Determining Factors Influencing Listening Test Item Difficulty and Predicting Reading Proficiency

Vahid ARYADOUST
Centre for English Language Communication, National University of Singapore
Outline

• Two studies
  1. Exploring oral text-related variables influencing item difficulty $\rightarrow$ ANFIS
  2. Classifying readers based on lexico-grammatical knowledge
Study 1: Predicting Item Difficulty

• How do you develop an in-class reading test?
• How do you make your tests easy or difficult for students?
• How do you design an easy/difficult reading or listening test item?
• Some items are easier and some are more difficult...
• Study 1: predicting listening test item difficulty.
Predicting Item Difficulty

- Item difficulty is affected by various factors.
- To design items, those factors must be considered.
- Need for theories to predict item difficulty.
- Developing a predictive theory of item difficulty.
- Listening test development has been “thin on substantive theory” (Stenner, Stone, & Burdick, 2011, p. 3)
Predicting Listening Test Item Difficulty 1

- TOEFL listening minitalks:
- negation
- referential words
- rhetorical organization
- fronted structures
- lexical overlap
- vocabulary
- length, and
Predicting Listening Test Item Difficulty 2

- **word-level** variables, e.g., vocabulary
- **sentence-level** variables, e.g., dependent clauses
- **discourse-level** variables, e.g., questions or statements; and
- **task-processing** variables, e.g., inference-making.
- 41% of variance in item difficulty (Kostin, 2004)
Predicting Listening Test Item Difficulty 3

- Rule-space methodology: Buck and Tatsuoka (1998) extracted 15 attributes and 14 interactions (in 5 categories)
- **Task identification attributes:** the ability to identify the task by determining what type of information to search for in order to complete it;
- **Context attributes,** such as the density of information in the text;
- **Information location attributes,** such as the ability to use previous items to help locate information;
- **Information processing attributes,** such as the ability to process very fast text automatically; and
- **Response construction attributes,** such as the ability to construct a response quickly and efficiently.
Predicting Listening Test Task Difficulty

- Surveys of teachers and students:
- Listener-related, speaker-related, and material- and medium-related factors (Boyle, 1984).
- Goh (1999): text-related (e.g., speech rate), listener-related (e.g., prior knowledge), speaker-related (e.g., accent), and environmental variables (e.g., physical condition).
Statistical Tools to Predict Item Difficulty

• T-tests? → limited
• Regression models (Freedle & Kostin, 1990, 1993, 1999; Ginther, 2000)
• Rule space methodology (Buck & Tatsuoka, 1998)
• The fusion model (Aryadoust, 2011; Y.-W., Lee & Sawaki, 2009; Sawaki, Kim, & Gentile, 2009).
• Artificial neural networks (Perkins, Gupta, & Tammana, 1995)
Problem pertaining to the Nature of these Data Analysis Tools 1

- Relying on prescriptive (deterministic) models such as linear regression
- Linearity violated: likely leading to disqualifying the theory-informed hypotheses
- "The validity of the studies in which multiple regression is used to predict item difficulty is not high…It is hypothesized that using variables in combination and introducing forms of nonlinearity might improve the validity of item difficulty studies" (Perkins et al., 1995, p. 35).

(Graph from: http://amosdevelopment.com/video/mixture/regression/MixtureRegression.html)
Problem pertaining to the Nature of these Data Analysis Tools 2

- Multidimensional item response theory (IRT) models
- the fusion and rule space models,
- Limitation: they require a large sample size, a requirement which is not always fulfilled (Aryadoust, 2011).
Current Study

• A Class of Neuro-fuzzy models
• Adaptive Neuro-Fuzzy Inference Systems (ANFIS)
Neuro-fuzzy

• Synergy: Artificial Neural Networks (ANN) and Fuzzy Set Theory
• Landín, Rowe, and York (2009, p. 325): “Rule sets generated by neuro-fuzzy logic are completely in agreement with the findings based on statistical analysis and advantageously generate understandable and reusable knowledge,” and conclude that “Neurofuzzy logic is easy and rapid to apply and outcomes [yield] knowledge not revealed via statistical analysis.”
A Fuzzy Inference System: Fuzzification
Types of Neuro-Fuzzy Models

• Mamdani (e.g., Mamdani & Assilian, 1975)
• Takagi-Sugeno-Kang rule-based hybrid systems.
• This study uses a variant of the latter: Adaptive Neuro-Fuzzy Inference System (ANFIS).
• ANFIS optimizes parameter estimation and output prediction and reduces estimation time.
• Approximately 80% to 90% of the data: training
• Rest: testing
Method

• **Participants**: 209 Iranian, Chinese, and Malay,
• Between 16 and 45 years old
• **Materials**: the *Official IELTS Practice Materials* (University of Cambridge Local Examinations Syndicate, 2007).
• 40 test items
• 4 sections
Data Analysis

• Rasch item difficulty (WINSTEPS);
• Item coding (expert judgments)
• ANFIS (37/3) (the MATLAB, Version 7.11 (Mathworks, Inc., 2011); Seven ANFIS models generated,
• Incorporating between 1 and 7 hypothesized explanatory variables.
• Models comprising between 2 and 6 variables, numerous sub-models consisting of all possible variable combinations were tested.
Item coding

- Word count
- Prepositional phrases
- Modal verbs
- Propositional density of oral texts
- Propositional density of test items
- Item format
- Information type
- they were normalized before the modeling to take values between zero and one
Goodness-of-fit Indices

- Coefficient of efficiency ($E_f$): ranges between negative infinity and one
- Squared correlation coefficient ($r^2$): ranges between zero and one, with values near one indicating good fit.
- Root mean squared error (RMSE): Lower RMSE values indicate smaller error terms.
- Mean Absolute Error (MAE): Lower MEA values indicate smaller error terms.

Data Analysis
Training Data

**TABLE 2 BEST ANFIS MODELS GENERATED IN THE TRAINING STAGE AND EVALUATED IN THE VALIDATION STAGE**

<table>
<thead>
<tr>
<th>Models</th>
<th>$E_r$</th>
<th>$r^2$</th>
<th>RMSE</th>
<th>MAE</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-variable</td>
<td>0.947</td>
<td>0.999</td>
<td>0.306</td>
<td>0.239</td>
<td>3, 6, 7</td>
</tr>
<tr>
<td>4-variable</td>
<td>0.905</td>
<td>0.971</td>
<td>0.410</td>
<td>0.317</td>
<td>2, 3, 6, 7</td>
</tr>
<tr>
<td>5-variable</td>
<td>0.981</td>
<td>0.990</td>
<td>0.182</td>
<td>0.151</td>
<td>2, 3, 4, 6, 7</td>
</tr>
<tr>
<td>6-variable(^a)</td>
<td>0.737</td>
<td>1.00</td>
<td>0.687</td>
<td>0.508</td>
<td>1, 2, 3, 4, 5, 7</td>
</tr>
<tr>
<td>6-variable(^b)</td>
<td>0.788</td>
<td>0.910</td>
<td>0.616</td>
<td>0.595</td>
<td>1, 2, 3, 4, 6, 7</td>
</tr>
<tr>
<td>7-variable</td>
<td>0.966</td>
<td>0.980</td>
<td>0.246</td>
<td>0.180</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
</tr>
</tbody>
</table>

*Note.* \(^a\) and \(^b\) two 6-variable models fit equally well.

Variables: 1 = item format. 2 = information type. 3 = word count. 4 = prepositional phrases. 5 = modals. 6 = propositional density 7 = item density.
## Testing Data

### TABLE 3 VALIDATION STAGES OF THE 5- AND 7-VARIABLE OPTIMAL MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Item</th>
<th>Observed Value</th>
<th>Simulated Value</th>
<th>$E_f$</th>
<th>$r^2$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-variable</td>
<td>2</td>
<td>-1.88</td>
<td>-1.574</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.35</td>
<td>0.401</td>
<td>0.890</td>
<td>1.113</td>
<td>0.371</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>1.32</td>
<td>1.361</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-variable</td>
<td>2</td>
<td>-1.88</td>
<td>-1.110</td>
<td></td>
<td></td>
<td>0.705</td>
<td>0.578</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.35</td>
<td>-0.575</td>
<td>0.722</td>
<td>0.994</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>1.32</td>
<td>1.575</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion and Conclusion 1

• ANFIS found that all variables exerted an effect on item difficulty.
• MCQs were easier than open-ended items.
• MCQs appear to tax test takers’ cognitive processes less than constructed-response items,
• Propositional density of both spoken texts and test items affected item difficulty: agrees with Buck and Tatsuoka (1998).
• Propositional density of items affected item difficulty → a likely source of construct-irrelevant variance (CIV)
Discussion and Conclusion 2

• Information type and use of modal verbs were also predictors of item difficulty → (Buck & Tatsuoka, 1998).

• Word count affected item difficulty → Freedle and Kostin (1993, 1999),
Future applications of Neuro-fuzzy Models (NFMs)

• ANFIS model: more intuitive and theory-informed
• Hold great promise for behavioral prediction in language assessment.
• Computerized adaptive language testing (CALT)
• Coh-Metrix: useful technique to explore oral/reading texts
Study 2: Reading Comprehension

- Reading: Textbase
- Reading: Grammar + Vocabulary + Strategies ...
- Lexico-grammatical knowledge predict reading
- Vocabulary knowledge + Grammar knowledge → Reading
- Can the relationship be (mathematically) modelled?
Data Mining: Predictive Modelling

• Data mining: discovery of knowledge in data
• Captures meaningful patterns (i.e., information) in data → to constitute understandable structures
• Some predictive modelling techniques:
  • 1) Linear Regression
  • 2) Classification And Regression Tree (CART)
  • 3) Artificial Neural Networks
• Represent three schools of thought
• Conjecturing – predicting relationships and results
Artificial Neural Networks (ANNs)

- Nonlinear mathematical models inspired by human brains (Haykin, 1998).
- Consist of interconnected units or neurons.
- Capable of pattern recognition, prediction, classification, and learning.
- Learn by example: The network finds out how to solve the problem by itself.

Advantages of ANN

• No assumption on the relationships between dependent and independent variables
• Normality, linearity, & homoskedasticity
Technical Information of the ANN

- One Hidden Layer
- Six Inputs → Six Units
- Mathematical Activation Functions: Hyperbolic & Tangent
- Error Function: Sum of Squares

<table>
<thead>
<tr>
<th>Learning</th>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum of Squares Error</td>
</tr>
<tr>
<td></td>
<td>Percent Incorrect Predictions</td>
</tr>
<tr>
<td></td>
<td>Stopping Rule Used</td>
</tr>
<tr>
<td></td>
<td>Training Time</td>
</tr>
<tr>
<td>Testing</td>
<td>Sum of Squares Error</td>
</tr>
<tr>
<td></td>
<td>Percent Incorrect Predictions</td>
</tr>
</tbody>
</table>
Thank You!
Epilogue

- A science is exact only as far as it employs mathematics. (Immanuel Kant)
- Math: contribution to our understanding of the world
- Physics, biology, chemistry, social sciences, and applied linguistics
Summary

• Assumptions of Linear Regression are restrictive.

• ID variables might be highly correlated in LT data.

• CART and ANNs offer promise for re-/examining some of the data-driven LT theories.

• Due to their flexibility, ANNs can predict DVs with lower degrees of error of measurement.
Predicting Listening Test Item Difficulty 5

• Information density (Dunkel, Henning, & Chaudron, 1993);

• Buck and Tatsuoka (1998): dividing the number of content words by the number of words surrounding the information necessary to answer the test item (found influential).

• Rupp, Garcia, and Jamieson (2001, p. 211): the type/token ratio (i.e., a ratio of function or grammatical units to lexical units), sentence length, word count, and item-text interactions (found influential but not robust)
Applying the fuzzified input to the rules’ antecedents

- $\mu_1(X_1 = A_{[low]}) = 0.40$; and $\mu_2(X_1 = B_{[high]}) = 0.80$ (Lotfi Zadeh, 1965)

- Rule$_1$: IF $X_1 = A_{[low]}$, THEN $Y = 2$

- Rule$_2$: IF $X_1 = B_{[high]}$, THEN $Y = 5$

- ($Y$ is the output and could take any value depending on the range of the data to be predicted).
Fuzzy Inference System: Two Inputs

- Inputs: $X_1 = 14$ and $X_2 = 18$. /MF: low & High
- The joint functions of the two inputs might be rewritten as:

  - Rule 1: IF $X_1 = A_{\text{low}}$ AND $X_2 = A_{\text{low}}$, THEN $Y_1 = 2$ (1)
  - Rule 2: IF $X_1 = A_{\text{low}}$, AND $X_2 = B_{\text{high}}$, THEN $Y_2 = 3$ (2)
  - Rule 3: IF $X_1 = B_{\text{high}}$, AND $X_2 = A_{\text{low}}$, THEN $Y_3 = 3$ (3)
  - Rule 4: IF $X_1 = B_{\text{high}}$, AND $X_2 = B_{\text{high}}$, THEN $Y_4 = 5$ (4)
Example of Two Inputs

- Low and high values of \( X_1 = 0.8 \) and \( 0.4 \).
- Low and high values of \( X_2 = 0.3 \) and \( 0.6 \).

- 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:

\[
Y = (\mu_1 \times Y_1) + (\mu_2 \times Y_2) + (\mu_3 \times Y_3) + (\mu_4 \times Y_4) 
\]

(see Vilém, 1989, 2005).
5-MF model

Discussion and Conclusion 4

• Path modeling: enabled the consideration of test section as moderating.
• Test section: representing a change in the purpose of the oral text (e.g., academic or non-academic input) & the speed of delivery
• Its effect supports the validity argument: Flowerdew (1994): texts with a high speed of delivery more difficult.
• path and ANFIS models converge to some degree, and also differ in some respects.
• ANFIS model: more intuitive and theory-informed