Cognitive Complexity in the Remote Association Test - Chinese Version

Su-Pin Hung, Po-Sheng Huang & Hsueh-Chih Chen


To link to this article: https://doi.org/10.1080/10400419.2016.1229988

Published online: 14 Nov 2016.

Article views: 168

View related articles

View Crossmark data

Citing articles: 1 View citing articles
Cognitive Complexity in the Remote Association Test - Chinese Version

Su-Pin Hung  
National Cheng Kung University

Po-Sheng Huang  
Hsuan Chuang University

Hsueh-Chih Chen  
National Taiwan Normal University

The remote association test (RAT) has been applied in various fields; however, evidence of construct validity for the original version and subsequent extensions of the RAT remains limited. This study aimed to elucidate the dimensionality and the relationship between item features and item difficulties for the RAT—Chinese Version (RAT-C) using the Rasch model and the linear logistic test model (LLTM). The revised 30-item RAT-C was administered to 475 undergraduates (263 women and 212 men) in 8 universities in Taiwan. Item features (including types of associations among stimulus words, and frequency and concreteness of target words) were recoded. The analysis found that the RAT-C measured a single latent construct, with all 30 items conforming to the Rasch model’s expectation. Furthermore, according to the LLTM analysis, most item features predicted Rasch item difficulty, suggesting that these features can explain why some items were more difficult than others and can be used to create new items with known item difficulty to tailor the difficulty level for different groups of participants in the future.

Remote association has long been treated as an indicator of the ability to solve unusual problems because the underlying associative hierarchy can predict creative behavior (Mednick, 1962). Two college-level versions of the remote association test (RAT; Mednick, 1968) have been constructed, each consisting of 30 items. A typical RAT item consists of three stimulus words. None of the three words is directly related to one another but they are all, in some way, related to a fourth target word. The respondent’s task is to identify this target word. For example, for the three stimulus words blood, music, and cheese, a correct response would be blue, which can be associated with the stimulus words either through a compound word (blue blood) or through category association (blues music and blue cheese). Reaching a solution requires respondents to simultaneously retrieve a number of associates from long-term memory and match them with the stimulus words. The RAT is easy to administer, score, and interpret, and is widely used to study various phenomena, such as creative thinking and problem solving (Smith, 1995; Smith & Blankenship, 1991), psychopathology (Foder, 1999), performance feedback (McFarlin & Blascovich, 1984), use of attentional resources (Ansburg & Hill, 2003), and the relationship between working memory capacity and long term memory (Ricks, Turley-Ames, & Wiley, 2007).

DIFFERENT VERSIONS OF THE RAT AND THEIR VALIDITY STUDIES

To avoid item overexposure, several versions of the RAT have been constructed (Bowden & Jung-Beeman, 2003; Bowden & Jung-Beeman, 2003; Bowden & Jung-Beeman, 2003).
Item response theory (IRT) offers several advantages over classical test theory. IRT models postulate item responses as a function of person measures and item parameters. Among IRT models, the family of Rasch models (Rasch, 1960) has been applied to creativity tests to evaluate their dimensionality and other psychometric properties (Huang, Chen, & Chen, 2012; Koizumi & Serafim, 2009; Teo & Waugh, 2010; Wang, Ho, Cheng, & Cheng, 2014; Wechsler, Vendramini, & Oakland, 2012). A few such studies have used other IRT models to assess the dimensionality of the RAT (Chermahini et al., 2012; Lee et al., 2014). One goal of this study was to further and deepen this effort, by both investigating the dimensionality of the RAT-C and examining the relationship between item features and RAT item difficulties.

In the dichotomous Rasch model, the probability of success for a person with ability \( \theta \) on an item with difficulty \( \delta \) is given by:

\[
P_{si} = \frac{\exp(\theta_s - \delta_i)}{1 + \exp(\theta_s - \delta_i)}
\]

(1)

The major task of Rasch analysis is to calibrate the person measures \( \theta_s, s = 1, \ldots, S \) and item difficulties \( \delta_i, i = 1, \ldots, I \) from item response data. In itself, the model does not explain why an item has such a difficulty level, but in practice it is of great value in attempting to explain or predict an item’s difficulty.

The linear logistic trait model (LLTM; Fischer, 1973) was developed to meet this goal. Let \( q_{ik} (k = 1, \ldots, K; i = 1, \ldots, I) \) denote the \( k \)th item feature (e.g., item length, cognitive complexity, etc.) of item \( i \). The item difficulty is then regressed on the \( K \) item features as:

\[
\delta_i = \beta_0 + \sum_{k=1}^{K} \beta_k q_{ik}
\]

(2)

where \( \beta_0 \) is the intercept and \( \beta_k \) is the regression weight for \( k \). A combination of Equations 1 and 2 leads to:

\[
P_{si} = \frac{\exp[\theta_s - (\beta_0 + \sum_{k=1}^{K} \beta_k q_{ik})]}{1 + \exp[\theta_s - (\beta_0 + \sum_{k=1}^{K} \beta_k q_{ik})]}
\]

(3)

This model is conventionally referred to as the LLTM and it has been widely applied to describe how item difficulties are affected by item features (Emetretson, 1998; Gorin, 2005; Janssen, Schepers, & Peres, 2004). The LLTM offers a number of advantages for assessing and ultimately enhancing construct validity at the item level (Emetretson & Daniel, 2008). First, decomposing
items into particular item features makes the latent construct more explicit and provides evidence in the validation process. Second, the regression weights for the item features can be used to predict item difficulty (see Equation 2) and their values reflect the relative importance of these item features. Third, item features are linked to item difficulties (Equation 2) and item difficulties are linked to item responses (Equation 1), meaning that item features are linked to item responses (Equation 3). Finally, if item features can accurately predict item difficulties, test developers are in a better position to create new items of targeted difficulty levels without field testing, to improve measurement precision with fewer items, and to adaptively and automatically create new items. To date, the LLTM approach has been used to create a wide range of new items (e.g., Raven matrix tasks, Hornke, Habon, & GmbH, 1986; reading comprehension items, Sonnleitner, 2008; mathematical problem solving items, Embretson & Daniel, 2008).

In this study, a Rasch model was applied to the LLTM to the RAT-C. Table 1 shows eight item features that were used to generate RAT-C items. Among them, the first seven features involve the association between stimulus words and target words, and the eighth feature is related to the target words. The first item feature was loosely synonymous association (Dailey, 1978), in which the stimulus and target words were loosely synonymous (e.g., route and track). Definitionally dependent associations (Dailey, 1978) consisted of cases such as red maple and autumn. For category association, one word was a type of the other (Holley & Dansereau, 1984), such as fruit and watermelon.” For state associations, one word was a characteristic of the other (Holley & Dansereau, 1984), such as seven colors and rainbow. For functional associations, the presented words had a functional association with the target word, such as candle and illuminate. For compound words, the pairs formed a two-word phrase, such as on-screen and couples. For idiomatic associations, the response was related to the stimuli through an idiom, such as eyes and window. Finally, for frequency and concreteness level, words were grouped into levels by word frequency (high/low) and the level of concreteness (high/medium), based on free association norms for 1200 Chinese words (Chen, 1999).

Not all of these item features were used to generate any given RAT-C item. For example, one RAT-C item had three two-character stimulus words (反射 [reflection] /打扮 [dress up] /猪八戒 [pig]) with the solution being the target word “镜子 [mirror]. The target word was a high-frequency word with a medium level of concreteness (Chen, 1999). Three association types were used to connect the stimulus words with the target word: DDA (mirror, reflection), FA (a mirror is commonly used for getting dressed), and IA (猪八戒照镜子 which means something like ‘catch 22’ in Chinese).

### METHODS

#### Instrument

The RAT-C was initially developed by Huang et al. (2009), and has since been revised into two separate 30-item college level versions (Huang et al., 2012). In this study, version A was used. An earlier criterion-related validity study showed that the RAT-C had a positive correlation with performance on insight problems and had no significant relationship with performance on a divergent thinking test (Huang et al., 2012).

For data analysis purpose, items were recoded for the presence of each item feature by the first two authors (the original RAT item developers), working independently. For each item, each of the three stimulus words was recoded dichotomously based on the presence of each of the first seven items features discussed previously and in Table 1: 1 if the relationship existed, and 0, otherwise. The target words were also recoded dichotomously on the eighth item feature: 1 if the word was at the high level of both frequency and concreteness, and 0 otherwise (based on the frequency and concreteness levels in Chen, 1999). Chen (1999) found that words that were at the high level of both frequency and concreteness aroused more idiosyncratic and divergent associates than did words that were at the low level of frequency and/or were more abstract. Thus, it was expected that high frequency and concreteness stimulus words would generate more associates, making it easier to identify the correct target word and thus reducing item difficulty.

#### TABLE 1

Eight Item Features of the RAT-C

<table>
<thead>
<tr>
<th>Item Feature</th>
<th>Description</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus Words</td>
<td>Loose Synonymous Association</td>
<td>The stimulus word is loosely synonymous with the target word.</td>
</tr>
<tr>
<td></td>
<td>Definitionally Dependent Association</td>
<td>The stimulus word is definitionally dependent on the target word.</td>
</tr>
<tr>
<td></td>
<td>Category Association</td>
<td>The stimulus word is ‘a type of’ the target word.</td>
</tr>
<tr>
<td></td>
<td>State Association</td>
<td>The stimulus word is a characteristic of the target word.</td>
</tr>
<tr>
<td></td>
<td>Functional Association</td>
<td>The stimulus word has a functional association with the target word.</td>
</tr>
<tr>
<td></td>
<td>Compound Word</td>
<td>The stimulus and target words form a compound word or phrase.</td>
</tr>
<tr>
<td></td>
<td>Idiomatic</td>
<td>The stimulus word is related to the target word through an idiom or slang.</td>
</tr>
<tr>
<td>Target Words</td>
<td>High Word Frequency and High Concrete Level</td>
<td>The target word is a high frequency and highly concrete word.</td>
</tr>
</tbody>
</table>
The item features were defined first, and then the recoding was conducted by the two authors independently. The percentage of agreement between the two authors was 0.80 in the first round. The recoding procedure was repeated two more times until consensus was achieved. The result of this recoding was a complete feature matrix describing the eight item features of each item. For illustrative purpose, Table 2 shows the feature matrix of a single RAT-C item. This item involved loosely synonymous association, definitionally dependent association, and functional association, but did not involve category, state, compound or idiomatic association. The target word was at the high level both in word frequency and concreteness.

### Participants
A total of 475 undergraduates (263 women and 212 men) from eight universities in Taiwan were recruited to take the RAT-C. The students came from different academic disciplines and their ages ranged from 17 to 28 years old (Mean = 19.57, SD = 1.62). All participants received credits toward fulfillment of a course requirement. They were allowed up to 15 min to complete the test. All participants completed the test in the allotted time.

### Data Analysis
The Rasch model was fit to the data using the software ConQuest 3.0.1 (Adams, Wu, & Wilson, 2012) to calibrate Rasch model parameters, infit MNSQ (weighted fit mean-square, primarily reflecting response patterns), and outfit MNSQ (unweighted fit mean-square, reflecting the influence of outliers). The infit and outfit MNSQ have an expected value of 1 when the data fits the model, and can range from 0 to infinity. Wright and Linacre (1994) suggested that a reasonable range of MNSQ was 0.6–1.4, with MNSQ values greater than 2 likely to distort or degrade measurement. The person separation reliability, being a signal-to-noise ratio calculated using an estimate of true variance over error variance and analogous to the KR-20 or Cronbach alpha coefficient, was used as an internal consistency index. Item separation reliability indicated how well the items were separated by the participants taking the test. The LLTM model was also fit to the data with the default design matrix in ConQuest replaced by the feature matrix for the RAT-C items. The likelihood ratio test and the Akaike information criterion (AIC) was calculated for model comparison, with lower AIC values implying a better model fit.

### RESULTS

#### Item Characteristics
The mean RAT-C raw score was 18.56 (SD = 5.23). The Cronbach alpha coefficient was .81, indicating good internal consistency. The first two columns in Table 3 show, for each item, the p-values (the proportion of correct responses), which ranged from 0.20 to 0.90, and the item-rest correlations, which ranged from 0.26 to 0.52. Item difficulties varied widely, and all items were positively related to the overall test score.

Rasch item difficulties and item fit statistics are displayed in Table 3. Items infit and outfit MNSQ ranged from 0.82 to 1.15, indicating that all items met the Rasch model’s expectations. In addition, the person separation reliability was .85, indicating that the 30 RAT-C items were able to separate the 475 participants to a moderate degree. The item separation reliability was .99, indicating that the RAT-C items were well separated by the participants. Item difficulties ranged from −2.14 logits (the easiest item) to 2.26 logits (the hardest item), with a mean difficulty of −0.74. Figure 1 presents the item-person map, locating both item and person parameters on the same scale. Person measures ranged from −1.74 logits to 2.57 logits. The item difficulties appeared to have been well-targeted when compared to the range of person measures, although slightly easy. In the future, it would be desirable to create a few very difficult items to better target high ability individuals.

#### LLTM Analysis of Item Features
The Rasch model had a log-likelihood deviance of 15,488 and an AIC of 15,550; whereas the LLTM had a log-likelihood deviance of 16,505 and an AIC of 16,527. Thus, the Rasch model had a better fit. This is not surprising because the Rasch model was more complicated. According to the LLTM analysis, the regression equation for the item difficulties was as follows:

<table>
<thead>
<tr>
<th>Item</th>
<th>Target Word</th>
<th>Loosely Synonymous Association</th>
<th>Definitionally Dependent Association</th>
<th>Category Association</th>
<th>State Association</th>
<th>Functional Association</th>
<th>Compound Word</th>
<th>Idiomatic</th>
<th>Frequency and High Concrete Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>喜事, 喜客、通知</td>
<td>喜帖</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
TABLE 3
Rasch Item Parameter Estimates

<table>
<thead>
<tr>
<th>Item</th>
<th>p-values</th>
<th>Item-rest correlation</th>
<th>Difficulty</th>
<th>SE</th>
<th>Outfit MNSQ</th>
<th>INFIT MNSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.33</td>
<td>2.26</td>
<td>0.10</td>
<td>1.03</td>
<td>1.05</td>
</tr>
<tr>
<td>2</td>
<td>0.36</td>
<td>0.39</td>
<td>1.37</td>
<td>0.09</td>
<td>1.03</td>
<td>1.01</td>
</tr>
<tr>
<td>3</td>
<td>0.59</td>
<td>0.39</td>
<td>0.22</td>
<td>0.09</td>
<td>1.05</td>
<td>1.04</td>
</tr>
<tr>
<td>4</td>
<td>0.84</td>
<td>0.40</td>
<td>-1.25</td>
<td>0.10</td>
<td>0.83</td>
<td>0.94</td>
</tr>
<tr>
<td>5</td>
<td>0.74</td>
<td>0.36</td>
<td>-0.57</td>
<td>0.09</td>
<td>1.07</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.57</td>
<td>0.36</td>
<td>0.35</td>
<td>0.09</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>7</td>
<td>0.43</td>
<td>0.40</td>
<td>1.00</td>
<td>0.09</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>8</td>
<td>0.46</td>
<td>0.47</td>
<td>0.87</td>
<td>0.09</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>9</td>
<td>0.67</td>
<td>0.35</td>
<td>-0.14</td>
<td>0.09</td>
<td>1.07</td>
<td>1.02</td>
</tr>
<tr>
<td>10</td>
<td>0.78</td>
<td>0.38</td>
<td>-0.80</td>
<td>0.10</td>
<td>1.04</td>
<td>0.97</td>
</tr>
<tr>
<td>11</td>
<td>0.59</td>
<td>0.42</td>
<td>0.26</td>
<td>0.09</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>12</td>
<td>0.41</td>
<td>0.42</td>
<td>1.12</td>
<td>0.09</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>13</td>
<td>0.57</td>
<td>0.26</td>
<td>0.33</td>
<td>0.09</td>
<td>1.13</td>
<td>1.10</td>
</tr>
<tr>
<td>14</td>
<td>0.59</td>
<td>0.52</td>
<td>0.24</td>
<td>0.09</td>
<td>0.86</td>
<td>0.90</td>
</tr>
<tr>
<td>15</td>
<td>0.87</td>
<td>0.41</td>
<td>-1.56</td>
<td>0.11</td>
<td>0.80</td>
<td>0.93</td>
</tr>
<tr>
<td>16</td>
<td>0.92</td>
<td>0.36</td>
<td>-2.14</td>
<td>0.12</td>
<td>0.81</td>
<td>0.95</td>
</tr>
<tr>
<td>17</td>
<td>0.86</td>
<td>0.36</td>
<td>-1.39</td>
<td>0.11</td>
<td>1.04</td>
<td>1.00</td>
</tr>
<tr>
<td>18</td>
<td>0.21</td>
<td>0.40</td>
<td>2.21</td>
<td>0.10</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>19</td>
<td>0.38</td>
<td>0.47</td>
<td>1.26</td>
<td>0.09</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>20</td>
<td>0.66</td>
<td>0.40</td>
<td>-0.08</td>
<td>0.09</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>21</td>
<td>0.77</td>
<td>0.39</td>
<td>-0.74</td>
<td>0.10</td>
<td>1.03</td>
<td>1.00</td>
</tr>
<tr>
<td>22</td>
<td>0.72</td>
<td>0.48</td>
<td>-0.42</td>
<td>0.09</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>23</td>
<td>0.64</td>
<td>0.47</td>
<td>0.01</td>
<td>0.09</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>24</td>
<td>0.65</td>
<td>0.30</td>
<td>-0.05</td>
<td>0.09</td>
<td>1.12</td>
<td>1.07</td>
</tr>
<tr>
<td>25</td>
<td>0.56</td>
<td>0.42</td>
<td>0.39</td>
<td>0.09</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>26</td>
<td>0.71</td>
<td>0.28</td>
<td>-0.35</td>
<td>0.09</td>
<td>1.15</td>
<td>1.07</td>
</tr>
<tr>
<td>27</td>
<td>0.88</td>
<td>0.27</td>
<td>-1.60</td>
<td>0.11</td>
<td>1.08</td>
<td>1.01</td>
</tr>
<tr>
<td>28</td>
<td>0.62</td>
<td>0.51</td>
<td>0.12</td>
<td>0.09</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>29</td>
<td>0.43</td>
<td>0.47</td>
<td>0.99</td>
<td>0.09</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>30</td>
<td>0.91</td>
<td>0.37</td>
<td>-1.91*</td>
<td>0.50</td>
<td>0.82</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Note. *The items are not provided to avoid compromising future applications of the test, but can be requested from the authors.

The asterisk indicates that this parameter has been constrained. Estimates were scaled by fixing item parameters (Mean = 0), according to typical Rasch model procedures.

\[
\delta_i = 0.42 \times \text{LSA} - 1.22 \times \text{DDA} - 0.58 \times \text{CA} \\
+ 0.04 \times \text{SA} + 0.25 \times \text{FA} - 0.08 \times \text{CW} \\
+ 1.24 \times \text{ID} - 0.43 \times \text{HH} + 0.016 
\]

The predicted difficulty of any item can be estimated with Equation 4 by inserting its feature scores. For example, the predicted difficulty of the item in Table 2 can be computed as:

\[
\delta_i = 0.42 \times 1 - 1.22 \times 1 - 0.58 \times 0 + 0.04 \\
\times 0 + 0.25 \times 1 - 0.08 \times 0 + 1.24 \times 0 - 0.43 \\
\times 0 + 0.16 = -0.39
\]

The Impact of Item Features on Item Difficulty

As expected, all but one regression weight for the eight features were statistically significant, indicating that they were useful in predicting item difficulties. The regression weight for state association was very close to zero (0.04), indicating it was not helpful for the prediction, given there were already seven other features. The positive regression coefficients for loosely synonymous association (0.42), functional association (0.25) and idiomatic association (1.24) indicated that the presence of these three features made items more difficult. Among the eight features, idiomatic association had the largest impact on item difficulty, in line with earlier research showing that performance on the RAT was largely determined by the ability to respond correctly to idiomatic items (Dailey, 1978).

Both loosely synonymous association and definitional-dependent association are types of semantic association (Dailey, 1978). However, these two item features had contrasting impacts on item difficulty: whereas loosely synonymously association contributed to item difficulty, the relatively strong negative regression coefficient found for definitionally dependent association (-1.22) indicated that this kind of association actually decreased item difficulty (while holding the other features constant). This was consistent with findings on free association responses (Chen, 1999), in which respondents were able to generate far more definitional-dependently associations than loosely synonymous associations. Finally, the regression weights for category association (-0.58) and high frequency and concreteness (-0.43) were very similar in magnitude, with both types decreasing item difficulty. The comparatively weak negative regression coefficient found for compound words (-0.08) was in line with the expectation that such lexical “neighbor words” were retrieved more easily than other words in the association process.

The correlation between the Rasch item difficulties and the LLTM-predicted item difficulties was .80, meaning that about 64% of the variance in the Rasch item difficulties was attributable to the eight item features. The percentage of explained variance was higher than found in previous studies (Emberston & Daniel, 2008; Hornke et al., 1986). Figure 2 displays the scatterplot of these two kinds of item difficulties and their 95% confidence intervals. All 30 items fell within the 95% confidence intervals, suggesting a high level of agreement between these two kinds of item difficulties. The high agreement and high R² justified the use of these eight item features to generate RAT items, and thus supported its item-level construct validity.

DISCUSSION

This study illustrated how eight item features can be used to generate RAT-C items, explain their difficulties, and provide evidence for item-level construct validity. The results have
FIGURE 1  Person-Item map. Notes. Person measures are on the left-hand side of the logit scale and item difficulties are on the right-hand side. Each “X” represents 0.7 participants.

FIGURE 2  Relationship between the Rasch difficulties and the predicted LLTM difficulties.
two major implications. First, the relative importance of the eight item features on item difficulties was evaluated, and together they accounted for over 60% of the variance in item difficulties. Thus, the RAT-C items can be decomposed into different combinations of the eight item features, so a more detailed understanding of the latent construct can be obtained. Second, the results support the use of the eight item features to create new RAT items in the future. There are altogether 256 (2^8) possible combinations of these eight item features. Each combination creates a distinct type of RAT items. Using Equation 4, users can create items with known difficult levels (no field testing needed) to tailor tests to the ability levels of individual participants or to meet other testing purposes (e.g., very difficult items could be generated for scholarship competitions).

There are limitations to this study. All of the participants in the study were Taiwanese university students. It is possible that Chinese-speaking participants who live outside of Taiwan may be unfamiliar with slang or idioms used in Taiwan. More research is needed to better understand the utility of the RAT-C for cross-cultural applications. In particular, an investigation of cross-cultural differential item functioning (Holland & Wainer, 1993) for the RAT-C would be very useful in supporting the generalizability of our findings. Continued focus can be dedicated to refining the RAT-C through the use of larger cross-validation samples in which respondents represent diverse target populations. Further research might fruitfully apply LTM extension models (e.g., the multicomponent latent trait model; Embretson, 1984) to the RAT-C dataset to determine whether other abilities are also required for successful performance on the RAT-C.

REFERENCES


