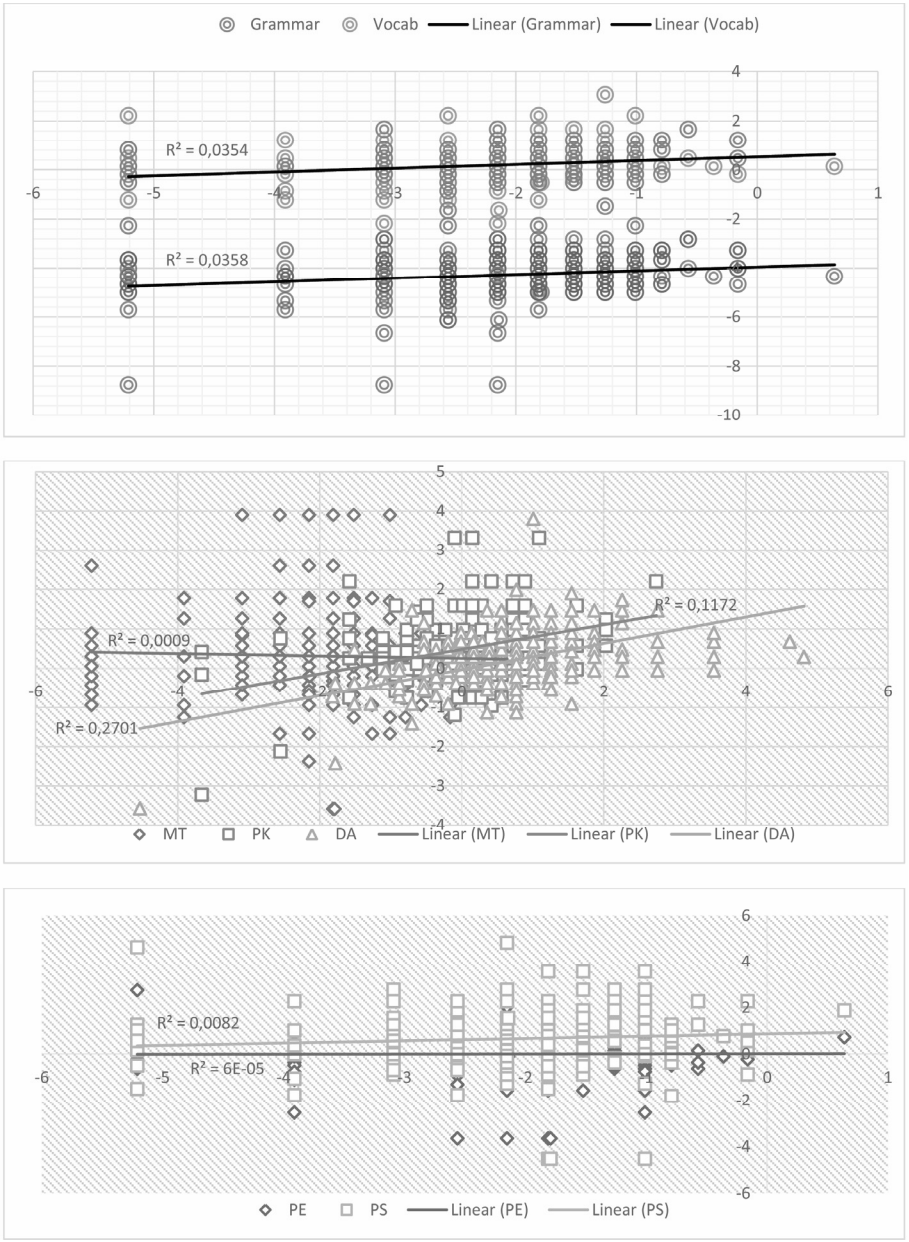

Figure 3:
Scatterplot of listening proficiency scores and input variables, representing nonlinear patterns.

Finally, positive and negative percentages and magnitudes for each variable in the chosen models were estimated. Sensitivity for the function Y = *f*(x,z,…) is estimated as follows:

$$\left|\frac{\partial_Y}{\partial_x}\right| . \frac{std(x)}{std(Y)} \tag{8}$$

, where

$\frac{\partial_Y}{\partial_x}$ = partial derivative of Y with respect to x,

std (x) = standard deviation of x, the input variable, and

std (Y) = standard deviation of Y, the output variable.

Sensitivity indicates the direction and strength of the correlation between input and output variables. For example, if this index is .6, then all else equal, a one-unit increase in x will result in a .6-unit increase in Y. This number is the sum of the magnitude of positive and negative effects: for example, if x positively impacts Y 75% of the time with magnitude 2 and negatively impacts Y 25% of the time with magnitude 1, the net effect is 1.25 (2*.75 – 1*.25 = 1.5 – .25).

Linear regression model

Linear regression analysis was performed on IBM SPSS computer package. EA-based symbolic regression was compared with linear regression in order to compare these methods' robustness. The input variables in the linear regression models were assessed for multicollinearity. Multicollinearity could arise from high correlations between the independent variables, as a result of which one of the variables would appear to be insignificant in the regression model and contribute no share of variance. When degrees of multicollinearity are high, the estimated model can be suboptimal, leading to the automatic elimination of one of the variables causing multicollinearity.

There are several ways to check multicollinearity; according to Callaghan and Chen (2008, p. 2), "the best practice for assessing linear dependencies in model data" includes examining eigenvalues and their corresponding condition values. Other (preliminary) methods include inspection of VIF (variance inflation factor) and tolerance (1/VIF), variance-decomposition proportions (VDPs), correlations between independent variables, and correlations between regression coefficients. As a rule of thumb, eigenvalues close to zero along with high condition indices indicate the presence of multicollinearity. Finally, as mentioned by Callaghan and Chen (2008, p. 3), "[a] suggested procedure for diagnosing collinearity is a high condition index, which is also associated with a high variance-decomposition proportion for two or more regression coefficient variances."

It should be noted that as in EA-based symbolic regression analysis, 70% of the data was used for model building (training) and 30% for validation. Regression coefficients, R^2, and adjusted R^2 (which is R^2 adjusted for the number of input variables) were estimated.

RESULTS

Preliminary analysis: Rasch measurement and multicollinearity check

To examine the tests' psychometric features and dimensionality, test data were subjected to Rasch model analysis. The data in all subscales and tests fitted the Rasch model reasonably well, providing evidence for psychometric reliability of the data (see Aryadoust, 2015b, for further information).

Since the study seeks to compare EA-based symbolic regression with linear regression, the possibility of multicollinearity in the data was assessed. First, the correlation between learners' ability measures, as estimated by the tests and MALQ subscales, were evaluated. Table 2 presents bivariate correlation coefficients for the tests and the MALQ subscales and Rasch model item reliability and separation indices alongside their person/item infit and outfit MNSQ statistics (Linacre, 2015b). Except for the high correlation between grammar and vocabulary knowledge tests (R = .899), coefficients were either negligible, weak (0.1 to 0.3), or moderate (0.31 to 0.7). In addition, the variables' skewness and kurtosis coefficients all fell between -2 and +2, indicating univariate normality. In addition, Rasch model item reliability and separation indices show that the tests reliably distinguish test takers of different ability levels, indicating minimum amount of measurement error. Rasch measurement and RSM average infit and outfit indices also indicate that items and test takers, on average, had a sufficiently good fit to the model.

Table 2:
Correlation Coefficients of the Input and Output Variables alongside Rasch Model Reliability and Fit Indices

	Grammar	MT	PE	PK	PS	DA	Vocabulary
Grammar	1						
MT	-.048	1					
PE	.005	.326**	1				
PK	.074	.351**	.342**	1			
PS	.062	.409**	.588**	.536**	1		
DA	.007	.323**	.435**	.504**	.520**	1	
Vocabulary	.899**	-.050	.005	.075	.063	.009	1
Listening	.189**	-.030	.008	-.021	.090	-.045	.188**
Rasch model reliability	0.98	0.98	0.88	0.98	0.88	0.98	0.98
Rasch model separation	6.55	7.72	2.66	7.62	2.66	7.63	6.55
Mean infit MNSQ for items	1.00	1.01	1.02	1.06	1.01	1.00	1.01
Mean infit MNSQ for persons	1.01	0.99	1.01	1.07	1.00	0.99	1.01
Mean outfit MNSQ for items	1.01	0.98	1.02	0.98	1.01	1.05	1.00
Mean outfit MNSQ for persons	1.00	1.02	1.01	0.96	1.00	1.05	1.00

Note: ** $p < 0.005$.

Linear regression model

Multiple linear regression models were generated and assessed through backward and stepwise methods of parameter estimation. To examine multicollinearity in each model, VIF and tolerance indices were estimated. As Table 3 demonstrates, the VIF indices for Grammar and Vocabulary in the regression models where these variables are present were extremely high and their tolerance statistics were close to zero, indicating a high possibility of multicollinearity. Additionally, the constant (intercept) had a large standard error of measurement, suggesting inflation of the variance. On the other hand, the VIF and tolerance indices were back to normal when either of Vocabulary and Grammar variables were excluded from the model. A follow-up analysis was conducted to determine what degree of multicollinearity might have affected the data.

Table 3:
Results of Regression Modeling Including Coefficients, t Values, and Collinearity Statistics

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	*p* value	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	5.336	10.240		.521	.603		
	Grammar	1.689	2.288	1.387	.738	.461	.001	911.122
	MT	-.049	.069	-.050	-.707	.480	.780	1.283
	PE	-.061	.103	-.046	-.591	.555	.627	1.595
	PK	-.129	.114	-.089	-1.135	.257	.628	1.592
	PS	.181	.082	.198	2.214	.028	.482	2.073
	DA	-.070	.112	-.049	-.623	.534	.621	1.610
	Vocabulary	-1.453	2.284	-1.195	-.636	.525	.001	911.686
2	(Constant)	-1.172	.343		-3.416	.001		
	Grammar	.235	.076	.193	3.076	.002	.982	1.018
	MT	-.045	.069	-.046	-.650	.516	.787	1.271
	PE	-.059	.102	-.045	-.572	.568	.627	1.594
	PK	-.128	.114	-.088	-1.126	.261	.629	1.591
	PS	.178	.081	.195	2.181	.030	.484	2.065
	DA	-.075	.112	-.053	-.671	.503	.624	1.602
3	(Constant)	5.168	10.223		.506	.614		
	Grammar	1.648	2.284	1.353	.722	.471	.001	910.286
	MT	-.053	.069	-.054	-.765	.445	.786	1.272
	PK	-.127	.114	-.088	-1.117	.265	.629	1.590
	PS	.161	.074	.176	2.169	.031	.585	1.709
	DA	-.082	.110	-.058	-.742	.459	.642	1.558
	Vocabulary	-1.411	2.280	-1.160	-.619	.537	.001	910.814

	(Constant)	-2.210	.084		-26.381	.000		
	MT	-.048	.068	-.049	-.700	.485	.793	1.261
	PK	-.125	.114	-.086	-1.105	.270	.629	1.589
*4	PS	.158	.074	.173	2.136	.034	.587	1.704
	DA	-.087	.110	-.061	-.794	.428	.645	1.551
	Vocabulary	.234	.076	.192	3.068	.002	.982	1.019
	(Constant)	-2.212	.084		-26.447	.000		
	PK	-.137	.112	-.094	-1.216	.225	.642	1.558
5	PS	.145	.072	.159	2.030	.043	.623	1.606
	DA	-.094	.109	-.066	-.861	.390	.650	1.538
	Vocabulary	.239	.076	.196	3.152	.002	.990	1.010
	(Constant)	-2.200	.083		-26.654	.000		
	PK	-.167	.107	-.115	-1.565	.119	.712	1.405
6	PS	.124	.067	.136	1.845	.066	.712	1.405
	Vocabulary	.243	.076	.200	3.210	.002	.994	1.006
	(Constant)	-2.242	.078		-28.642	.000		
7	PS	.068	.057	.074	1.191	.235	.996	1.004
	Vocabulary	.238	.076	.196	3.142	.002	.996	1.004
8	(Constant)	-2.200	.070		-31.567	.000		
	Vocabulary	.244	.076	.201	3.226	.001	1.000	1.000

Note. * Optimal model.

As Table 4 presents, the eigenvalues for the independent variables were larger than zero when either Vocabulary or Grammar was excluded from the analysis, indicating serious multicollinearity of these two variables. Several condition indices of Vocabulary also fall above 30, suggesting high multicollinearity. Additionally, the matrix of VDPs (right hand side) gives the variables whose regression coefficients have been degraded by multicollinearity; that is, there are multiple values greater than 0.50 which is the accepted index for multicollinearity. Vocabulary is, therefore, involved in severe multicollinearity and DA causes moderate levels of multicollinearity.

The optimal model among those presented in Table 3 is Model 4, according to its R^2 value and the number of variables predicting listening comprehension. In this mode, Vocabulary and PS were found to predict listening proficiency ($p < 0.05$), whereas Grammar knowledge was not ($R^2 = 0.06$, adjusted $R^2 = 0.041$, $p = 0.010$). This model yielded a constant of -2.210 ($p < 0.001$) and β coefficients of .158 ($p = 0.034$) for PS and .234 ($p = 0.002$) for Vocabulary, indicating that if Vocabulary and PS increase by one unit, then on average Listening comprehension scores will increase by 15% and 23%, and thus giving the following solution:

$$\theta_L = -2.210 + 0.158 \times \xi + .234 \times \delta \quad (9)$$

Table 4:
Eigenvalues, Condition Indices, and VPs

Model	Dimension	Eigenvalue	Condition Index	Variance-decomposition proportions (VDPs)							
				Constant	Grammar	MT	PE	PK	PS	DA	Vocab
1	(Constant)	3.542	1.000	.00	.00	.02	.01	.02	.02	.02	.00
	Grammar	1.556	1.509	.00	.00	.02	.12	.00	.01	.05	.00
	MT	1.016	1.868	.00	.00	.02	.00	.00	.00	.00	.00
	PE	.652	2.330	.00	.00	**.88**	.05	.00	.01	.09	.00
	PK	.574	2.483	.00	.00	.00	.47	.09	.02	.38	.00
	PS	.378	3.063	.00	.00	.04	.01	**.62**	.06	.47	.00
	DA	.282	3.547	.00	.00	.01	.34	.26	**.88**	.00	.00
	Vocabulary	**2.216E-005**	**399.827**	**1.00**	**1.00**	.01	.00	.00	.00	.01	**1.00**
2	(Constant)	3.393	1.000	.00	.00	.02		.03	.03	.02	.00
	Grammar	1.222	1.667	.00	.00	.07		.01	.02	.13	.00
	MT	1.016	1.828	.00	.00	.02		.00	.00	.00	.00
	PK	.644	2.296	.00	.00	**.79**		.02	.00	.23	.00
	PS	.383	2.976	.00	.00	.08		.31	.27	**.62**	.00
	DA	.342	3.149	.00	.00	.01		**.63**	**.67**	.00	.00
	Vocabulary	**2.218E-005**	**391.169**	**1.00**	**1.00**	.01		.00	.00	.00	**1.00**
3	(Constant)	2.886	1.000	.03		.04		.04	.04	.04	.01
	MT	1.060	1.650	.07		.07		.00	.00	.05	**.63**
	PK	.721	2.001	.44		.00		.00	.00	.27	.25
	PS	.638	2.126	.03		**.83**		.03	.00	.12	.10
	DA	.357	2.844	.31		.06		.06	**.61**	.48	.01
	Vocabulary	.339	2.918	.12		.00		**.87**	.34	.05	.00
4	(Constant)	2.563	1.000	.05				.05	.05	.05	.01
	PK	1.010	1.593	.05				.00	.01	.10	.71
	PS	.720	1.887	.45				.00	.00	.25	.28
	DA	.368	2.639	.31				.02	**.64**	**.54**	.00
	Vocabulary	.339	2.749	.14				**.92**	.30	.07	.00
5	(Constant)	2.188	1.000	.09				.08	.08		.03
	PK	.927	1.536	.00				.02	.03		**.94**
	PS	.542	2.009	**.87**				.05	.23		.03
	Vocabulary	.342	2.530	.04				**.85**	.66		.00
6	(Constant)	1.589	1.000	.21					.20		.10
	PS	.889	1.337	.04					.11		**.89**
	Vocabulary	.522	1.744	.75					**.70**		.01
7	(Constant)	1.209	1.000	.40							.40
	Vocabulary	.791	1.237	.60							**.60**

Equation 9 was tested across the validation sample, yielding an extremely low R^2 of 0.004.

On the other hand, when Vocabulary was excluded and Grammar was included, the results differed. The regression model with Grammar knowledge as one of the independent variables was also significant (R^2 = 0.062, adjusted R^2 = 0.039, p = 0.073), and yielded a constant of -1.172 (p = 0.001) and β values of .235 (p < 0.01) for Grammar and .178 (p = 0.030) for PS, thus giving the following solution:

$$\theta_L = -1.410 + 0.178\times\xi + 0.178\times\zeta \quad (10)$$

Solutions 9 and 10 indicate that Vocabulary and Grammar (alongside PS) do exert a significant impact on listening comprehension, but their effects are masked by multicollinearity when both are simultaneously included in the equation.

EA-based symbolic regression

Eureqa integrated with Amazon EC2 Secure Cloud quickly reached a maturity index of 95%, generating multiple solutions. Table 5 presents these solutions from the cross-validation or testing sample, plus the solutions' fits as generated in the validation phase. It should be noted that the training samples usually had a sufficiently high fit and therefore presenting the fit of training models is unnecessary; on the other hand, the fit of the model to the cross-validation data is of paramount importance as it can determine the actual performance of the models/solutions. According to the fit statistics and complexity indices (R = 0.775; R^2 = 0.597[3]; MSE = 0.118; MAE: 0.579; complexity = 27), the fittest solution is mathematically expressed as follows:

$$\theta_L = 10\times\text{sma}(\delta, 37)^2 + 2.288133097\times\tanh(2.444468119 + \text{sma}(\mu, 6) + \text{wma}(\zeta, 4)) \quad (11)$$

The solution uses addition, multiplication, exponent, simple moving average (sma), and weighted moving average (wma) operators, and includes vocabulary knowledge (δ), grammar knowledge (ζ), and planning and evaluation (μ). Four out of five MALQ subscales (ξ, η, κ, and λ) did not have a marked influence on listening proficiency in this model. The results presented with regard to equation (11) are referring to solution 3 that can be found in Table 5.

The other finding from Table 5 is that high complexity does not necessarily yield good fit to the data. For example, the most highly complex model is mathematically expressed as follows:

$$\theta_L = 0.2102301606\times\text{sma}(\zeta, 35) + 0.7382124855\times\text{sma}(\delta, 22)\times\tanh(5.785283279+\mu) - 15.97487679\times\text{wma}(\kappa, 34) - 0.7382124855\times\text{step}(\text{wma}(\kappa, 33))\times\tanh(5.785283279 + \mu) \quad (12)$$

[3] It should be noted that the R^2 values for EA-based models and linear regression models are directly comparable.

Table 5:
Competing Solutions with Varying Degrees of Fit Generate in Round One
(Cross-Validation Sample Results)

Model	R	R^2	MSE	MAE	Complexity
1. θ_L = 0.2102301606×sma(ζ, 35) + 0.7382124855×sma(δ, 22)×tanh(5.785283279 + μ) - 15.97487679×wma(κ, 34) - 0.7382124855×step(wma(κ, 33))×tanh(5.785283279 + μ)	0.746	-31.05	9.472	2.482	47
2. θ_L = 0.2166821275×sma(ζ, 35) + 0.7382124855×sma(δ, 22)×tanh(4.251668213 + μ) - 14.96266279×wma(κ, 34) - 0.7382124855×step(wma(κ, 33))	0.749	-27.207	8.334	0.749	39
3. *θ_L = 10×sma(δ, 37)^2 + 2.288133097×tanh(2.444468119 + sma(μ, 6) + wma(ζ, 4))	**0.775**	**0.597**	**0.118**	**0.575**	**27**
4. θ_L = 3.460672997×sma(δ, 37) + 2.531999428 × tanh (2.40368155 + sma(μ, 8) + wma(ζ, 4))	0.739	0.540	0.135	0.739	24
5. θ_L = 10×sma(δ, 37)×wma(δ, 37) + 2.134266438×tanh(ζ + μ)	0.718	0.503	0.146	0.718	19
6. θ_L = 3.550279297×sma(δ, 37) + 2.309212838×tanh(ζ + μ)	0.660	0.430	0.168	0.660	16
7. θ_L = 0.3986282781×wma(ζ, 10) + ξ×δ×sma(μ, 9)	0.528	0.225	0.228	0.528	15
8. θ_L = 0.396138302×sma(ζ, 5) - 0.2056761054×μ×δ	0.422	0.152	0.250	0.422	12
9. $\theta_L = \frac{7.108555685}{(\zeta - \delta - \mathrm{sma}(\mu, 9))}$	0.128	0.011	0.292	0.128	11
10. θ_L = 0.3680446258×ζ - 0.09528322737×μ×δ	-0.111	0.1683	0.328	0.444	9
11. $\theta_L = \frac{7.632352241}{(\zeta - \mu - \delta)}$	-0.170	0.073	0.345	0.459	8
12. θ_L = 0.4225062325×ζ×μ×δ	-9.097	0.176	2.983	1.538	7
13. $\theta_L = \frac{10 \times \mu \times \delta}{\zeta}$	-11.79	0.068	3.779	1.658	6
14. θ_L = 10×ζ×μ×δ"	-901.38	0.176	266.63	10.834	5

Note: δ = vocabulary knowledge; ζ = grammar knowledge; μ = planning and evaluation; κ = mental translation. * chosen for the second round.

However, Solution 9 fits the data poorly (R = 0.746; R^2 = -31.05; MSE = 9.472; MAE: 2.482), although it includes a wider range of input variables: ζ, δ, μ, and κ.

Finally, a sensitivity analysis was carried out to examine the relative impact of each variable in the solution on the output variable, θ_L. Table 6 presents the sensitivity and positive / negative indices of δ, μ, and ζ under "Best model_Round 1." Sensitivity percentages of δ and μ are 0.376, 0.283, respectively, with positive sensitivity percentages of 100% and negative sensitivity percentages of 0.00%, indicating that an increase in δ and μ will always lead to an increase in listening proficiency and never to a decrease. ζ displays a slightly different pattern: it has a markedly higher impact on listening proficiency (sensitivity = 4.824) and 93% of the time it positively correlates with listening proficiency, but 7% of the times it negatively correlates with listening proficiency.

In the second round of analysis, Solution 3 (θ_L = 10×sma(δ, 37)^2 + 2.288133097× tanh(2.444468119 + sma(μ, 6) + wma(ζ, 4))) was used as the initiating equation of the EA algorithm. This algorithm iterated through numerous generations and reached a maturity of 95%. Figure 4 presents 17 solutions with three graphs: observed and predicted values, including trend lines, the slopes of which represent R^2 values; output and row, which plot observed data against model-estimated values; and "age-fitness Pareto optimization", which plots error of measurement (1 – fit) against generation (age).

Table 7 reports the fit of the models to the cross-validation sample in the second round and complements Figure 4, presenting the fit statistics and complexity of the solutions from the second round. The two most complex models fit the data poorly; for example, fit of the most complex solution (complexity = 23) is: R = 0.752; R^2 = -24.423; MSE = 7.512; MAE: 2.261. The best solution, which combines optimal fit to the validation

Table 6:
Sensitivity Analysis of the Variables in the Best Solutions Estimated in Rounds One and Two (Cross-Validation Sample Results)

Model and variable	Sensitivity	% Positive	Positive magnitude	% Negative	Negative magnitude
Best model_Round 1					
δ	0.376	100%	0.376	0.00%	0.00
μ	0.283	100%	0.283	0.00%	0.00
ζ	4.826	93%	5.122	7.00%	1.0029
Best model_Round 2					
ζ	0.2538	92%	0.273	8.00%	0.044
μ	2.778	100%	2.778	0.00%	0.00
ξ	0.337	0.00%	0.00	100%	0.337
δ	5.169	100%	5.169	0.00%	0.00

Table 7: Competing Solutions with Varying Degrees of Fit Generate in Round Two (Cross-Validation Sample Results)

Solution	R	R^2	MSE	MAE	Complexity
1. $\theta_L = \frac{(63.11113365\times \text{sma}(\kappa, 20) - 10\times \text{sma}(\mu, 28) - 15.14590125\times \text{sma}(\delta, 37))}{\zeta - 2.224635872}$	0.752	-24.423	7.512	2.261	23
2. $\theta_L = \frac{(66.08053004\times \text{sma}(\kappa, 20) - 10\times \text{sma}(\mu, 28) - 10\times \text{sma}(\delta, 37))}{\zeta - 1.97358747}$	0.727	-26.759	8.202	2.364	21
3. * $\theta_L = \frac{(\xi - 13.88244441\times \text{sma}(\mu, 28) - 19.3979306\times \text{sma}(\delta, 37))}{\zeta - 2.54512072}$	**0.800**	**0.640**	**0.107**	**0.240**	**20**
4. $\theta_L = \frac{(\xi - 10\times \text{sma}(\mu, 28) - 17.91068312\times \text{sma}(\delta, 37))}{\zeta - 2.485238406}$	0.785	0.614	0.113	0.252	18
5. $\theta_L = \frac{(-10\times \text{sma}(\mu, 28) - 15.46442487\times \text{sma}(\delta, 37))}{\zeta - 2.485995364}$	0.761	0.557	0.130	0.266	17
6. $\theta_L = \frac{(-10\times \text{sma}(\mu, 28) - 10\times \text{sma}(\delta, 37))}{\zeta - 2.227888843}$	0.747	0.450	0.162	0.284	15
7. $\theta_L = \frac{(-1.214292923 - 10\times \text{sma}(\mu, 28) - 10\times \text{sma}(\delta, 37))}{\zeta - 2.527440201}$	0.720	0.443	0.164	0.282	16
8. $\theta_L = \frac{(3.9039382e - 253\times \mu - 10\times \text{sma}(\delta, 34))}{\zeta - 2.088696584}$	0.632	0.374	0.184	0.333	13
9. $\theta_L = \frac{(-0.1860286571\times \mu - 10\times \text{sma}(\delta, 34))}{\zeta - 2.080214824}$	0.631	0.374	0.184	0.329	14
10. $\theta_L = \frac{10\times \text{sma}(\zeta, 5)}{(25.7418284 + \delta - \mu)}$	0.423	0.156	0.249	0.381	11
11. $\theta_L = 0.3916419231\times \text{sma}(\zeta, 5) - 0.2448197734\times \mu\times \delta$	0.409	0.153	0.250	0.380	12
12. $\theta_L = \frac{10\times \zeta}{(26.83158053 + \mu - \delta)}$	0.161	-0.053	0.311	0.431	8
13. $\theta_L = -1.520453269 - 0.000210038612\times \zeta\times \mu\times \delta$	-0.176	-0.035	0.306	0.427	9
14. $\theta_L = \frac{-0.7771937288\times \mu\times \delta}{\zeta - 1.645699436}$	-0.068	-0.041	0.307	0.425	10
15. $\theta_L = 0.4223048209\times \zeta\times \mu\times \delta$	0.176	-9.096	2.983	1.538	7
16. $\theta_L = \frac{10\times \mu}{(\delta - \zeta)}$	0.062	-11.150	3.590	1.606	6
17. $\theta_L = 10\times \zeta\times \mu\times \delta$	0.176	-901.388	266.639	10.834	5

Note: δ = vocabulary knowledge; ζ = grammar knowledge; μ = planning and evaluation. * The best solution.

Model
Observed vs. Predicted
Output vs. Row
Error/Complexity Pareto
Model: θ = (a*sma(κ, 20) - b*sma(μ, 28) - c*sma(δ, 37))/ζ - d
Text: θ = (63.1111*sma(κ, 20) - 10*sma(μ, 28) - 15.1459*sma(δ, 37))/ζ - 2.22464
Model: θ = (a*sma(κ, 20) - b*sma(μ, 28) - c*sma(δ, 37))/ζ - d
Text: θ = (66.0805*sma(κ, 20) - 10*sma(μ, 28) - 10*sma(δ, 37))/ζ - 1.97359
Model: θ = (ζ - a*sma(μ, 28) - b*sma(δ, 37))/ζ - c
Text: θ = (ζ - 13.8824*sma(μ, 28) - 19.3979*sma(δ, 37))/ζ - 2.54512
Model: θ = (ζ - a*sma(μ, 28) - b*sma(δ, 37))/ζ - c
Text: θ = (ζ - 10*sma(μ, 28) - 17.9107*sma(δ, 37))/ζ - 2.48524
Model: θ = (-10*sma(μ, 28) - a*sma(δ, 37))/ζ - b
Text: θ = (-10*sma(μ, 28) - 15.4644*sma(δ, 37))/ζ - 2.486
data
model
frontier
model

Model: θ = (-a - b*sma(μ, 28) - c*sma(δ, 37))/ζ - d
Text: θ = (-1.21429 - 10*sma(μ, 28) - 10*sma(δ, 37))/ζ - 2.52744
Model: θ = (-10*sma(μ, 28) - a*sma(δ, 37))/ζ - b
Text: θ = (-10*sma(μ, 28) - 10*sma(δ, 37))/ζ - 2.22789
Model: θ = (-a*μ - b*sma(δ, 34))/ζ - c
Text: θ = (-0.186029*μ - 10*sma(δ, 34))/ζ - 2.08021
Model: θ = (a*μ - b*sma(δ, 34))/ζ - c
Text: θ = (3.90394e-253*μ - 10*sma(δ, 34))/ζ - 2.0887
Model: θ = a*sma(ζ, 5) - b*μ*δ
Text: θ = 0.391642*sma(ζ, 5) - 0.24482*μ*δ
data
model
frontier

Model: $\theta = 10*\mathrm{sma}(\zeta, 5)/(a+\delta-\mu)$
Text: θ = 10*sma(ζ, 5)/(25.7418 + δ - μ)
Model: $\theta = -a*\mu*\delta/\zeta - b$
Text: θ = -0.777194*μ*δ/ζ - 1.6457
Model: $\theta = -a-b*\zeta*\mu*\delta$
Text: θ = -1.52045 - 0.000210039*ζ*μ*δ
Model: $\theta = 10*\zeta/(a+\mu-\delta)$
Text: θ = 10*ζ/(26.8316 + μ - δ)
Model: $\theta = a*\zeta*\mu*\delta$
Text: θ = 0.422305*ζ*μ*δ
data
model
frontier

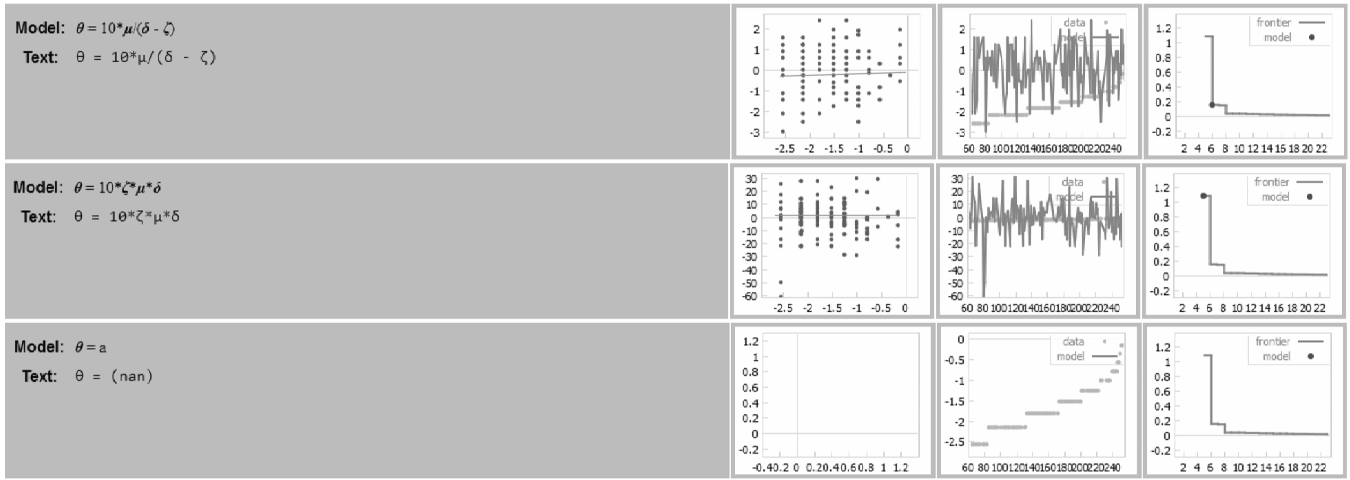

Figure 4:
Graphic representation of the models estimated in the second round of EA-based symbolic regression.

sample (R = 0.800; R^2 = 0.640; MSE = 0.107; MAE: 0.240) and relatively low complexity (20), is expressed as follows:

$$\theta_L = \frac{\left(\xi - 13.88244441 \times \mathrm{sma}\left(\mu, 28\right) - 19.3979306 \times \mathrm{sma}\left(\delta, 37\right)\right)}{\zeta - 2.54512072} \tag{13}$$

This solution uses subtraction, multiplication, division, and SMA functions to map ζ, δ, μ, and ξ onto listening proficiency. Relative to its parent, Solution 11, Solution 13 has evolved to comprise an additional input variable (planning and evaluation, ξ) and has better fit to both training and cross-validation data.

Table 6 presents the sensitivity statistics of Solution 3 under "Best model_Round 2." Sensitivity indices of δ and μ are 2.778 and 5.169, respectively, with positive sensitivity percentages of 100% and positive magnitudes of 2.778 and 5.169, respectively, suggesting that an increase in δ and μ will always lead to an increase in listening proficiency. Similarly, ζ has a high impact on listening proficiency (sensitivity = 0.254. positive sensitivity percentage = 92%). In sum, a nonlinear and relatively simple solution consisting of lexico-grammatical tests and two MALQ dimensions (ζ, δ, μ, and ζ) fitted the data well in the second round of EA-based symbolic regression, suggesting that listening proficiency can be predicted much more accurately by EA-based symbolic regression than linear regression models.

DISCUSSION

This study set out to investigate the potential of EA-based symbolic regression in a listening assessment context. EFL learners' listening proficiency was modeled as a function of their vocabulary and grammar knowledge as well as their metacognitive strategies. The study's findings are discussed in terms of their implications for model-building in listening comprehension theory and predictive modeling in language assessment.

Implications for language assessment research

Predictive modeling in language assessment has been led by theories. Researchers often subject various amounts of data to psychometric validation, postulating a linear relationship between input and out variables. However, the cognitive and predictive theories applied in studies rarely indicate the functional form of the mathematical relationship between input and output variables. Indeed, although linear models can sometimes yield good fit, it would be highly counterintuitive to propose simple linear relationships among cognitive assessment data.

To obtain maximum value from the data and to search for unified models in language assessment, nonlinear science and optimization techniques such as genetic and evolutionary algorithms offer great potential. As such, one of the applications of EA-based symbolic regression is to reassess those linear predictive models which have been refuted

due to their poor or lack of fit. This line of research would quite likely rectify or improve the data-driven theories generated in previous studies.

In addition, when the data contains outliers, the slope parameter in linear regression models is significantly affected. Researchers can examine the magnitude of this impact by fitting the regression model under two conditions: (1) leaving out outliers, and (2) including outliers. When outliers influence the model, the R^2 value decreases greatly under condition 2, whereas if the outlier exerts no sizeable influence, the R^2 value will have no significant change. It is important to note that large data sets are largely robust to outliers, but if a large data set – like that in the present study – deviates from linearity and is rife with outliers, then linear models would not be able to capture the complexity of the data. Nevertheless, the research should initially ensure that the data is not actually influenced by error of measurement; otherwise, it is likely that the predictive techniques would only lead to models fitting the measurement or sampling error rather than the data/variables of interest. In the present study, the psychometric quality of the data was examined through Rasch measurement, and as such, the non-linear model applied does not model error of measurement. It is suggested that researchers apply such psychometric validation methods before using any predictive model (see Aryadoust & Liu, 2015).

Another implication of the study is that multicollinearity has to be taken into account when using linear regression models. Researchers using linear regression should initially establish the lack high dependence between independent variables. If high dependence is found (e.g., high correlation statistics), it is important to carry out tolerance and VIF analysis to ascertain whether the variables are multicollinear. If there is a large amount of multicollinearity in data, researchers should remove the effect prior to linear modeling; although there are various methods to cancel out multicollinearity effects, most often researchers have to remove one of the independent variables, because other methods have proved to be less effective. However, this can provide a "contorted" model which might be (highly) different from otherwise well-established theoretical models. As the present study has shown, one of the most useful techniques to address this problem is the application of non-linear models, which do not fall prey to interdependence of independent variables.

Implications for listening comprehension

Among the solutions generated by EA-based symbolic regression, Solution 3 was found to be both well-fitting and relatively simple. This model included vocabulary knowledge, grammar knowledge, planning and evaluation (PE), and problem solving (PS). Its high R^2 (.64) and correlation coefficient (.775) suggest a strong association between lexico-grammatical knowledge and metacognitive strategies and listening comprehension, which is markedly higher than the associations previous research has reported (Goh & Hu, 2014; Vandergrift & Goh, 2012).

The linear regression models, by contrast, included only problem solving and grammar knowledge, and yielded two models with significant or near-significant *p* values but non-substantive predictive power. As a result, a researcher using linear regression would tend

to argue that vocabulary knowledge and metacognitive strategies do not predict listening performance.

The present study is one of the first studies to use both lexico-grammatical knowledge and metacognitive strategies to predict listening comprehension. The study's findings corroborate and improve the precision of previous studies on metacognitive strategies in listening (e.g., Goh & Hu, 2014; Goh, 2000) and also establish a nonlinear relationship between EFL listening and lexico-grammatical knowledge.

Metacognitive strategies

Problem solving

This study contradicts Goh and Hu (2014) in finding a negative correlation between problem-solving (PS) strategies and listening comprehension. However, the substantive meaning of this finding may agree with Goh and Hu's predictions. Goh and others (Goh & Hu, 2014; Vandergrift & Goh, 2012) have postulated an inverse relationship between lexico-grammatical knowledge and PS:

$$x = a \times z,\ a < 0; \tag{14}$$

this proportionality also holds in $y = f(x, z, \ldots)$

x = lexico-grammatical knowledge,

z = PS, and

y = listening.

Test takers employ PS as a compensatory strategy when they are unable to understand some parts of the message due to their limited vocabulary and grammar resources (Goh & Hu, 2014). When learners have extensive lexico-grammatical knowledge, they do not need to avail themselves of compensatory strategies such as PS. Since previous research on the MALQ has not included vocabulary or grammar tests or lexico-grammatical knowledge variables, PS appears to positively influence listening proficiency. Including these variables, on the other hand, allows the results to capture the incomplete lexico-grammatical knowledge and accordingly inefficient cognitive strategies implied by PS (Goh & Hu, 2014). Goh and Hu's research may have found an inverse relationship between PS and listening comprehension had they included variables measuring lexico-grammatical proficiency.

Planning and evaluation

The current study also contradicts Goh and Hu (2014) in finding that planning and evaluation (PE) strategies – which aid listeners in planning for listening and appraising their performance after listening – positively correlated with listening proficiency. PE can significantly impact listening test performance, particularly when test takers have received test taking strategy training (Vandergrift, 2004). Explicit instructions provided by teachers help learners plan for test questions and activate related test taking knowledge. This strategy seems to be highly useful in "while-listening-performance" tests such as

BEC. These tests require students to multitask: they must plan to read the items once prior to the listening experience; listen to the text and simultaneously reread the test items; and supply or choose the best answer to the test items (Aryadoust, 2013). PE's high sensitivity index in Solution 13 suggests that test takers used this strategy effectively to enhance their test performance.

Mental translation, directed attention, person knowledge

Mental translation (MT), directed attention (DA), and person knowledge (PK) had no statistical significance in predicting listening proficiency. MT is particularly useful for low-ability learners or beginners, who must translate the oral text into their first language as a decoding strategy (Vandergrift, 2004). However, the learners in the present study had received English instruction and test-taking training for several years, and were therefore less likely to need these strategies.

As in Goh and Hu (2014), DA lacked statistical significance in predicting listening proficiency. It appears likely that BEC's policy of providing test takers a second chance to listen to the recorded stimuli encourages test takers not to attempt to listen harder or concentrate more, even if they missed some part of the text on first listen. Furthermore, since the learners had received training on BEC, they are likely to have known that anything they missed could be heard a second time. This suggests that the utility of DA varies with the nature and requirements of the task.

PK is a scale which measures students' awareness of their listening level and anxiety in general, which does not seem to relate closely to test taking.

Lexico-grammatical knowledge

Grammar and vocabulary knowledge were the most important predictors of listening proficiency. This finding supports the proposition that lexico-grammatical knowledge helps listeners execute comprehension strategies, and that deficiencies in vocabulary or grammar resources adversely affect learners' listening comprehension (Field, 2009). The study also reinforces previous comprehension research by psycholinguists finding that lexico-grammatical knowledge plays an important role in listening test taking (e.g., Kintsch, 1998), and contradicts studies that have found these knowledge resources to be less important (e.g., McKeown et al., 1983).

In reading comprehension, word meaning is retrieved during comprehension from the mental lexicon and is then associated with the perceived words (Buck, 2001; Buck & Tatsuoka, 1998). Syntactic rules are then infused into the string of words to recode the decoded message and make sense of them (Buck, 2001). Similar mechanisms seem to be operational during listening comprehension, the output of which is associated with lexico-grammatical resources.

One of the gaps in listening theory is that the types of relationships between different listening components have generally not been articulated. This has led to vague and purely descriptive models with no mathematical expressions, which contrast with the

prescriptive and powerful models developed in other fields of science, such as physics. In language assessment, however, no predictive study has been able to fully replicate the results of previous studies, partly due to the mathematical models used and partly due to the numerous unknown influences on human behavior data, such as differences between cultures and education systems (see Bodie, 2013; Bodie, Worthington, Imhof, & Cooper, 2008). Valuable attempts have been made by, for example, Bachman and Palmer (1996) to unpack influential parameters in test taking behavior. However, these attempts have generally not been treated rigorously in subsequent research. For example, Bachman and Palmer's strategic competence framework makes no mention of the nature of the mathematical relationships between test performance and its predictors, but validation researchers have generally assumed these relationships to be linear.

A word of caution seems appropriate regarding the use and implications of predictive modeling. Predicting cognitive skills such as listening using explanatory variables such as lexico-grammatical knowledge and metacognitive strategies does not indicate that listening comprehension can be viewed as a process based merely on these mechanisms. Rather, it suggests that EFL learners' vocabulary and grammar knowledge and test-taking strategies are nonlinearly associated with their listening performance; strategic EFL learners with large lexico-grammatical repertoires and good test-taking strategies are more successful in listening comprehension tests, and the mathematical representation of this warranted assumption is Solution 13. To improve the precision of this solution, other variables which have been shown to affect listening comprehension, such as certain demographic variables and working memory capacity, should be identified and added to the equation.

In addition, the sample in the present study comprised Chinese students. Based on their first language, these students might have different strategies than students with other languages as their first language. Depending on the culture or first language, different models and correlations might be found for different samples. Thus, the current findings might not be generalizable to other samples. Previous studies where such relationships were not found used multiracial and multiethnic participants (e.g., Vandergrift et al., 2006), and this diversity in the sample might be a significant intervening factor. Future research into such highly diverse sample should factor in ethnicity or first language before drawing any inference from data analysis procedures.

Conclusion

The present study reported on one of the first applications of EA-based symbolic regression to language assessment problems, which achieved high accuracy and yielded a theoretically valid solution. EA-based symbolic regression can achieve much higher precision if influential variables in a predictive model are reliably identified, operationalized, and measured. Future research into predictive modeling in language assessment can apply EA-based symbolic regression in highly researched areas which, nevertheless, have returned inconclusive results, such as the predictive validity of second language tests.

Exploring the potential of and nonlinear science holds great potential for language assessment. The commonly applied static models do not appear to explain the non-static universe of language learning and assessment.

Acknowledgement

I would like to express my thanks to the two reviewers of *The Psychological Test and Assessment Modeling* for providing constructive feedback on earlier versions of this article. I would also like to thank Ms Wang Yan for her assistance in data collection.

References

Alamir, M. (1999). Optimization based nonlinear observers revisited. *International Journal of Control, 72*(13), 1204–1217.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrica, 43, 561-573.

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

Aryadoust, V. (2015a). Predictive modelling in Coh-Metrix-based writing research: A proposal for genetic-programming-based symbolic regression. In V. Aryadoust & J. Fox (Eds.) *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim*. Newcastle: Cambridge Scholars Publishing.

Aryadoust, V. (2015b). Fitting a mixture Rasch model to EFL listening tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing, 15*(3), 216-238.

Aryadoust, V., & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing, 24,* 35-58.

Bachman, L., F., & Palmer, A., S. (1996). *Language testing in practice.* New York: Oxford University Press.

Baghaei P., & Aryadoust, V. (2015). Modeling test method effect with a multidimensional Rasch model. *International Journal of Testing*, *15*(1), 71-87.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, 18(5). Retrieved from http://pareonline.net/getvn.asp?v=18&n=5

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. MS-19). Princeton, NJ: ETS.

Bodie, G. D. (2013). Issues in the measurement of listening. *Communication Research Reports, 30*(1), 76-84.

Bodie, G. D., Worthington, D. L., Imhof, M., & Cooper, L. (2008). What would a unified field of listening look like? A proposal linking past perspectives and future endeavors. *International Journal of Listening, 22,* 103-122.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. London, UK: Erlbaum.

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.

Callaghan, K. J., & Chen, J. (2008). Revisiting the collinear data problem: An assessment of estimator 'Ill-Conditioning' in linear regression. *Practical Assessment Research & Evaluation*, *13*(5). Available online: http://pareonline.net/getvn.asp?v=13&

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of a listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal, 77*(2), 180-191.

Flavell, J.H., Miller, P.H., & Miller, S.A. (1993). *Cognitive development* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Fogel, D. B., & Corne, D. W. (1998). An introduction to evolutionary computation for biologists. In D. B. Fogel (Ed.), *Evolutionary computation: The fossil record* (pp. 19-38). Piscataway, NJ: IEEE Press.

Fogel, L. J. (1999). *Intelligence through simulated evolution: Forty years of evolutionary programming.* New York: John Wiley.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28,* 55–75.

Goh, C. C. M., & Hu, G. (2014). Exploring the relationship between metacognitive awareness and listening performance with questionnaire data. *Language Awareness*, *23*, 255-274.

Guyon I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Gwiazda, T. G. (2006). *Genetic algorithms reference Vol.1 crossover for single-objective numerical optimization problems*. Poland: Lomianki.

Hair J. F., Black W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis. A global perspective* (7th Ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Holland, J. (1975). *Adaptation in natural and artificial systems.* University of Michigan Press, Ann Arbor, Michigan.

Holland, J. (1975). *Adaptation in natural and artificial systems.* Ann Arbor, Michigan: University of Michigan Press.

Keith, T. (2006). *Multiple regression and beyond*. Boston: Pearson.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge: Cambridge University Press.

Koza, J. (2010). Human-competitive results produced by genetic programming. *Genetic Programming and Evolvable Machines, 11*, 251–284.

Koza, J. R., Keane, M. A., Streeter, M. J. (2003). *Evolving inventions*. Scientific American.

Kubinger, K. D., Rasch, D., & Yanagida, T. (2011). A new approach for testing the Rasch model. *Educational Research and Evaluation, 17*(5), 321-333.

Leinweber, D. J. (2007). Stupid data miner tricks. *The Journal of Investing, 16*, 15–22.

Lin, E. (2009). *Eureqa, the robot scientist (w/ Video)*. Retrieved from the Physics Organization website at http://phys.org/news179394947.html

Linacre, J. M. (2015a). WINSTEPS: Rasch model computer programs [computer program]. Chicago, IL: Wisteps.com.

Linacre, J. M. (2015b). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

McKeown, M.G., Beck, I. L., Omanson, R.C., & Perfetti, C.A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Literacy Research, 15*(3), 3-18.

McRee, R. K. (2010). *Symbolic regression using nearest neighbor indexing.* Proceedings of the 12th annual conference companion on Genetic and evolutionary computation, GECCO 10, 1983–1990.

Michalski, R. S. (2000). Learnable evolution model: Evolutionary processes guided by machine learning. *Machine Learning, 38*, 9–40.

Nicoară, E S. (2009). Mechanisms to avoid the premature convergence of genetic algorithms. *Economic Insights: Trends and Challenges, 61*(1), 87–96.

Nutonian (n.d.a). *Eureqa desktop quick start guide*. Retrieved from http://www.nutonian.com/products/eureqa-desktop/quick-start/

Nutonian (n.d.b). *Eureqa desktop user guide*. Retrieved from http://formulize.nutonian.com/documentation/eureqa/user-guide/prepare-data/

Pardoe, H. R., Abbott, D. F., & Jackson, G. D. (2012). Sample size estimates for well-powered cross-sectional cortical thickness studies. *Human Brain Mapping, 34*(11), 3000-3009.

Potomkin, M., Gyrya, V., Aranson, I., & Berlyand, L. (2013). Collision of microswimmers in a viscous fluid. *Physical Review: statistical, nonlinear, and soft matter physics, 87*(5), 3005 1-8.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. [Reprint, with a Foreword and Afterword by Benjamin D. Wright. Chicago: University of Chicago Press, 1980.]

Schmidt, M., & Lipson, H. (2008). Coevolution of fitness predictors. *IEEE Transactions on Evolutionary Computation*, *12*(6), 736–749.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81–85.

Schmidt, M., & Lipson, H. (2010). Comparison of tree and graph encodings as function of problem complexity. In Thierens, D., Beyer, H., Bongard, J., Branke, J., Clark, J. A., Cliff, D., C., et al. (Eds.), *GECCO 07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, (vol. 2) (pp. 1674–1679). London. ACM Press.

Schmidt, M., & Lipson, H. (2011). Age-fitness Pareto optimization. *Genetic Programming Theory and Practice, 8*, 129–146.

Schmidt, M., & Lipson, H. (2013). Eureqa [Computer package], Version 0.99 beta. www.nutonian.com

Slater, M., Rovira, A., Southern, R., Swapp, D., Zhang, J. J., Campbell, C., & Levine, M. (2013). Bystander responses to a violent incident in an immersive virtual environment. *PLoS ONE, 8*(1), e52766.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*(1), 3-25.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge.

Vandergrift, L., Goh, C. C. M, Mareschal, C., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire (MALQ): Development and validation. *Language Learning, 56*, 431–462.

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 2*, 1-25. Ann Arbor, MI: University of Michigan English Language Institute.

Zeng, Y. (2012). *Metacognition and self-regulated learning (SRL) for Chinese EFL listening development* (Unpublished doctoral dissertation). Nanyang Technological University, Singapore.

Zhang, L. M., Goh, C. M. C., & Kunnan, A. J. (2014). Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach. *Language Assessment Quarterly, 11*(1), 76-120.