Application of Rasch Model for Cloze Test Quality Analysis in Language Teaching Situation

Chang, Xinping[1] and Cao, Jiaying[2]

[1] Sun Yat-sen University, Guangzhou 510275, China
[2] Sun Yat-sen University, Guangzhou 510275, China
flscxp@mail.sysu.edu.cn

Abstract. The fundamental function of language testing is to facilitate language teaching and learning. Hence, teachers need to be informed of effective measurement methods. Quality control of routine language tests in daily teaching situations needs to be given more attention. This research employs the Rasch model together with a content validity analysis framework to look into the quality of cloze items in a Grade Seven English mid-term test paper with the purpose to introduce quantitative test quality control measures to the daily teaching situations to help discriminate item suitability to the learners. To be specific, the following five indicators of *unidimensionality, local independence, item fitness to the model, suitability of items to test-takers* and *measurement bias between groups* in the Rasch model are employed for the examination of the performance of a group of Grade Seven students in their mid-term test paper in a junior high school in Jiangsu Province. The results show that these cloze items in the test paper are of high quality in general. However, poor differentiation of language ability for the above-average level students and measurement bias between male and female students are detected among certain items. The examination process and the findings concerning the cloze items under investigation have provided powerful evidence that the Rasch model has its special advantages in discriminating items in terms of item difficulty and person ability on the one hand, it also demonstrates its edge on classical measures in terms of the amount of detailed and multi-faceted information it provides. Discussions around the findings and their implications for language testing use in daily language instruction are provided at the end of the paper.

Keywords: Rasch Model, Cloze Items, The Content Validity Analysis Framework, Ministeps, Language Instruction.

1 Introduction

Quality control for language tests in daily language teaching situations is generally neglected in the language testing field due to the fact that teachers are generally fully occupied by heavy teaching load. Hence, in general, seldom do teachers conduct item analysis on their tests. Consequently, teachers in practice have no idea about the quality of their tests for their students. The present study attempts to employ the Rasch model

Q. Zhang (ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)*,
Atlantis Highlights in Social Sciences, Education and Humanities 23,
https://doi.org/10.2991/978-94-6463-494-5_27

together with a supplementary content validity analysis framework to analyze items in a cloze test used at a certain stage of daily language teaching situation to see whether such kind of analysis can help to discriminate the effectiveness of the items in achieving the expected purpose of assessment so as to provide teachers and students with useful information on language teaching and learning. Rasch model was developed by the Danish mathematician Georg Rasch in the 1960s as a means to predict the relationship between an item and a latent trait for achievement testing among school children. It is one of the several models developed on the basis of item response theory, an approach for quantifying latent traits based on the assumption that a person's response to an item is a function of the difference between his/her abilities and the characteristics of the item. This model is widely used for the analysis and validation of educational and psychological tests and for scaling respondents (Dhyaaldian et al., 2022), because it can provide information on the suitability of test items to test-takers and can provide a more global view of what test-takers know (Demars, 2010, p. 79).

Cloze is an item type developed by Wilson Taylor in 1953 based on Gestalt psychology and the linguistic phenomenon of redundancy. Cloze items were originally used to measure the readability of language material and the reading comprehension ability of students, and were later widely used to estimate the comprehensive language ability of students. The ability to complete the cloze items is generally considered to be a comprehensive embodiment of language use (Li, 1997, p. 234).

Based on a Grade Seven English mid-term test paper and students' performance on it, the present study uses the Rasch model to analyze the quality of cloze items in the test paper by examining the following two aspects of the cloze test, i.e. the content validity and the fitness of the cloze items for the targeted group of students. The objectives are achieved by exploring the performance from the following five dimensions: *unidimensionality, local independence, item fitness to the model, items' suitability to test-takers*, and *measurement bias between groups* with the purpose to demonstrate the applicability of this model to the assessment of ordinary small-sample test quality control, and to introduce quantitative test quality control measures to discriminate item suitability in commonly-used language testing methods like cloze.

2 Literature Review

2.1 Rasch Model

As a construct validation tool, Rasch model assumes that the most economical and effective predictor of a latent trait is the relationship between the two variables: item difficulty and person ability. Item difficulty is concerned with the qualities of the item whereas person ability is concerned with qualities of the person. Georg Rasch believed that "a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one" (1960, p. 117). That is, Rasch model works from the principle that the key predominant underlying attribute of the measurement situation is the latent

trait expressed in the items, elicited from the candidates, and recorded in the performances when each test candidate and each test item interact (Bond & Fox, 2001, p. 107).

When the data fit the Rasch model, the model demonstrates a number of functions. First, according to Rasch, the model has a property called specific objectivity, which means that the comparison between two persons should not be influenced by the specific items used for the comparison, and the comparison between two items is person-free (cited from Wu et al., 2017, p. 111). Similarly, Smith (2003, p. 9) also mentioned that the parameter estimates in Rasch model are neither sample nor test dependent. Second, it can report estimates of the difficulty level of each item and the ability level of each person on the same interval scale in a common metric, making it possible for comparisons between items, between persons, and between items and persons (Wu et al., 2017, p. 124). Third, Rasch model can be used to combine the estimated measure of any person with that of any item to produce expected response values. These expected response values can be compared with the observed responses, and then combined in various useful ways to produce direct tests of the fit of an item, a person, etc. to the model (Smith, 2003, p. 9). For example, Rasch model can be used to identify such under-fit items as being easy items while more capable persons get them wrong, or to identify such under-fit persons as having high-level ability while performing unexpectedly badly on easy items. This function of the model, therefore, may provide teachers with information concerning the problem of test items or students' gaps in their knowledge or their failure to concentrate on the item, etc. Fourth, some scholars pointed out that Rasch model coupled with item banking testing programs can help to develop more adaptable tests, because different levels of students can take different tests which are suitable to each of them and the results can still be compared on the same scale (Waugh, 2011, p. 2). Given these advantages, the Rasch model is widely used in the field of language testing (Hughes & Hughes, 2020, p. 251).

In terms of the measurement facets involved, research using the Rasch model can be broadly divided into two-facet and multifaceted research. Two facet measurement model covers two aspects of the measurement situation for assessment with objective items: the item difficulty level and the person's ability level. Measurement models that involve three or more facets consider many factors, such as scoring criteria, rater bias, students, items, etc., and are mainly used for assessment of subjective items, such as speaking, writing, translation, reading and writing (Peng et al., 2022). The current study aims to explore the cloze items, so a two-facet Rasch model is adopted.

2.2 Cloze Items

Cloze items are believed to be able to measure learners' comprehensive language ability (Li, 1997, p. 234). In view of the fact that in China a lot of large-scale examinations such as the Senior High School Entrance Examination, the National College Entrance Examination, the Test for English Majors-Band 4, etc. have employed this item type frequently with different adapted forms, this method keeps its attraction among researchers (Chang and Zhu, 2017; Chang, 2020; Xu et al., 2022).

Rasch model in modern measurement theory provides a new way to analyze the quality of this kind of test item and has stimulated a number of studies on its applicability and actual application to the quality analysis of cloze items in different kinds of tests. Oliveira et al. (2012), for instance explored the application of this model to the analysis of cloze items; Chen and Zhou (2018) used this model to examine the cloze test in the National Matriculation English Test (NMET) in China; Gan et al. (2018) applied it to the investigation of the 2015 and 2016 Test for English Majors-Band 4 (TEM4) in China; Baghaei and Ravand (2019) applied it to the examination of National University Entrance Examination in Iran; Mokshein et al. (2019) used it for the quality analysis of the English Paper 1 items of a Primary School Achievement Test administered by the Ministry of Education to Grade 6 students in public and private schools in Malaysia, and Dhyaaldian et al. (2022) attempted it at a cloze test in a mid-term exam in a reading comprehension course in a university in Iraq. However, most of the studies are around cloze tests in well-established large-scale tests as illustrated above, whereas little research is found on the analysis of cloze items in daily teaching situations from the perspective of using technical language testing quality control methods to assess its effectiveness in assessment, to diagnose the potential inadequacies of such tests and their possible effect on language teaching process.

Rasch model, with the aim of supporting true measurement, provides a mathematical framework which can estimate the probability of a specific response according to person ability and item difficulty parameters, and place them both on an interval scale. Such kind of equal interval measures can allow teachers to confidently compare the progress of, and between, students that are located at different portions of a single trait. In addition, research shows that a smaller number of targeted items can provide more reliable measures than a larger number of items with this framework (Embretson & Hershberger, 1999), which means that this model can be used with small samples because the expected performance of a person on an item can be inferred from each person's ability measure and the difficulty of items which are expressed on the same scale. This facilitates decisions on teaching content adjustment in daily practice. Furthermore, many other indices can also be used to evaluate the validity and reliability of the test method with this framework. Thus, applying this model to item quality control in language teaching situations can not only help teachers to assess the relative difficulty level of each item and the suitability of items to test-takers, but also can help to identify problematic items and test-takers' unusual behavior so as to facilitate better test development in teaching situations and teaching instruction.

3 Research Methods

3.1 The Participants

In this study, 68 Grade Seven students of a private junior high school in Wuxi, Jiangsu Province, participated in a mid-term test in which a cloze test method was employed as one of the measures of the students' comprehensive English ability after half a semester's learning. Among the group, 28 of them are girls and 40 of them are boys.

3.2 The Instrument

The cloze test for the analysis here was one of the integrative items in an English mid-term test paper employed to measure the students' progress in the first half of the 2023 academic year in the private junior high school mentioned above, and this test paper was supposed to be used only for the students in this school. The cloze test was developed by the teachers themselves in this school, and it takes 10% of the total score of 100 for the whole test paper. It contains ten objective multiple-choice questions concerning lexical-grammatical use ability. In the test paper, these ten items are numbered from 36 to 45 with each blank taking one score if filled successfully and a zero if not. Students are required to choose one appropriate answer from the four possible options designed for each blank.

3.3 Frameworks for the Analysis

Two frameworks are used for the analysis: the Rasch model for the main analysis and the content validity analysis framework (Li, 1997) as a supplement. To gain an initial understanding of the items, the two aspects of the content validity analysis framework (characteristics of the input and test point validity) were employed as the references for the qualitative examination of the items. Then, five Rasch model indicators were used as the framework for the statistical analysis of the data by employing the software Ministeps 5.5.0 (developed by Mike Linacre and downloaded from https://www.winsteps.com) for the quantitative analysis of students' performance on the items. Specific objectives of the five indicators are listed as below:

1. *Unidimensionality* was analyzed by means of the Principal Component Analysis of the Residuals, with the aim of examining whether the items that make up the cloze are measuring the same underlying trait or not.
2. *Local independence* was analyzed by the "Largest residual correlations for items", with the aim of examining whether there is local dependence between pairs of items or persons.
3. *Item fitness to the model* was examined with the aim of testing whether the Rasch model fits the data.
4. *Suitability of items to test-takers* was investigated by means of Bubble chart, reliability coefficient and separation coefficient, the distribution of students' ability and item difficulty in person-item map.
5. *Measurement bias between groups* was described by the differential item functioning (DIF) value.

3.4 Research Procedure

In order to understand the quality of these cloze items, several steps are involved. First, a qualitative content validity analysis was conducted. Then, data collected from students' performance on the cloze test were analyzed from the five aspects of *unidimensionality, local independence, item fitness to the model, items' suitability to test-takers*,

and *measurement bias between groups* by using the Rasch Model technique. In addition, students' background information was also collected as supplementary evidence for the discussion.

4 Results of the Analysis

4.1 Analysis Based on the Modified Content Validity Analysis Framework

To examine whether the content of the test constitutes a representative sample of targeted domain from a qualitative perspective, a modified framework was established by drawing on the characteristics of the input part from the test task characteristics framework proposed by Bachman and Palmer (1996, p.49) and the test point validity framework including test point level and focus factor proposed by Li (1997, p.250), which are both widely used to examine the content validity of a test (e.g., Zhang & Zhao, 2011; Xu & Deng, 2023).

4.2 Characteristics of the Input

The input consists of the material that the test takers processed (Bachman & Palmer, 1996), and its characteristics can be approached from the aspects of theme, genre, length, readability, etc. (e.g., Gu et al., 2009; Zhang & Zhao, 2010). In this study, CohMetrix 3.0, which is freely available to researchers from http://cohmetrix.com/, was used to assist the description of the cloze test text. In addition, the cloze test text was processed with all blanks completed, and Chinese characters were removed from the text.

The characteristics of the cloze test text are as follows: The genre of the passage is a narrative and the theme is about the story of Steven Hawking, which meets the requirements of the "*The English Curriculum Standard for Compulsory Education (2022 Edition)*" for junior high schools. The number of words of the passage is 156. The length of the cloze test in this study is in the range (i.e. between 150 and 350) suggested by Alderson (2000, p.256). Readability could be represented by Flesch Reading Ease. The higher the value, the easier the text is. The Flesch Reading Ease of the passage is 76.281, which belongs to “fairly easy” style of readability (Flesh, 1948, p. 230). Zhang and Zhao (2010, p. 32) mentioned that the Flesh Reading Ease of achievement test should be within 50-70. Thus, the cloze test text here is a little bit easier than the reasonable range of achievement test suggested by Zhang and Zhao.

4.3 Test Point Level and Focus Factor Analysis

Many scholars believe that rational deletion is superior to random fixed-distance deletion (Bachman, 1985; Li, 1997). To control the quality of cloze test, Li (1997) pointed out that careful consideration should be given to the decision of item deletion. Hence, a follow-up test point validity framework (TPVF) was constructed to qualitatively examine the items from two aspects: the test point level and the focus factor (ibid). The

TPVF comprises four different levels, ranging from lower to higher: word level (W), phrase level (P), sentence level (S), and discourse level (D), which indicates the resources required for answering the item. The lower-level item requires less resources while higher level item requires more resources. Thus, a word-level item can be answered solely by considering individual words without referring to the surrounding context whereas answering a discourse-level item will require the consideration of the context at a broader discourse level. Thus, a higher test point level indicates a greater test point validity. The focus factor is categorized into three domains: grammar, collocation, and meaning. Lower-level items emphasize the examination of the grammar factor, whereas higher-level items emphasize the scrutiny of the meaning factor (Li, 1997, p. 250). Based on the definitions of the four levels and the three focus factors, two coders (the two researchers in this study) coded the test point and the focus factor for each item individually. Holsti's method (Holsti, 1969) was used in this study to calculate the inter-coder reliability. The inter-coder reliability is 0.95 (>0.8). For the one item with inconsistent coding judgement, the code was determined after discussion and agreement between the two coders. Based on the two factors, a preliminary qualitative analysis of the ten items in this cloze test produces the following result as listed in Table 1.

Table 1. Test points distribution in the cloze test.

Item Number	36	37	38	39	40	41	42	43	44	45
Test Point Level	D	D	S	S	D	D	D	D	S	D
The Focus Factor	M	M	M	G	G	M	G	M	M	M

Note: D = discourse; S = sentence; P = phrase; M = meaning; C = collocation; G = grammar

It can be seen from Table 1 that no test points in this cloze test were set at the word level or the phrase level, whereas the discourse level items account for 70%, followed by the sentence level items for 30%. Thus, it can be concluded that the test points' levels are concentrated on the higher levels of sentence and discourse whereas the focus factors are mainly around the meaning aspect. This tendency out of the qualitative analysis is in general in line with the orientation advocated in *The English Curriculum Standard for Compulsory Education (2022 Edition)*, which indicates that this test demonstrates good validity in a general sense in terms of its designing, and provides us a general picture of the test points distribution. However, whether such designing can be well reflected in students' performance need more in-depth examination. Thus, we attempted a Rasch analysis to examine the quality of the items in detail in the following part.

5 Rasch Analysis

5.1 Unidimensionality Analysis

To conduct Rasch analysis, two basic requirements need to be met, among which, unidimensionality is one of them. Based on the Rasch model, useful measurement involves

the examination of only one human attribute at a time (Bond & Fox, 2001, p. 32). Thus, the collected data must be approximate to unidimensionality, which means that most of the test items must provoke data along the same underlying construct of the student (Smith, 2003, p. 9).

We used Ministeps 5.5.0 software to help us with the analysis by employing the method of the principal component analysis of the residuals, and eigenvalue is the main indicator we used. The result is displayed in Table 2.

Table 2 shows that the eigenvalue extracted from the first factor is 1.69, and those for the 2nd, 3rd, 4th, and 5th factors are 1.47, 1.30, 1.14, and 1.08 respectively, all of which do not exceed 3. It was proposed that when the minimum eigenvalue in the residual model is less than 3, the unidimensionality of the items is true (cited from Yuan, 2016, p.142). In addition, "In the unexplained variance, a 'secondary dimension' must have the strength of at least 3 items, so if the first contrast has 'units' (i.e., eigenvalue) less than 3 (for a reasonable length test) then the test is probably unidimensional" (Linacre, 2023, p. 652). Therefore, the cloze items of this English test paper meet the Rasch unidimensionality, and the Rasch model can be used to further analyze the test items and the students.

Table 2. Standardized residual variance in eigenvalue units.

	Eigenvalue	Observed	Expected
Total raw variance in observations	17.1497	100.0%	100.0%
Raw variance explained by measures	7.1497	41.7%	39.7%
Raw variance explained by persons	3.2137	18.7%	17.8%
Raw variance explained by items	3.9360	23.0%	21.8%
Raw unexplained variance (total)	10.0000	58.3%	60.3%
Unexplained variance in 1st contrast	1.6874	9.8%	16.9%
Unexplained variance in 2nd contrast	1.4653	8.5%	14.7%
Unexplained variance in 3rd contrast	1.3026	7.6%	13.0%
Unexplained variance in 4th contrast	1.1432	6.7%	11.4%
Unexplained variance in 5th contrast	1.0769	6.3%	10.8%

5.2 Local Independence

The second requirement for the use of Rasch model is that the data must demonstrate local independence, i.e., the probability of responding correctly to one item must not be influenced by the particular response to another item (Smith, 2003). There are currently no well-documented suggestions of the critical values which should be used to indicate local independence. Linacre (2023), held that high positive residual correlations in the table of "Largest residual correlations for items" (p. 449) may indicate local dependency between pairs of items or persons. Calkin et al. (2023) held that residual correlations exceeding 0.20 can reflect local dependency. Thus, in this study, the critical value used by Calkin et al. (2023) is adopted.

From Table 3, we can see that none of the residual correlations between items in this study exceed 0.20, thus the ten items here are not locally dependent.

Table 3. Largest residual correlations for items.

Correlation	Item	Item
.16	item 38	item 41
-.33	item 36	item 43
-.32	item 39	item 41
-.26	item 37	item 40
-.26	item 39	item 45
-.25	item 41	item 43
-.23	item 42	item 45
-.22	item 40	item 41
-.21	item 40	item 42
-.21	item 40	item 44
-.20	item 37	item 45
-.19	item 38	item 43
-.19	item 37	item 44
-.19	item 40	item 45
-.18	item 38	item 39
-.18	item 36	item 44
-.16	item 39	item 43
-.15	item 39	item 44
-.15	item 38	item 45
-.14	item 39	item 42

5.3 Item Fitness to the Model

Rasch model assumes that only if the data match the predicted value of the model, the results obtained by the analysis can be of practical value, which indicates that the functions of the Rasch model mentioned above can only hold if the item response data fit the model. Therefore, to claim the benefit of using the Rasch model, the data must fit the model to begin with (Wu et al., 2017, p. 122). Fit statistics are often used to make judgement about whether data fit the Rasch model (p. 140). In the Rasch model, two indicators, INFIT (i.e. an information-weighted fit statistic) and OUTFIT (i.e. an outlier-sensitive fit statistic), are typical fit statistics used to reflect how well the data fit the model. Linacre (2023) proposed a relatively broad range of fitness, as shown in Table 4. Therefore, this study uses a range of 0.5-1.5.

Table 4. Interpretation of parameter-level mean-square fit statistics. (Linacre, 2023, p.686)

Fit statistic	Interpretation
>2.0	Distorts or degrades the measurement system.
1.5-2.0	Unproductive for construction of measurement, but not degrading.
0.5-1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliability and separations.

Table 5 displays the Rasch model fit statistics for the cloze items in this test paper. It can be seen that the INFIT MNSQs are all within the acceptable range of 0.5-1.5, and so do the OUTFIT MNSQs except for item 39, item 40, and item 41, indicating that on the whole, the way students perform on these items is consistent with the model prediction, and the students' performance and ability are quantified. Among the three items' OUTFIT MNSQs, those of item 39 and item 40 are larger than the acceptable scope. Thus, these two items are poor fitting items with the Rasch model, indicating that there is some interference in measuring the English ability of students in these items. The OUTFIT MNSQ of item 41 is smaller than the acceptable scope, indicating that it is an "over-fit" item (MNSQ substantially less than 1). Wu et al. (2017, p. 154) suggested that fit statistics should be used as an indication for detecting problematic items rather than for setting concrete rules for accepting or rejecting items. For underfit items, such data should be used for assisting test designers in re-examining the items to look for sources of misfit, and then choosing to improve or reject items if sources of misfit can be identified. For over-fit items, the acceptance or rejection should be decided after examining the specific items. Blind use of such statistics to reject items should be avoided. Following this suggestion, item 39, item 40 and item 41 will be further analyzed below.

Table 5. Fit statistics of cloze items.

INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Item
1.27	1.79	5.24	6.23	item 39
1.19	1.66	1.61	2.37	item 40
.74	-.94	1.29	.68	item 36
1.25	.83	.84	-.05	item 43
1.00	.08	1.02	.17	item 42
.88	-.55	.82	-.41	item 38
.85	-.68	.75	-.57	item 37
.81	-1.53	.75	-.79	item 45
.72	-.79	.37	-.91	item 41
.72	-1.46	.52	-1.49	item 44

5.4 Person Most-Unexpected Observations

The analysis above helps us find two under-fit items: item 39 and item 40. Further analysis needs to be conducted to examine what causes their unexpected behavior. According to Hughes and Hughes (2020), one possible cause is from certain candidates' performance, and we can test that by seeing whether the item can fit the model well when we drop the data of certain candidates who perform unusually on it. Another possibility is from the item itself. Thus, unexpected responses on item 39 and item 40 were analyzed below.

```
MOST MISFITTING RESPONSE STRINGS
         |Person
         |3211 63221 542211165333255345436
OUTMNSQ  |5183536752838439641193297673408O  Item
      high--------------------------------
   5.24 A|...................1.....11.11.1  4 item 39
   1.61 B|0.0.0.....................11.1.  5 item 40
   1.29 C|.0..................00.........  1 item 36
    .84 D|................0.00...0..0.....  8 item 43
   1.02 E|...0.0...00.....................  7 item 42
    .82 e|......0.0....00..................  3 item 38
    .75 d|.......0...00...0................  2 item 37
    .75 c|.......................1....... 10 item 45
    .37 b|...................0....0......  6 item 41
    .52 a|..............0.0..............  9 item 44
         |---------------------------------low
         |3211563221854221116533325534543 6
         |5183 36752 384396411932976734080
```

Fig. 1. Most misfitting response strings.

Figure 1 shows a listing of the most unexpected responses for the ten cloze items with the rows demonstrating the unexpected responses of the most misfitting persons to the listed items. Following Smith (2003), the dot "." indicates that the responses are in the expected range and responses producing large residuals are shown with the actual response. As shown in Figure 1, the two items of primary interest are item 39 and item 40.

For item 39, the most misfitting responses are from persons labeled with numbers 51, 56, 37, 54, 40, 60. When the data of these persons are dropped, the INFIT MNSQ and OUTFIT MNSQ of item 39, as shown in Table 6, are 0.94 and 0.74 respectively. Both are within the acceptable range, which indicates that item 39 fits well to the model. Thus, we may conclude that there is no problem with item 39, but the six students mentioned above may have knowledge concerned with this test point though their ability was not expected by the model to answer this item correctly.

Table 6. Fit statistics for item 39 when six persons are deleted.

INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Item
.94	-.30	.74	-.48	item 39

Further analysis of item 39 shows that the ability it measures belongs to the sentence level ability. The context of item 39 is: “and the black holes suck up everything around __ - even light.” The correct choice is C. (them). The six students whose levels of ability were relatively poor but answered item 39 correctly. This might indicate that these students have a good command of knowledge of personal pronoun at the sentence level.

Similarly, in Figure 1, it can be seen that the most misfitting responses are from persons labeled with numbers 35, 18, 5, 38, 54, 43 for item 40. When the data of these persons are dropped, the INFIT MNSQ and OUTFIT MNSQ of item 40, as shown in Table7, are 0.87 and 0.76 respectively. Both are within the acceptable range, which indicates that item 40 fits well to the model. Thus, we may conclude that there is no problem with item 40 as well. For persons labeled with numbers 5, 18, 35, they were expected by the model to answer item 40 correctly, but they failed. On the contrary, for persons labeled with numbers 38, 54, 43, they unexpectedly answered item 40 correctly.

Table 7. Fit statistics for item 40 when six persons are deleted.

INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Item
.87	-.1.19	.76	-.1.03	item 40

Further analysis of item 40 shows that the ability it measures belongs to the discourse level ability. The sentence that item 40 lies in is “Hawking's studies changed ____ people look at the universe.” The correct choice is C (How). Students who were expected by the model to answer this item correctly but failed might indicate that these students had gaps of knowledge in this level or they might not give sufficient attention to this item.

The findings above show that these cloze items have good fit to the Rasch model and thus it is possible to scale students and cloze items on an interval unidimensional scale.

5.5 Measurement Error of Test-takers’ Abilities for Each Item

Bubble chart can visually present the fit statistics of each item and the standard error for each item. In the chart, the horizontal axis represents the mean square fitting value of items, and the vertical axis represents the difficulty measurement of items. Items on the top are more difficult ones and those towards the lower end of the scale are easier ones. One bubble represents one item. The coordinate of the bubble is the set of the mean square fitting value and the difficulty measurement value of the item. Also, the relative bubble sizes are set by the standard errors of the measures. The smaller the bubble, the smaller the standard error, and hence, the more accurate the measurement result is. Thus, the Bubble chart can help to identify the items with large measurement errors for test-takers’ abilities and can reflect how the measurement error is correlated with the difficulty level for each item, which can help to determine whether the items are suitable for these test-takers or not.

Although the fit statistics of the data and the Rasch model generally refer to INFIT MNSQ and OUTFIT MNSQ, the INFIT MNSQ is usually used as the main reference since the OUTFIT MNSQ is sensitive to extreme values (Chen & Zhou, 2018). Therefore, in this study, the horizontal axis of the bubble chart is set as the INFIT MNSQ, and the vertical axis is the difficulty measurement value of items. Figure 2 shows that the fit statistics of all the ten items are in the acceptable range of 0.5-1.5, which means good fit to the model. In addition, it can be seen that the difficulty measurement value of these cloze items is distributed between -3 logit and 3 logit, and the difficulty of most items is concentrated below 0 logit, which indicates that these cloze items are mainly lower-intermediate difficulty level items, which can provide accurate measurements for lower-intermediate level students. Besides, in Figure 2, the item with the largest bubble is item 41, indicating that the measurement error of estimating students' abilities of item 41 is the largest. The vertical axis value shows that item 41 is located in the lowest part of the vertical axis, which means that it is the easiest item. Thus, the reason for the largest measurement error of item 41 is that it is such an easy item for these test-takers that test-takers of various levels of abilities can answer correctly. In addition, the reason for the large measurement error of item 41 can also be seen from the Wright Map (see Figure 3), and the number of students at the level of item 41 is small. It is suggested that a small number of easy, non-discriminating items may be kept at the beginning of a test to give candidates confidence and reduce the stress they feel (Hughes and Hughes, 2020), however, in this test, there are too many lower-intermediate difficulty level cloze items while most students' ability level is generally higher than lower-intermediate level, thus item 41 might be deleted or modified.

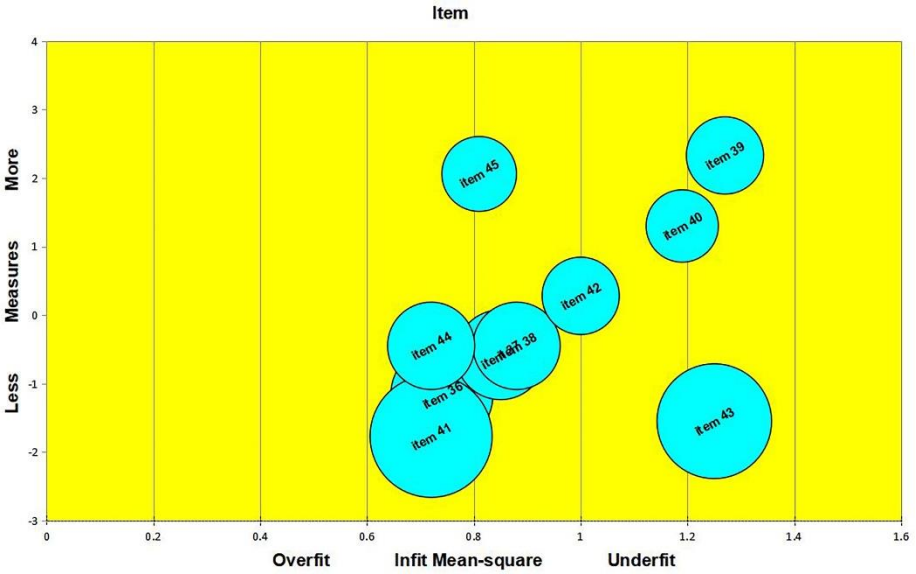

Fig. 2. Bubble chart representing measurement error for each item.

5.6 Reliability Coefficient and Separation Coefficient Analysis

Item reliability and separation statistics are used to verify the item hierarchy. When item separation value is low (< 3, item reliability < 0.9), it implies that a larger person sample size is needed in order to confirm the item difficulty hierarchy of the instrument. Person separation value is used to classify people. When this value is low (< 2, person reliability < 0.8) together with a relevant person sample, the instrument may be considered to be not sensitive enough to distinguish between high and low performers, which implies that more items may be needed. To be specific, the relationship between person reliability and the discrimination levels on persons is as follows: 0.9 = 3 or 4 levels. 0.8 = 2 or 3 levels. 0.5 = 1 or 2 levels (Linacre, 2023, p. 764).

Table 8. Item reliability coefficient and separation coefficient.

Person	10INPUT TOTAL	COUNT	10MEASURED MEASURE	REALSE	INFIT IMNSQ	INFIT ZSTD	OUTFIT OMNSQ	OUTFIT ZSTD
MEAN	47.6	68.0	.00	.37	.94	-.2	1.32	.5
P.SD	13.7	.0	1.38	.07	.21	1.1	1.35	2.2
REAL RMSE .38 TRUE SD 1.33 SEPARATION 3.50 Item RELIABILITY .92								

Table 9. Person reliability coefficient and separation coefficient.

Person	68INPUT TOTAL	COUNT	68MEASURED MEASURE	REALSE	INFIT IMNSQ	INFIT ZSTD	OUTFIT OMNSQ	OUTFIT ZSTD
MEAN	7.0	10.0	1.35	1.02	.96	.0	1.21	.1
P.SD	2.0	.0	1.48	.27	.39	.9	1.68	1.1
REAL RMSE 1.05 TRUE SD 1.05 SEPARATION 1.00 Person RELIABILITY .50								

The reliability and the separation coefficients of these cloze items are 0.92 and 3.50 respectively, as can be seen from Table 8, which indicates that the item difficulty range of these items is not narrow. Table 9 shows that the reliability and separation coefficients of the students who participated in this test are 0.50 and 1.00 respectively, which belongs to low person separation, indicating that these cloze items have low discriminating power in distinguishing students with different ability levels. According to the *Circular on Strengthening the Management of School Examinations during Compulsory Education* issued by Ministry of Education of the People's Republic of China in 2021, the mid-term test should discriminate students' ability into four or five levels. Therefore, the cloze items of this mid-term test paper in this study did not meet such requirements. From the results of the bubble chart and the Wright map, it is obvious that more items of above-average difficulty level need to be included whereas the proportion of items with below-average difficulty level needs to be reduced so that students whose ability is above the average level can be distinguished.

5.7 Distribution of Students' Ability and Item Difficulty in Person-Item Map

The 'Person-Item' map (often called Wright map) is a powerful graphical tool used to present the results of multiple-choice tests by placing the difficulty of the test items on the same measurement scale as the test takers' ability. This helps the user to conduct a comparison between the test takers and items and better understand whether the test measured appropriately or not. Here, in Figure 3, we can see that person ability and item difficulty are calibrated on the same scale, with the persons being located on the left-side of the scale according to ability measures and the items on the right-side according to item difficulty measures. The line in the center of the map is the logit scale, an interval-level measurement scale on which the distances at any point on that vertical scale are of equal size. This logit scale is the measurement unit common to both person ability and item difficulty. Along the logit scale, M stands for Mean, indicating average level; S stands for One Standard Error, meaning one standard deviation from the mean; T stands for Two Standard Errors, meaning two standard deviations from the mean. In this study, the scale from top to bottom corresponds to the gradual decrease of both student's ability level and the difficulty level of the corresponding item. According to the definition of item difficulty, when a person is located next to an item on the item-person map, the person has 50% chance of answering the item correctly (Wu et al., 2017, p. 119). That is, 50% threshold is used routinely in Rasch analysis to indicate a person's mastery (Bond & Fox, 2001, p. 45). Therefore, a person will have a greater than 50% chance of successful performance on items that are located lower than his/her location, and a person will have a less than 50% probability of successful performance on items that are located higher than his/her location.

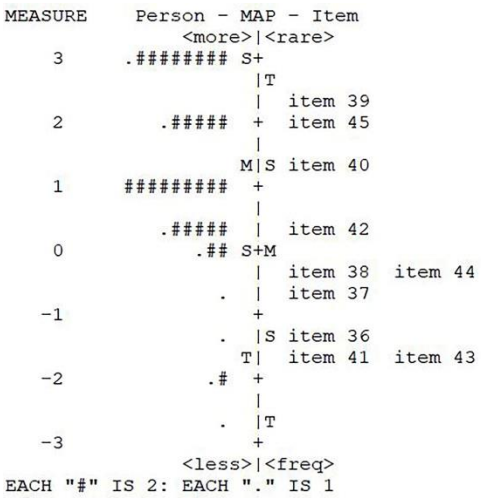

Fig. 3. Wright map of cloze items.

Here in Figure 3, the hash sign "#" represents 2 students, the dot "·" represents 1 student. It can be seen from the figure that a large gap exists between item 40 and item 42,

and also a large gap above item 39, indicating a lack of items that match the abilities of students whose level is above-average. This makes it impossible for students whose ability belong to above-average levels to be accurately measured. Obviously, item 39 is the most difficult item, while item 41 and item 43 are the easiest ones among the ten cloze items. The distribution of persons is larger than the distribution of items thresholds, which indicates that the students' ability range is wider than the items' difficulty range. To put it another way, the person distribution is top heavy in comparison with the item distribution. Thus, from a general perspective, we can conclude that the cloze items are not well-targeted to the test-takers and their inadequacies have been diagnosed. The cloze test needs more items with higher difficulties so that the abilities of high-level students can be estimated more precisely.

5.8 DIF Analysis

DIF (differential item functioning) refers to the difference in item functioning, which mainly reflects the difference in the performance of different groups of students on the same item. The two values in the DIF can be used to determine whether there is a DIF in an item: the DIF contrast value, and the p-value. The DIF CONTRAST is the difference between the two groups in terms of item difficulty. This value should be at least 0.5 logits for DIF to be noticeable. "Prob." shows the probability of observing this amount of contrast by chance. For statistically significant DIF on an item, Prob.$\leqslant$0.05 (Linacre, 2023, p.471). So, if the absolute value of the DIF contrast value of an item is greater than 0.5, and the p-value is less than 0.05, it can be concluded that the item has DIF. That is, the difference in students' performance on the same item is not caused by the difference in the student's ability but by the student's certain identity, such as gender, age, etc.

Table 10. DIF between students' genders.

Person CLASS	DIF MEASURE	Person CLASS	DIF MEASURE	DIF CONTRAST	Rasch-Welch Prob.	Item
M	-1.40	W	-.86	-.55	.5099	item 36
M	-.42	W	-.86	.43	.5733	item 37
M	-.86	W	.03	-.89	.2044	item 38
M	1.63	W	3.49	-1.86	.0104	item 39
M	1.49	W	1.07	.42	.4705	item 40
M	-1.73	W	-1.87	.14	.8938	item 41
M	.61	W	-.23	.83	.2045	item 42
M	-1.73	W	-1.28	-.45	.6261	item 43
M	-.23	W	-.85	.63	.4088	item 44
M	2.42	W	1.63	.79	.1952	item 45

As can be seen from Table 10, for item 39, the absolute value of the DIF contrast value is 1.86, which is greater than 0.5, and the p-value of item 39 is 0.01, which is less

than 0.05. Thus item 39 has a DIF between students' genders. That is, for boys, item 39 is more difficult than it is for girls. Thus, there is a certain degree of injustice in item 39.

6 Discussion

This study aims to explore and illustrate how quality control analysis around ordinary language tests in daily teaching situation can be conducted by employing the Rasch model together with a supplementary content validity analysis framework and how such quality control analysis can provide insights to the improvement of daily test development and language teaching.

The analysis was purported to examine the quality of the items in a cloze test of Grade Seven mid-term English test paper. The examination process and the findings concerning the cloze items under investigation have provided powerful evidence that the Rasch model has its special advantages in discriminating items in terms of item difficulty and person ability on the one hand, it also demonstrates its edge on classical measures in terms of the amount of detailed and multi-faceted information it provides. Specifically, the Rasch model has demonstrated its following advantages in evaluating test items:

1. By means of the Principal Component Analysis of the Residuals, we can gain a clear picture of whether the items are measuring the same latent trait or not by examining the indicator of unidimensionality. In this analysis, the ten cloze items under investigation meet the unidimensional hypothesis, and hence, we can conclude that they meet the purpose of measuring the assumed ability for this level of students;
2. By examining the "Largest residual correlations for items", local dependency between pairs of items or persons can be detected. In this analysis, there is no local dependence in the ten cloze items. Thus, we can conclude that the ten cloze items satisfy the second requirement of the Rasch model.
3. The model fit statistics can provide us information about the objectiveness of cloze items. In this study, the cloze items under examination fit the Rasch model well, confirming the objectiveness of these items measurement in their evaluation of the assumed students' abilities at this stage.
4. The bubble chart, the reliability coefficient, the separation coefficient and the Wright map provide us with evidence from different angles to help us with the decision about the difficulty level of the ten cloze items. Thus, in this study, we can conclude with more confidence and in a more elaborate way that these items can cover the ability range of most of the intermediate and low-level ability students. Hence, one the one hand, we can conclude that these items can provide accurate measurement for the intermediate and low-level ability students, but on the other hand, however, it was found that such items do not fit students whose abilities are at the above-average levels. To meet the requirement of *Circular on Strengthening the Management of School Examinations during Compulsory Education*, which requests that mid-term test should discriminate students into four or five levels, such cloze items need to be improved so as to differentiate among the above-average level students.

Therefore, some cloze items with lower-intermediate level difficulty can be deleted, and items with high level difficulty can be added;

5. The DIF provides us with information concerning item functioning for different groups of students. In this study, a gender bias on item 39 against male students was detected from the analysis of the students' performance on this item. Further analysis of item 39 shows that this item measures the language ability at the sentence level, particularly focusing on the examination of grammar. This reminds instructors that in language teaching practice, peer learning with mixed-gender activities can be encouraged so as to reduce gender differences in grammar learning.

It is obvious that findings with such amplitude coming from Rasch-model-based analysis are very inspiring for language instructors in classrooms due to its demonstration of details and the rich information it provides.

7 Conclusion

The discussion above has indicated that the information obtained from this Rasch-model-based analysis is quite rich and powerful in its ability of illustration, although the analysis is only based on a small test. Following the general purpose of this study which aims to introduce quantitative test quality control measures to daily teaching situations to help discriminate item suitability to the learners, two specific objectives were also set at the beginning which are the examination of the content validity and the fitness of the cloze items for the targeted group of students. With the demonstration of the analysis above, the following conclusions are thus drawn as stated below.

Firstly, both the qualitative analysis and the quantitative examination concerning the validity of the ten items show that the overall quality of these cloze items is generally high, but the differentiation of these items still needs to be improved, and some items also need to be further polished. Secondly, the detailed quantitative examination process helps to detect in detail why some items fail to match the abilities of the targeted students and why some students fail to perform certain items as expected, providing more detailed information for language teachers in practice and making it possible for more confident improvement both in testing and teaching. The whole process and the results have illustrated powerfully the applicability and efficacy of the Rasch model together with the supplementary content validity analysis framework in helping us to conduct quality control of commonly-used language tests in daily language teaching. It is obvious that the Rasch model, in particular, demonstrates a very strong discriminating power in helping to pinpoint problematic items and provide multi-faceted information around the items under investigation. Such rich information is very useful for both test designers and test users in item bank construction and language teaching, especially when such kind of information is introduced to language teachers in practice. Hence, there is no doubt that it will be a good resource for teachers to refer to when they want to diagnose their students and their own teaching.

In addition, the process also indicates the necessity of introducing packaged quality control measures into daily-used test items in routine language teaching. In this study, for instance, the focus factors of many comparatively higher-level difficulty

cloze items, such as item 39, item 40, and item 42, mainly examine grammatical knowledge, indicating that grammar is still a difficult area for students of this stage, and more efforts still need to be given to this aspect in daily classroom instruction. Also, the findings in DIF analysis reveal that there are significant differences in the abilities of students with different genders in some items. Such findings could remind teachers of designing stratified or adapted assignments for students of different levels or genders. It can be seen that such subtle characteristics of the items and the students can be easily neglected without a close examination of the item quality in daily test use, and negative effects in teaching and learning may be induced due to neglect in the long run. Thus, from the perspective of language teaching and learning, which is also the fundamental function of language testing, it is suggested that a Rasch-model-based package for daily testing analysis with strong operability characterized by user-friendly instructions and report sheets could be further developed so as to make it easily accessible to ordinary teachers and to bridge the gap between theory and practice. It is hoped that the advancement of new technology will change the complicated test quality control measures into handy tools for daily language teaching use. From what we can see about the applicability of the Rasch analysis at present, the realization of the idea does not seem to be very far.

References

1. Dhyaaldian, S. M. A., Al-Zubaidi, S. H., Mutlak, D. A., Neamah, N. R., Albeer, A. A. M. A., Hamad, D. A., Al Hasani, S. F., Jaber, M. M. and Maabreh, H. G.: Psychometric Evaluation of Cloze Tests with the Rasch Model. International Journal of Language Testing 12(2), 95-106 (2022)
2. DeMars, C.: Item Response Theory (Understanding Statistics). Oxford University Press, New York (2010)
3. Li, X. J.: [The Science and Art of Language Testing]. Hunan Education Publishing House, Changsha (1997)
4. Rasch G.: Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen (1960)
5. Bond, T. and Fox, C. M.: Applying the Rasch Model: Fundamental Measurement in the Human Sciences, LEA (2001)
6. Wu, M., Tam, H. P. and Jen, T. H.: Educational Measurement for Applied Researchers: Theory into Practice. Springer Singapore (2017)
7. Smith, R. M.: Rasch Measurement Models-Interpreting WINSTEPS and FACETS Output. JAM Press, Minnesota (2003)
8. Waugh, R.: Applications of Rasch Measurement in Education (Education in a Competitive and Globalizing World). Nova Science Pub Inc, New York (2011)
9. Hughes, A. and Hughes, J.: Testing for Language Teachers (Third Edition). Cambridge University Press, New York (2020)
10. Peng, R.Z., Zhu, C.G. and Wu, W. P.: Study on the quality of the Intercultural Competence Test: A Rasch model analysis. Foreign Language World 5, 12-19，79 (2022)
11. Chang, X.P. and Zhu, C. Q.: Text-based Grammatical Completion: Its Validity Analysis and Intended washback. Contemporary Foreign Language Studies 05, 16-20+47+110 (2017)

12. Chang, X.P.: Strategies and Patterns in Processing a Grammar-completion Task. Journal of Beijing International Studies University 01, 96-110 (2020)
13. Xu, J., Tang, S. Z. and Y. H.: A Study of the Assessment Value of Cloze from the Perspective of Gaokao Assessment Framework. Journal of Tianjin Normal University (Elementary Education Edition) 04, 46-51 (2022)
14. Oliveira, K., AAAA, D. S., Boruchovitch, E. and Rueda, F.: Reading Comprehension: Differential Item Functioning Analysis of a Cloze Test. Ppar Research 25(2), 221-229 (2012)
15. Chen, Y. and Zhou, R.: The application of Rasch Model in the quality analysis of cloze items in NMET. Foreign Language Testing and Teaching 1, 39-47，64 (2018)
16. Gan Q., Yang, Y. and Li, Y. Q.: Rasch model based comparative study on validity between multiple-choice cloze questions and banked cloze questions in TEM-4. Journal of Chengde Petroleum College 20(4), 49-53 (2018)
17. Baghaei, P. and Ravand, H.: Method Bias in Cloze Tests as Reading Comprehension Measures (Article). SAGE Open, 9(1) (2019)
18. Mokshein, S. E., Ishak, H. and Ahmad, H.: The use of Rasch measurement model in English testing. Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan 1(1), 16-32 (2019)
19. Embretson, S. E. and Hershberger, S. L.: The New Rules of Measurement: What Every Psychologist and Educator Should Know. Mahwah, NJ: Lawrence Erlbaum Associates, Inc (1999)
20. Bachman, L. F. and Palmer, A.: Language testing in practice. Oxford University Press, New York (1996)
21. Zhang, S. K. and Zhao, T.: A study of the content validity of fast reading comprehension section of the HSK test paper in 2007. Journal of Southwest Minzu University (Humanities and Social Sciences Edition) 202, 107-110 (2011)
22. Xu, Y. and Deng, Y. L.: Examining the Content Validity of Translation Task in Large-Scale Language Tests: A Case of CET. Shandong Foreign Language Teaching 44(3), 26-37 (2023)
23. Gu, X. D., Li, Z. F. and Zhang, S. K.: A study of the content validity of fast reading section of the English CET-4. Journal of Southwest Minzu University (Humanities and Social Sciences Edition) 30(01), 258-263 (2009)
24. Zhang, S. K. and Zhao, T.: A study of the content validity of English reading comprehension section of the final exam in Southwest Minzu University. Journal of Southwest Minzu University (Humanities and Social Sciences Edition) 201, 30-33 (2010)
25. Ministry of Education of the People's Republic of China.: [The English Curriculum Standard for Compulsory Education] (2022 Edition). Beijing Normal University Publishing Group, Beijing (2022)
26. Alderson, J. C.: Assessing Reading. Cambridge University Press, Cambridge (2000)
27. Flesch, R.: A new readability yardstick. Journal of Applied Psychology 32(3), 221-233 (1948)
28. Bachman, L. F.: Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. TESOL Quarterly 19(3), 535–556 (1985)
29. Holsti, O. R.: Content Analysis for the Social Sciences and Humanities. Random House (1969)
30. Yuan, J. Study on the quality of the university English placement test. Journal of Southeast University (Philosophy and Social Science) 201, 142-145 (2016)
31. Linacre, J. M.: A User's Guide to WINSTEPS MINISTEP: Rasch-Model Computer Programs. UpToDate, from https://www.winsteps.com/a/Winsteps-Manual.pdf, last accessed 2023/6/25

32. Calkin, C. J., Numbers, K., Sachdev, P. S., Brodaty, H. and Medvedev, O. N.: Measuring distress in older population: Rasch analysis of the Kessler Psychological Distress Scale. Journal of Affective Disorders 330, 117-124 (2023)
33. Ministry of Education of the People's Republic of China.: Circular on Strengthening the Management of School Examinations during Compulsory Education. Communique of Ministry of Education of the People's Republic of China 11, 29-30 (2021)

Appendix

Stephen Hawking was a British scientist. Many people think he was the 36 scientist since Albert Einstein.

As a scientist, Hawking was most 37 for studying black holes. He thought that the universe started with the Big Bang and will 38 in black holes. Black holes suck up everything around 39 — even light.

Hawking's studies changed 40 people look at the universe. He also wrote 41 to help more people understand the universe. A Brief History of Time is his most popular book.

42 , people looked up to Hawking not only because he was smart but also because he had a strong will.

Hawking had a serious 43 that started when he was 21. He could not walk or talk. However, his illness didn't 44 him from living a meaningful life. He traveled around the world and wrote many books. He 45 played himself on many TV shows.

Just as his children said after his death, Hawking's works and spirit will live on for many years.

36. A. tallest	B. poorest	C. happiest	D. greatest
37. A. worried	B. famous	C. hungry	D. late
38. A. call	B. cut	C. hand	D. end
39. A. it	B. him	C. them	D. us
40. A. why	B. that	C. how	D. when
41. A. books	B. messages	C. letters	D. songs
42. A. Instead	B. Because	C. Though	D. However
43. A. idea	B. friend	C. illness	D. life
44. A. stop	B. leave	C. help	D. protect
45. A. just	B. even	C. never	D. well

Answers:
36-40: DBDCC
41-45: ADCAB

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.