

# Item Analysis of a Multiple-Choice Rational Cloze Test

Katsumasa SHIMADA

## 1. Introduction

The basic unit of any test is the test item. At the item-writing stage we focus on the degree to which each item is properly written so that it measures what it intends to measure. *A priori* item analysis is concerned with the validity of each item. Henning (1987) and Brown (1996) list a number of general guidelines for language testers to take into consideration. It is not until the test has been tried out or piloted on an appropriate sample of examinees, and until the data of the performance were collected and analysed, that we can be certain of whether each item is good or poor. In order to make a test better, we must eliminate or improve some items which were judged as weak by *a posteriori* item analysis.

Some statistics have been used in order to find the poor items. Traditional item analysis based on classical test theory (CTT) have employed item difficulty, item discriminability with sample separation, point-biserial correlation coefficients. Latent trait measurement (also called item response theory (IRT)) is an extension of CTT, which was developed to overcome some problems entailed in CTT: sample-dependent and test-dependent. The latent trait theory describes the relation-

ship between testees' performances on a particular item and the latent trait by using an item response model. The model is a mathematical function which relates the probability of a correct response on a particular test item to person ability and to item difficulty. The items whose response pattern does not fit the model can be classified as misfitting items.

The purpose of this investigation is to submit data from a multiple-choice (MC) rational cloze test that Shimada (1997) invented to CTT item analysis and IRT measurement in order to identify weak items and to compare the weak items identified by the two different approaches.

## **2. Method**

### **2.1 Subjects**

One hundred and twenty-seven college students enrolled in English I at a private university participated in this investigation. Thirty-one freshmen majored in economics, 28 sophomores or juniors were business administration majors, and 68 freshmen or sophomores majored in sociology.

### **2.2 Materials and procedure**

A MC rational cloze test was developed (see Appendix A). A passage from Chavez-Oller, Chihara, Weaver and Oller's (1985) appendix was selected for its appropriate difficulty and content. The topic of the passage deals with the problems college students might encounter in leaving home for the first time. The first three sentences were left intact as the lead-in. Blanks were rationally placed within a range of 6 to 28 words (at an average ratio of 1: 11).

According to the range of context required for closures, words de-

## Item Analysis of a Multiple-Choice Rational Cloze Test

leted can be classified into four categories: within clause (syntax); within clause (lexis); across clause, within sentence; across sentence, within text (based on Bachman, 1985; Jonz, 1990). Seven words were deleted in each of the four categories described above, resulting in that the test involves twenty-eight blanks in total. The categorisation of each blank and key is shown in Appendix B.

The candidates were required to choose a correct answer from four options. They were allowed up to 60 minutes to choose from the options the word which was felt to fit the context best.

Test Data Analysis Program (TDAP) version 1.0 developed by Ohtomo and Nakamura (1996) was used to analyse the data.

### **3. Result and Analysis**

#### **3.1 Comparison between CTT and IRT**

Item difficulty was determined as the proportion of correct responses. Items with a proportion of correct answers less than .33 or greater than .67 were rejected as weak items (Henning, 1987: 50; Perkins and Miller, 1984: 23; Reynolds, Perkins and Brutton, 1994: 3)<sup>1</sup>. Two different methods were employed to compute item discriminability: item discriminability with sample separation and point-biserial correlation. For the sample separation method, 27 per cent of the total sample for both the upper and the lower group were used. Items with a discriminability index of less than .30 were rejected as weak (Brown, 1996: 70)<sup>2</sup>. Items with a point-biserial correlation coefficient of less than .25 were rejected as weak (Henning, 1987: 53; Perkins and Miller, 1984: 23; Reynolds, Perkins and Brutton: 1994: 4).

The item analysis indices are shown in Table 1. The mean for each index falls within the acceptable range. Fifty-seven per cent of the items

had to be rejected because their proportion correct was greater than .67 or less than .33. Forty-six per cent were weak items according to the sample separation indices, and 32 per cent were rejected because they

Table 1: Summary of Item Analysis

category	no.	difficulty	x	discrimination	x	r pbi	x	xxx	difficulty	t	x
1-L-b	26	0.213	x	0.265	x	0.237	x	xxx	2.066	1.461	
1-L-c	21	0.890	x	0.353		0.390			-1.722	-3.561	?
1-L-c	23	0.457		0.412		0.355			0.800	0.100	
1-L-f	4	0.244	x	0.353		0.280			1.867	1.013	
1-L-f	25	0.323	x	0.294	x	0.316			1.432	0.368	
1-L-h	6	0.913	x	0.235	x	0.249	x	xxx	-2.020	-2.150	?
1-L-h	27	0.646		0.353		0.341			-0.063	0.824	
1-S-c	3	0.370		0.235	x	0.241	x		1.199	1.016	
1-S-d	15	0.551		0.118	x	0.127	x		0.377	1.767	
1-S-d	18	0.512		0.529		0.381			0.554	-0.180	
1-S-e	10	0.953	x	0.118	x	0.263			-2.743	-2.184	?
1-S-f	19	0.756	x	0.441		0.470			-0.654	-2.304	?
1-S-f	28	0.394		0.118	x	0.086	x		1.088	2.518	x
1-S-h	7	0.906	x	0.176	x	0.271			-1.914	-2.338	?
2-a	8	0.291	x	0.706		0.602			1.597	-2.337	?
2-b	20	0.638		0.647		0.580			-0.025	-2.323	?
2-c	12	0.717	x	0.588		0.439			-0.428	-1.747	
2-c	24	0.496		0.147	x	0.180	x		0.624	1.343	
2-e	2	0.764	x	0.353		0.349			-0.702	-0.576	
2-e	17	0.874	x	0.176	x	0.236	x	xxx	-1.553	-0.363	
2-g	1	0.528		0.382		0.380			0.483	-0.189	
3-a	13	0.157	x	0.235	x	0.252			2.476	1.438	
3-c	22	0.953	x	0.118	x	0.235	x	xxx	-2.743	-3.008	?
3-d	9	0.472		0.353		0.301			0.729	0.535	
3-d	14	0.520		0.529		0.443			0.518	-0.798	
3-e	5	0.827	x	0.147	x	0.184	x	xxx	-1.136	0.808	
3-e	11	0.480		0.749		0.520			0.694	-1.068	
3-e	16	0.780	x	0.441		0.414			-0.802	-1.333	
mean		0.594	57%	0.342	46%	0.326	32%	18%	0.000	-0.474	32%
lower		0.581		0.286		0.286			0.019	-0.261	
higher		0.607		0.398		0.365			-0.019	-0.687	
h-1 diff.		0.026		0.112		0.079			-0.038	-0.426	

## Item Analysis of a Multiple-Choice Rational Cloze Test

produced point-biserial correlations less than .25. Five out of 28 items were rejected by item difficulty, item discriminability with sample separation and point-biserial methods: 5, 6, 17, 22 and 26.

As for IRT, Approximation Procedure (PROX) was used for the Rasch one-parameter calibration of test items. After person ability for each person and item difficulty for each item were calibrated, misfitting indices were calibrated on the assumption that the likelihood of success or failure, or degree of misfit, in responding to an item is shown as a function of the distance of person ability from item difficulty. T-value was used to express the measure of misfit. Items with t-value of 2.00 or above are considered misfits to the model, and a negative t-value of -2.00 or below is considered overfits to the model.<sup>3</sup> (Henning, 1987: 123). Final item difficulty and t statistics are also shown in Table 1. The result shows that 32 per cent (9 out of 28) of the items were misfits: 6, 7, 8, 10, 19, 20, 21, 22, 28.

Item type in terms of misfitting can be categorised into four as shown in Table 2. Eight items were accepted by both methods (Type A), while other 8 items were rejected by both methods (Type D). Eleven items were rejected by CTT but accepted by IRT approach (Type C), on the other hand, 1 item was accepted by CTT but rejected by IRT method (Type B). Items shown in the parentheses were rejected at least one classical item analysis.

Table 2: Classification of the Test Items

Type	CTT	IRT	Nos.	Items
A	accept	accept	8	1, 4, 9, 11, 14, 18, 23, 27
B	accept	reject	1	20
C	reject	accept	3(8)	5, 17, 26, (2, 3, 12, 13, 15, 16, 24, 25)
D	reject	reject	2(6)	6, 22, (7, 8, 10, 19, 21, 28)

We favour item misfit statistics based on IRT over CTT for detecting weak items. In CTT, item difficulty is sample dependent: it is stable only for groups of similar levels. Ability measurement according to CTT is dependent on the test items given. In IRT, the Rasch model uses a logit function to convert item difficulty to an equal interval scale so that it is independent of ability differences of any particular sample of examinees.

### 3.2 Comparison between higher-order and lower-order items

In a cloze test, whether it is a fixed-ratio or rational, as Bachman (1985) argues, “not all deletions in a given cloze passage measure exactly the same abilities” (p.535), resulting in that testees perform differently on different test items. It is hypothesised that item difficulty will be ordered according to level of context required for closure, with ‘within clause’ being the easiest and ‘across sentence, within text’ being the most difficult.

We divided all the test items into two categories: lower-order and higher-order items. ‘Within clause (syntax)’ and ‘within clause (lexis)’ fall into lower-order items, while ‘across clause, within sentence’ and ‘across sentence, within text’ are categorised as higher-order items.

The hypothesis, namely that the higher-order items are more difficult than lower-order items, was examined by comparing the two sets of mean. In order to investigate whether there is a significant difference between the two sets of mean, one-tail paired sample t-test was used. The result indicates that there is not a significant difference between the two categories (CTT:  $t = -0.391$ ,  $df = 13$ ; IRT:  $t = 0.087$ ,  $df = 13$ ). Item analysis also revealed that there is no difference of item discrimination ( $t = -1.560$ ,  $df = 13$ ) and point-biserial correlation ( $t = -1.604$ ,  $df = 13$ )

## Item Analysis of a Multiple-Choice Rational Cloze Test

between the two categories. The reason why the hypothesis was rejected may be because MC format offers options, which encourage blind guessing. It is possible to say that difficulty level of each item is determined by what options are given.

### 3.3 Correlation between difficulty and misfit

Rasch analysis showed a curious effect. Figure 1 shows a result from the Rasch misfit analysis of the 28 items in the cloze test. The scattergram shows a moderate positive correlation between item misfit and final item difficulty of the 28 items within the cloze test ( $r=0.668$   $t=4.578$ ,  $df=26$ ,  $p<.001$ ). The result indicates that the degree of misfitting gets higher as the item becomes more and more difficult. It implies that random guessing increases with increasing difficulty. As Heaton (1988) points out, it is true that "candidates rarely make wild guesses" (p.27), but poor candidates might make random guessing for difficult items.

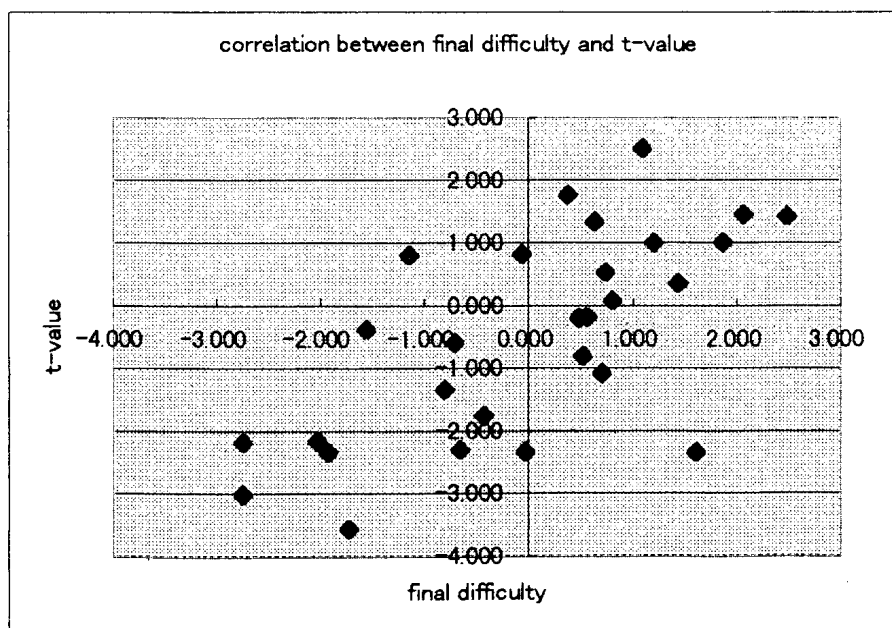


Fig.1: Correlation between Difficulty and Misfit

#### 4. Conclusion

The basic unit of any test is the test item. In order to make a test better, we need to identify weak items and modify them. We have identified and compared the weak items of a MC rational cloze test by CTT and IRT measurement. We have found that some items were accepted by CTT but rejected by IRT method, and others were judged as weak by CTT but considered to be acceptable by IRT approach. On the assumption that item difficulty will be ordered according to level of context required for closure, higher-order and lower-order items were compared. The result shows that higher-order items are not always more difficult than lower-order items. We also noted that item difficulty correlated moderately with item misfit. This result implies that random guessing increases with increasing difficulty.

#### Notes

1. Brown (1996) states that "those that fall in a range between 0.3 and 0.7 are usually considered acceptable" (p.69-70). Taking random guessing into consideration, the acceptable range might be set between 0.46 and 0.80.
2. Hughes (1987) states that "items that show correlations of 0.3 or more are generally considered satisfactory" (p.160).
3. Henning (1987) points out that overfitting items do "not constitute grounds for the rejection of items". However, he also states that there is some suspicion that overfitting items may "exhibit less stable difficulty or ability estimates on recalibration" (p.123).

#### References

- Bachman, L.F. (1985) "Performance on Cloze Tests with Fixed-Ratio and Rational Deletions" *TESOL Quarterly* 19: 535-556.
- Brown, J.D. (1996) *Testing in Language Programs*. Prentice Hall Regents.



## Item Analysis of a Multiple-Choice Rational Cloze Test

- Chavez-Oller, M.A., T.Chihara, K.A.Weaver, and J.W.Jr.Oller (1985) "When are cloze items sensitive to constraints across sentences?" *Language Learning* 35: 181-206.
- Heaton, J.B. (1988) *Writing English Language Tests*. Longman.
- Henning, G. (1987) *A Guide to Language Testing*. Newbury House.
- Hughes, A. (1989) *Testing for Language Teachers*. Cambridge University Press.
- Jonz, J. (1990) "Another Turn in the Conversation: What Does Cloze Measure?" *TESOL Quarterly* 24: 61-83.
- Ohtomo, K. and Y.Nakamura (1996) Test Data Analysis Program version 1.0. Taishukan Shoten.
- Perkins, K., L.D.Miller (1984) "Comparative analyses of English as a Second Language reading comprehension data: classical test theory and latent trait measurement" *Language Testing* 1: 21-34.
- Reynolds, T., K.Perkins, and S.Brutten (1994) "A comparative item analysis study of a language testing instrument" *Language Testing* 11: 1-13.
- Shimada, K. (1997) "The Validity of Multiple-Choice Rational Cloze Tests" *CELES Bulletin* 27: 139-146.

### Appendix A: Multiple-Choice Rational Cloze Test

DIRECTIONS: Fill in the blanks with a suitable word. Choose the correct answer from the options; {A, B, C, D}.

#### Joe leaves for college

Joe is a freshman, and he is having all the problems that most freshmen have. As a matter of fact, his problems started before he even left home. He had to do a lot of things he didn't like to do just because he was going to go away to college. He had his eyes examined and he had his cavities filled, <sup>1</sup>{A.because, B.although, C.as, D.and} he hates to go to a <sup>2</sup>{A.college, B.dentist, C.shop, D.home}, and he got his watch fixed <sup>3</sup>{A.to, B.from, C.by, D.with} a neighbourhood jeweller. Then, at his mother's suggestion, he had his father's tailor <sup>4</sup>{A.borrow, B.buy, C.present, D.measure} him for a suit. He didn't have a suit made, though, because his <sup>5</sup>{A.father, B.tailor, C.mother, D.friend} wouldn't let him order one. "You're still growing, son," he said.

“You’re growing <sup>6</sup>{A. such, B. very, C. so, D. as} fast that you’d outgrow a suit in no time. Buy yourself a pair <sup>7</sup>{A. of, B. with, C. for, D. on } slacks and a sports jacket. Klein’s has such a large selection that I’m sure you will find something you like <sup>8</sup>{A. one, B. it, C. here, D. there}.” Joe’s father always suggested Klein’s for <sup>9</sup>{A. clothes, B. example, C. her, D. good}.

Joe went to Klein’s in order <sup>10</sup>{A. that, B. to, C. in, D. of} please his father, but he didn’t <sup>11</sup>{A. like, B. say, C. find, D. take} anything that he liked there so <sup>12</sup>{A. she, B. he, C. small, D. that} went to another store to buy <sup>13</sup>{A. for, B. another, C. pair, D. the} slacks. He took them out of the box as soon as he got <sup>14</sup>{A. home, B. out, C. store, D. up} so that his father wouldn’t notice <sup>15</sup>{A. there, B. where, C. that, D. it} they came from.

When Joe was all ready to leave for school, his <sup>16</sup>{A. father, B. mother, C. teacher, D. friend} suggested that he visit all his relatives. “What do you want me to do that for?” he asked, and she <sup>17</sup>{A. rejected, B. asked, C. suggested, D. answered}, “To say good-bye.” She made him go to see his cousins in Bellevue, <sup>18</sup>{A. but, B. or, C. and, D. then} his Uncle Ned in Plaintown and his Great-Aunt Lizzie who lives in the southern part of the state. He <sup>19</sup>{A. had, B. would, C. didn’t, D. was} want to visit all those people, <sup>20</sup>{A. so, B. but, C. and, D. when} he did it anyway because of <sup>21</sup>{A. their, B. his, C. her, D. Klein’s} mother’s insistence.

On the day that <sup>22</sup>{A. he, B. she, C. they, D. Klein} left for college, his sister helped <sup>23</sup>{A. for, B. with, C. her, D. him} pack his clothes. She let him borrow her suitcase because he didn’t have <sup>24</sup>{A. it, B. one, C. that, D. suitcase} of his own. When everything was all ready, he got his father to <sup>25</sup>{A. meet, B. bring, C. show, D. drive} him to the station, and the <sup>26</sup>{A. whole, B. all, C. rest, D. his} family went along. Of course, his mother insisted on kissing him good-bye in <sup>27</sup>{A. case, B. front, C. terms, D. spite} of his embarrassment. As soon as the train pulled into the station, Joe jumped on and hurriedly found his seat. By the time it pulled out, he <sup>28</sup>{A. had, B. started, C. was, D. could} already contemplating his new life away from home.

《注 cavity : 虫歯の穴 jeweler : 貴金属商 insistence : しつこさ  
embarrassment : 困惑 contemplate : 熟考する》

## Item Analysis of a Multiple-Choice Rational Cloze Test

### Appendix B: Categorisation for Cloze Items and Keys

No.	category	Key
1.	2-g	B (although)
2.	2-e	B (dentist)
3.	1-S-c	C (by)
4.	1-L-f	D (measure)
5.	3-e	A (father)
6.	1-L-h	C (so)
7.	1-S-h	A (of)
8.	2-a	D (there)
9.	3-d	A (clothes)
10.	1-S-e	B (to)
11.	3-e	C (find)
12.	2-c	B (he)
13.	3-a	D (the)
14.	3-d	A (home)
15.	1-S-d	B (where)
16.	3-e	B (mother)
17.	2-e	D (answered)
18.	1-S-d	C (and)
19.	1-S-f	C (didn't)
20.	2-b	B (but)
21.	1-L-c	B (his)
22.	3-c	A (he)
23.	1-L-c	D (him)
24.	2-c	B (one)
25.	1-L-f	D (drive)
26.	1-L-b	A (whole)
27.	1-L-h	D (spite)
28.	1-S-f	C (was)

### Appendix C: Formulae

As for the CTT analysis, the formulae used are as follows:

item difficulty:  $p = \frac{\sum C_r}{N}$

where,  $p$  = item difficulty

$\sum C_r$  = the sum of correct responses

$N$  = the number of examinees

item discrimination with sample separation:  $D = \frac{U_c - L_c}{n}$

where,  $D$  = item discriminability

$U_c$  = the number of correct responses in the group

$L_c$  = the number of correct responses in the lower group

$n$  = the number of examinees in the upper (lower) group

point-biserial correlation:  $r_{pbi} = \frac{(\bar{x}_p - \bar{x}_q)\sqrt{pq}}{s_x}$

where,  $r_{pbi}$  = the point biserial correlation

$\bar{x}_p$  = the mean total score for examinees who pass the item

$\bar{x}_q$  = the mean total score for examinees who fail the item

$s_x$  = the standard deviation of test scores

$p$  = the proportion of examinees who pass the item

$q$  = the proportion of examinees who fail the item

In the IRT analysis, the following formulae were used:

$$p = \frac{\exp(b-d)}{1 + \exp(b-d)}$$

where,  $p$  = probability

$b$  = person ability

$d$  = item difficulty

$$z = \frac{x-p}{\sqrt{p(1-p)}}$$

where,  $z$  = the standardised residual

$x$  = the observed item response

$p$  = the expected response

$$t = \left( \ln \frac{\sum z^2}{df} + \frac{\sum z^2}{df} - 1 \right) \sqrt{\frac{df}{8}}$$

where,  $t$  = misfit statistics

$z$  = the standardised residual

## **Item Analysis of a Multiple-Choice Rational Cloze Test**

**Katsumasa SHIMADA**

### **Abstract**

The basic unit of any test is the test item. In order to make a test better, we need to identify weak items and modify them. The purpose of this investigation is to identify and compare the weak items of a cloze test by classical test theory (CTT) and item response theory (IRT) measurement.

A four-option multiple-choice (MC) rational cloze test was developed. According to the range of context for closures, words deleted can be categorised into two types: higher-order and lower-order items. Item difficulty, item discriminability with sample separation, and point-biserial correlation were used for CTT analysis. As for IRT analysis, PROX was employed for the Rasch one-parameter calibration of test items.

We found that some items were accepted by CTT but rejected by IRT method, and others were judged as weak by CTT but considered to be acceptable by IRT approach. In order to ascertain whether item difficulty will be ordered according to level of context required for closure, higher-order and lower-order items were compared. The result shows that higher-order items are not always more difficult than lower-order items. It is claimed that a MC format encourages wild guessing so that difficulty level of each item is determined by what options are given. We

also noted that item difficulty correlated moderately with item misfit. This result implies that random guessing increases with increasing difficulty.