The Reading Matrix
Vol. 5, No. 1, April 2005

A FRESH LOOK AT THE EVALUATION OF ORDERING TASKS IN READING COMPREHENSION: *WEIGHTED MARKING PROTOCOL*

Salim Razi
Email: salimrazi@comu.edu.tr

Abstract

After briefly discussing techniques such as 'the cloze test' and 'gap-filling' employed in assessing reading, the main focus of the paper resides in the scoring process of 'ordering tasks', where students are asked to re-arrange the order of sentences given in incorrect order. Since the evaluation of such tasks is quite complex, Reading Teachers rarely use them in their tests. According to Alderson, Reading Teachers frequently tend to mark these tasks either wholly right or totally wrong since the partial marking process is very time-consuming. In this respect, the readers of this paper will be introduced to a new approach, namely *Weighted Marking Protocol*, developed by the author for ordering tasks in order to achieve a fairer evaluation. This new approach makes it possible for Reading Teachers to reward their students for all right answers and not penalize them just for a single mistake.

Language Testing

Testing takes places at every stage of our lives, as well as in the language learning process. McNamara (2000) argues that there are a number of reasons for administering language tests, which play a powerful role in an individual's social and working life. Language teachers work with language tests since they need to evaluate their learners; language testing is also carried out for research purposes.

Language tests have been categorised under four headings by Alderson (1996) and Harmer (2001) as 'placement', 'diagnostic', 'progress or achievement', and 'proficiency', that is; to place learners in the right classes according to their level, to find out where learners have difficulties in a language course, to explore progress or reflect how well the students are learning a particular subject, and to give a general idea about a student's proficiency in the target language. This paper does not go into the details of language testing, however, the main focus is on how to assess reading comprehension and especially how to solve the problem of partial marking in ordering tasks.

Testing Reading

Alderson (1996) proposes that Reading Teachers feel uncomfortable in testing reading. To him, although most teachers use a variety of techniques in their reading classes, they do not tend to use the same variety of techniques when they administer reading tests. However, Alderson implies that there are similarities between the design of classroom activities and test items, so it is reasonable to expect that any teaching activity can easily be turned into a testing item, and vice versa. Despite the variety of testing techniques, none of them is subscribed to as the best one. Alderson (1996, 2000) considers that no single method satisfies reading teachers since each teacher has different purposes in testing.

Alderson (2000) identifies three different types of questions based on Pearson and Johnson (1978). According to this categorisation, 'textually explicit' questions are the ones in which the respondent is able to find both the question information and the correct answer. On the other hand, through 'textually implicit' questions the respondents are expected to find the answer by combining information across sentences. The last type is 'script-base', (or 'scriptually implicit') questions, in which the respondent needs to refer to her background knowledge since the text does not contain the correct answer itself. The respondents' schemata should fit the tester's schema in order to answer scriptually implicit questions.

Techniques for Testing Reading

As there appears to be no best method for testing reading (Alderson, 2000), then reading teachers should be aware of what they need to test in terms of selecting the most appropriate testing method for their students; discrete-point techniques when they intend to test a particular subject at a time, or integrative techniques when the aim is to see the overall picture of a reader. Following is a brief survey of the most frequently used testing techniques.

Cloze Test Versus Gap-Filling Test

The distinction between 'the cloze test' and 'gap-filling test' has been neglected and as a result, many professionals who deal with these terms tend to use them interchangeably. Alderson (2000: 207) defines the former as "...typically constructed by deleting from selected texts every n-th word ... and simply requiring the test-taker to restore the word that has been deleted". Alderson states that 'n' usually differs from intervals of every 5th word to every 12th word however; 'n' is a number between 5 to 11 according to Weir (1990) and just 5 to 7 according to McNamara (2000). On the other hand, two decades prior to this definition, Alderson (1979) defines cloze procedure in three different ways. It is defined as "… the systematic deletion of words from text…" *Systematic* is not defined here. The second definition divides *systematic* into "…two types of systems: either a random (or, better, pseudo-random) deletion of words, or a rational deletion". The last definition is known as "… the deletion of every fifth word from the text" (Alderson, 1979: 219). Nevertheless, according to Cohen (2001), the origins of this technique date back to the end of the 19th century and he points out that research by Chávez-Oller et al. (1985) indicates "... that cloze [is] sensitive to constraints *beyond* 5 to 11 words on either side of a blank" (Cohen, 2001: 521). Alderson (2000) states that, according to research, in order to achieve reliable results there should be at least 50 deletions in a cloze test.

To prepare a cloze test, the tester is required to decide which word to delete first; the other deletions follow this systematically, such as the deletion of every 6th word after the first deletion *(See Appendix A)*. Cloze tests are easy to prepare but since testers cannot control which words to delete, except the first one, they do not know confidently what their tests measure (Alderson, 2000). Cohen (1998) concludes that cloze tests do not assess global reading ability but they do assess local-level reading. Such tests are easy to mark since the testers expect to see the words that they deleted beforehand. They are also recommended to accept other answers which make sense.

In order to prepare a gap-filling test, however, the tester is required to decide which words to delete one by one *(See Appendix B)*. The deletion of the words is not based on any system, yet preparing a gap-filling test is as easy as preparing a cloze test. The decision of which words to delete is done on a rational basis so the tester can control the test. However, gap-filling tests were criticised by Weir (1993) since this type of test does not require

extracting information by skimming. Since the tester knows which words have been deleted in a gap-filling test, she may have a tendency to assume that these words are essential to meaning (Alderson, 2000). The marking process of gap-filling tests is almost the same as the one in cloze test process.

C-Tests

As an alternative integrated approach (Weir, 1990, 1993) the *C-Test* is acceptable in that it "... is based upon the same theory of closure or reduced redundancy as the cloze test" (Alderson, 2000: 225). Test-takers are asked to restore the second half of every second word deleted beforehand *(See Appendix C)*. Alderson (2000) and Cohen (2001) point out that C-tests are more reliable and valid than cloze tests in terms of assessing but are thought to be more irritating than cloze tests. In the marking process, the testers do not face difficulties since they expect to see the restored word.

The Cloze Elide Test

Another alternative integrated approach is called the *Cloze Elide Test* by Alderson (1996, 2000). This technique was introduced as the 'Intrusive Word Technique' and is also called as "...'text retrieval', 'text interruption', 'doctored text', 'mutilated text' and 'negative cloze'..." (Alderson, 2000: 225). The tester inserts words and the test-taker is asked to find the words that do not belong to the text *(See Appendix D)*. It is important to be sure that the inserted words do not belong to the text. Otherwise, the test-takers will not be able to identify the inserted words.

Weir (1993) proposes that C-tests are seen as puzzles rather than language tests in some part of the world where these tests are used in national examinations. This type of test is likely to be used, not for comprehension, but for a measure of comprehension. "The number of correctly identified items was taken as a measure of reading speed" (Alderson, 2000: 226).

Multiple-Choice Test Items

Another technique that Weir (1990, 1993), Alderson (1996, 2000), Ur (1996), Cohen (1998), and Hughes (2003) discuss is 'multiple-choice'; a common device for text comprehension. Ur (1996: 38) defines multiple-choice questions as consisting "... of a stem and a number of options (usually four), from which the testee has to select the right one". Alderson (2000: 211) states that multiple-choice test items are so popular because they provide testers with the means to control test-takers' thought processes when responding; they "… allow testers to control the range of possible answers …"

On the other hand, he argues that distractors may trick deliberately, which results in a false measure. Also, being a good reader does not guarantee being successful in a multiple-choice test since this type of test requires a separate ability. Cohen (1998: 97) also criticises the way that test-takers do "…not necessarily link the stem and the answer in the same way…" that the tester assumes. So the test-takers may reach the correct answer by following false reasoning. Alderson (2000) points out that test-takers are provided with possibilities that they might not otherwise have thought of.

However time-consuming it is to prepare a multiple-choice test, it is easy to evaluate, as it is a machine-markable technique. Weir (1990) mentions that multiple-choice questions are fashionable since marking them is totally objective.

Summary Tests

Cohen (1998) defines summarization tasks as more direct, since test-takers are required to use the strategies that they do not tend to use in non-test conditions. In *the free-recall test* (also called *immediate-recall test*), the test-takers are given a text, asked to read it, then leave it and write down everything they can remember. Alderson (1996) argues that free-recall tests are usually scored according to Meyer's (1975) *recall scoring protocol*, where the text is divided into idea units and the relationship between these idea units is examined. On the other hand, the *Summary Test* is accepted as a more familiar variant in which the test-takers are expected to summarise the main ideas of the text they read beforehand. It is possible to score them like free-recall tests or the summary can be scored on a scale.

The problem with summary tests is whether the writing skill or the reading skill is being tested. The solution to this problem proposed by Alderson (2000) is asking the test-takers to write the summary in their first language or by presenting a number of summaries and asking them to select the best summary. The latter is appropriate if the tester does not speak the same native language as the test-takers or if the aim is to test first language reading.

The Gapped Summary

To overcome the problems of Summary Tests, another approach called the *Gapped Summary* was introduced by Alderson (2000) in which the test-takers read a text for a limited time period, and then read a summary of the same text that includes some missing key words without referring to the text. Students must restore the missing words from the original text. The advantage of this technique is that the test-takers are not tested for their writing abilities. Scoring can be done by following the same process as in cloze or gap-filling tests.

Dichotomous Items (True-False Technique)

Test-takers are asked to state whether the given statement is true or false by referring to the text. This technique is well known as the *true or false* technique. According to Alderson (2000), the ease of construction makes this technique popular.

Alderson (2000) and Hughes (2003) argue that the problem with this technique is a 50% possibility of guessing the right answer without comprehending the target text. The tester may reduce this chance to 33.3% by adding one more statement such as 'not given'. However, such statements actually tend to test the ability of inferring meaning rather than comprehension. Testers therefore need to make sure what they intend to measure. Another alternative to solve the guessing problem is asking the test-takers firstly to state whether the statements are true or false, and secondly asking them to correct the false ones.

Not only is designing such tests easy (Ur, 1996) but scoring is quite easy. Such tests can be designed as machine-markable items in multiple choice format (Alderson, 2000).

Editing Tests

It is also possible to present tests in which errors have been introduced deliberately similar to a proof-reading task in real life (Alderson, 2000). The nature of the error identifies

whether it is testing the reading skill or linguistic ability. In these *editing tests*, the test-takers are asked to identify the errors and then correct them. Testers can manipulate the test by deleting a word from the text without replacing it with a gap so test-takers are required to find out where the missing word is first, and then write it in the place it belongs. Although similar to editing tasks, professionals criticise these since they provide wrong information to learners of a foreign language (Sezer, 2002). In the scoring process, the tests takers may be given points for each error that they identify.

Questions and Answers

In order to check the comprehension of any reading text, the teacher may ask 'open-ended' or 'closed' questions related to the text. In open-ended questions, test-takers are asked to write down every detail related with the question. On the other hand, Ur (1996) argues that if test-takers are provided with closed questions, the marking process will be easier since there will be fewer possible correct answers. The tester needs to prepare a detailed answer key for such tests. Also it is important to decide whether to take into account or ignore grammatical mistakes.

Short-Answer Tests

Weir (1993) points out that short-answer tests are extremely useful for testing reading comprehension. According to Alderson (1996, 2000), 'short-answer tests' are seen as 'a semi-objective alternative to multiple choice'. Cohen (1998) argues that open-ended questions allow test-takers to copy the answer from the text, but firstly one needs to understand the text to write the right answer. Test-takers are supposed to answer a question briefly by drawing conclusions from the text, not just responding 'yes' or 'no'. The test-takers are supposed to infer meaning from the text before answering the question. Such tests are not easy to construct since the tester needs to see all possible answers. Hughes (2003: 144) points out that "[t]he best short-answer questions are those with a unique correct response". However, scoring the responses depends on thorough preparation of the answer-key. Hughes (2003) proposes that this technique works well when the aim is testing the ability to identify referents.

Matching

In this technique, test-takers are provided with two sets of stimuli that need to be matched against each other. Multiple-matching items are similar to multiple-choice test items since there are distractors. In multiple-matching tests, each item acts as a distractor except one. According to Alderson (2000: 219) since "... there is only one final choice", giving more alternatives than the matching task requires is more sensible. He also mentions that they are difficult to construct because of the need to prevent unintentional choices. Matching questions have also been criticised since they offer distractors that the test-takers would not otherwise consider. The scoring process of this task is easy as the test-takers gain points for each correct matching.

Ordering Tasks

Through 'ordering tasks', test-takers are asked to put the scrambled words, sentences, paragraphs or texts into correct order. Although they test "... the ability to detect cohesion, overall text organisation or complex grammar..." (Alderson, 2000: 221) there are problems in administering this test type. Alderson argues that firstly, the test-takers may propose another sensible order different from the tester's. The tester is recommended to accept all unexpected but sensible orders or rewrite the test in order to provide only one possible correct order.

The second problem in ordering tasks occurs while scoring. The tester will probably have difficulties in giving marks to those who answer only half the test in the correct order. Usually, these answers are marked wholly correct or wholly wrong. Alderson reports the following comment on evaluation of ordering elements:

> "... the amount of effort involved in both constructing and in answering the item may not be … worth it, especially if only one mark is given for the correct version" (Alderson et al., 1995: 53 in Alderson, 2000: 221).

Alderson concludes that if ordering tasks are marked in terms of partial credit, then the marking process becomes unrealistically complex and error-prone. Since ordering tasks are difficult to construct and the scoring process is problematic, they are rarely used.

Aim of the Current Study

As discussed, there exist many different techniques to test reading skill. By considering both the advantages and the disadvantages of each, the tester can decide which one to use, bearing in mind the preparation, administration, and evaluation (scoring) of the test.

All testing techniques except ordering tasks provide a sensible scoring scheme to testers. As Alderson (2000) mentions, testers face problems when marking ordering tasks since testing professionals think it unfair to evaluate this type of question according to the traditional method of marking it completely right or completely wrong. The aim of this article is to find a sensible solution for scoring ordering tasks that is based on Alderson's (2000: 221) question:

> "If a student gets four elements out of eight in correct sequence, how is such a response to be weighted? And how is it to be weighted if he gets three out of eight in the correct order?"

So, how can we reward students for right answers and not penalize them just for a single mistake?

In the following sections, a new approach will be discussed, namely *Weighted Marking Protocol*, to solve the partial marking problem in ordering tasks. Answers of test-takers will be scored by using both the traditional approach and *Weighted Marking Protocol* which will enable us to compare the scores.

Evaluating Ordering Tasks

Let us examine the following ordering task. The text and the eight statements which are required to be put into the correct order are followed by the answer key and the answers of eight test-takers. First the marks of these eight test-takers will be given according to the traditional approach and then the marks will be given according to the new approach developed by the author, namely *Weighted Marking Protocol*.

It was almost midnight. John was still awake because he did not have to get up early in the morning. His favourite actor's movie on TV had just finished. The bell rang. He opened the door. It was his flat-mate, Tom. He had forgotten his keys at home in the morning. He seemed too tired to chat with John so he went to bed as soon as possible. John felt lonely and decided to go to bed. He went to the bathroom and brushed his teeth. When he came into his bedroom, he noticed some candies on the table. He ate a few of them. The candies reminded him of his childhood. Since he did not want to sleep, he decided to look at some old photos. He felt sad when he saw his ex-girlfriend Laura in a photo. He remembered the days they had spent together. He checked his watch and went to bed.

(Source: Original)

Figure 1: An Example of an Ordering Task

Put the scrambled sentences into the correct order that they happened.

(20 points)

(…..) A. John ate some candies.
(…..) B. John felt sad.
(…..) C. Tom went to bed and John felt lonely.
(…..) D. John watched a film on TV.
(…..) E. John remembered his childhood.
(…..) F. The bell rang and Tom came home.
(…..) G. John looked at the photos.
(…..) H. John brushed his teeth.

In Figure 1, an example of an ordering task is seen in which the test-takers are expected to put the eight statements in the correct order that they happen. Table 1 presents both the answer key of the ordering task and the answers of eight fictitious test-takers.

Table 1: Answer Key and Answers of the Test-takers

Answer Key	Test-takers							
	1	2	3	4	5	6	7	8
5	5	8	3	3	3	4	7	4
8	8	3	8	8	8	3	4	7
3	3	1	1	6	6	8	2	2
1	1	6	6	1	1	6	6	6
6	6	2	2	2	7	1	1	1
2	2	7	7	7	2	7	3	3
7	7	4	4	4	4	2	8	8
4	4	5	5	5	5	5	5	5

Table 1 shows answer key and answers of eight test-takers.

Scoring According to the Traditional Approach

The scoring process of ordering tasks either wholly right or totally wrong is called as *traditional approach* in this article. When the test-takers' answers in Table 1 are compared

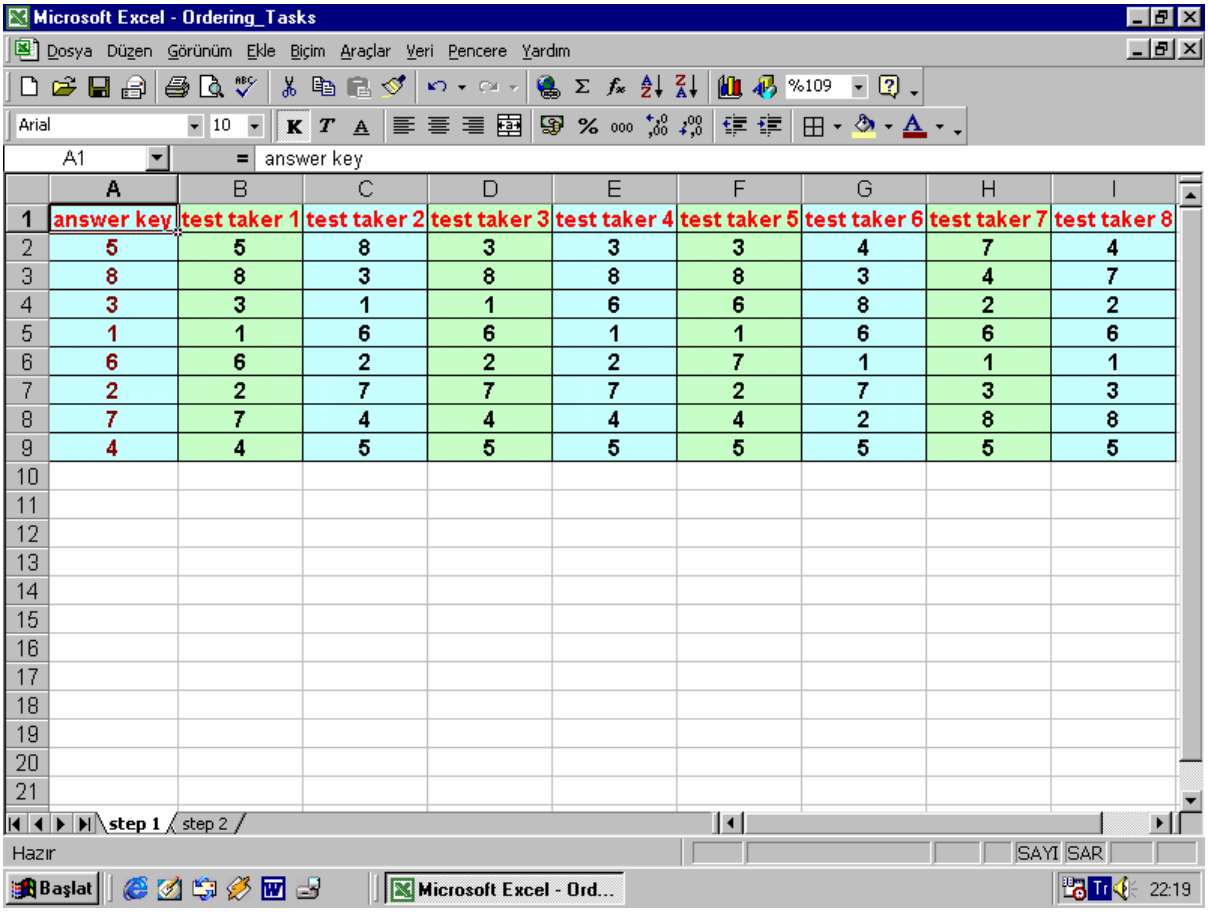
with the answer key, it is seen that only Test-taker 1 orders the sentences totally right which means that she deserves 20 points. The other seven test-takers have at least one mis-ordering or more. Since the traditional approach pressures the tester to evaluate as wholly right or totally wrong, these seven test-takers get '0' points even though some of them have partial right orderings.

Scoring According to the New Approach: *Weighted Marking Protocol*

According to the traditional approach, all, except Test-taker 1 get '0' points. However, Test-taker 2 has only one mis-ordering. If she had placed sentence 'D' as the first one, instead of last, she would have got full points. This indicates that Test-taker 2 generally comprehended the text by making only one mistake. On the other hand, Test-taker 8 seems to have trouble in comprehending the text since her order does not make any sense. It is not fair to give '0' points to both Test-taker 2 and 8 since their comprehension of the text differs.

The new approach, namely *Weighted Marking Protocol*, aims to mark ordering tasks in a fairer way that makes more sense. Here, the tester feeds the answers into a computer preferably through Microsoft Excel. It is also possible to feed the answers into a computer through Microsoft Word or SPSS (Statistical Package for Social Sciences). The following figure provides an example of the first step in this process using Microsoft® Excel 2000.

Figure 2: Input of Answers into Computer

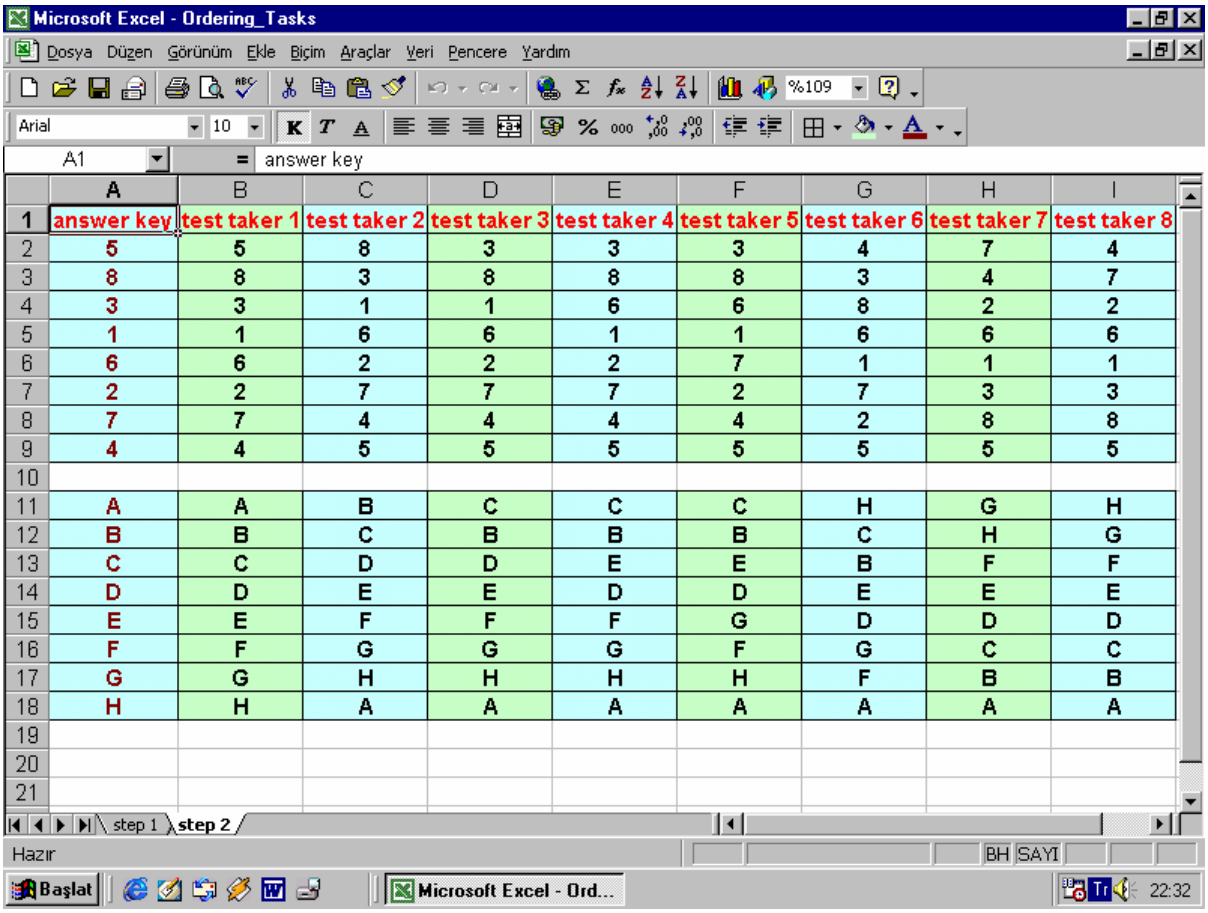


	A	B	C	D	E	F	G	H	I
1	answer key	test taker 1	test taker 2	test taker 3	test taker 4	test taker 5	test taker 6	test taker 7	test taker 8
2	5	5	8	3	3	3	4	7	4
3	8	8	3	8	8	8	3	4	7
4	3	3	1	1	6	6	8	2	2
5	1	1	6	6	1	1	6	6	6
6	6	6	2	2	2	7	1	1	1
7	2	2	7	7	7	2	7	3	3
8	7	7	4	4	4	4	2	8	8
9	4	4	5	5	5	5	5	5	5

Obviously, it is not easy to see how many mistakes the test-takers made at this step. To make it easier, the tester needs to change the answers into letters in alphabetical order

based on the answer key. For the above example, ‘5’ needs to be changed to ‘A’, ‘8’ = ‘B’, ‘3’ = ‘C’, ‘1’ = ‘D’, ‘6’ = ‘E’, ‘2’ = ‘F’, ‘7’ = ‘G’ and finally ‘4’ = ‘H’; so that the tester can see the answers in alphabetical order to help her correct the mistakes in the test-takers’ answers. The transformation can be done easily by pressing Ctrl+H at the same time and writing the values.

Figure 3: Test-takers’ Answers in Alphabetical Order



	A	B	C	D	E	F	G	H	I
1	answer key	test taker 1	test taker 2	test taker 3	test taker 4	test taker 5	test taker 6	test taker 7	test taker 8
2	5	5	8	3	3	3	4	7	4
3	8	8	3	8	8	8	3	4	7
4	3	3	1	1	6	6	8	2	2
5	1	1	6	6	1	1	6	6	6
6	6	6	2	2	2	7	1	1	1
7	2	2	7	7	7	2	7	3	3
8	7	7	4	4	4	4	2	8	8
9	4	4	5	5	5	5	5	5	5
10									
11	A	A	B	C	C	C	H	G	H
12	B	B	C	B	B	B	C	H	G
13	C	C	D	D	E	E	B	F	F
14	D	D	E	E	D	D	E	E	E
15	E	E	F	F	F	G	D	D	D
16	F	F	G	G	G	F	G	C	C
17	G	G	H	H	H	H	F	B	B
18	H	H	A	A	A	A	A	A	A
19									
20									
21									

Once the answers are in alphabetical order, it is easy to see where they made mistakes. The next duty is to find out the students who made mistakes in ordering. In the above example, Test-takers 2, 3, 4, 5, 6, 7, and 8 all made at least one mistake. Now the tester should correct their answers in the shortest possible way. For example, when ‘A’ is moved above ‘B’ in ‘Test-taker 2’s answer, it will be in the correct order, so Test-taker 2 requires only ‘1’ correction step. On the other hand, Test-taker 3 requires ‘2’, Test-taker 4 requires ‘3’, Test-taker 5 requires ‘4’, Test-taker 6 requires ‘5’, Test-taker 7 requires ‘6’ and Test-taker 8 requires ‘7’ corrections in order to achieve the correct order. At the end of this process, the tester gets the number of the correction steps for each test-taker. The formula is:

Score 1= number of statements ‘minus’ number of corrections

There are at most 7 corrections in an eight-statement task so test-takers who have 7 corrections should get ‘0’. Test-takers who need 5 and 6 corrections also deserve ‘0’ points since such orders do not make sense. Their score is called *‘probable minimum score’*. The next formula is:

Score 2 = Score 1 'minus' number of probable minimum score

In an eight-statement ordering task, maximum 'Score 2' is '5'. Now, the tester needs to arrange the rank of the test. For example, Figure 3 is a part of a test that is equal to '20' points. Since the total amount of the ordering task in the test is 20 points, this number is divided by maximum Score 2.

Score 3 = total points of ordering task 'divided by' Score 2

So Score 3 for the above-mentioned test is *20 'divided by' 5 = 4*
The final formula giving the scores for this ordering task is:

Score = Score 2 X Score 3

Table 2 shows the rank of ordering tasks in a '20' point section.

Table 2: Rank of Ordering Tasks

Number of corrections	Ordering Task Score
0	20
1	16
2	12
3	8
4	4
5	0
6	0
7	0

According to Table 2, Test-taker 1 gets 20 points, but instead of all the other getting zero points, Test-taker 2 gets 16, Test-taker 3 gets 12, Test-taker 4 gets 8, Test-taker 5 gets 4, and only Test-takers 6, 7, and 8 get none. These eight test-takers are examples of all the possible categories in an eight-statement ordering task. Table 3 compares the scores using the different approaches.

Table 3: Comparison of the Scores

Test-takers	Scores *Traditional Approach*	Scores *Weighted Marking Protocol*
Test-taker 1	20	20
Test-taker 2	0	16
Test-taker 3	0	12
Test-taker 4	0	8
Test-taker 5	0	4
Test-taker 6	0	0
Test-taker 7	0	0
Test-taker 8	0	0

It is clearly seen in tables 2 and 3 that testers can partially evaluate the ordering task. It is proposed therefore that Reading Teachers reward their students for right answers in ordering tasks instead of penalizing them with zero marks just for a single mistake.

Conclusion

This new approach towards ordering tasks enables testers to make a partial evaluation. According to the traditional approach, test-takers get full-points or zero in such a section. This means that test-takers who answer half the ordering task in the correct order are equated with those who have no mistakes, or those who have no sensible order. *Weighted Marking Protocol* does not require advanced computer knowledge; every teacher can do it by following the steps indicated. It does require a little more time than the traditional approach but is not so time-consuming. The major benefit of this new approach is that it enables teachers to reward students according to their right answers in ordering tasks, and thus give credit where credit is due.

Special Acknowledgements
I would like to express my gratitude to Asst. Prof. Dr. Ismail Hakki Erten, Asst. Prof. Dr. Ece Zehir Topkaya; and to anonymous reviewers of the Reading Matrix whose encouraging criticism of the earlier version of the article made it possible for me to revise it. My special thanks go to Graham Lee for proofreading the article.

References

Alderson, J. C. (1979). “The cloze procedure and proficiency in English as a foreign language.” *TESOL Quarterly*, 13, 2, 219-227.

Alderson, J. C. (1996). “The testing of reading.” In C. Nuttall (Ed.) *Teaching reading skills in a foreign language,* 212-228. Oxford: Heinemann.

Alderson, J. C. (2000). *Assessing Reading.* Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A. and Oller, Jr. J. W. (1985). “When are cloze items sensitive to constraints cross sentences?” *Language Learning,* 35, 2, 181-206.

Cohen, A. D. (1998). “Strategies and processes in test taking and SLA.” In L. F. Bachman and A. D. Cohen (Eds.) *Interfaces between second language acquisition and language testing research,* 90-111. Cambridge: Cambridge University Press.

Cohen, A. D. (2001). “Second language assessment.” In M. Celce-Murcia (Ed.) *Teaching English as a Second or Foreign Language,* 515-534. Boston: Heinle & Heinle.

Harmer, J. (2001). *The Practice of English Language Teaching.* Essex: Pearson Education Limited.

Hudges, A. (2003). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

McNamara, T. (2000). *Language Testing.* Oxford: Oxford University Press.

Meyer, B. (1975). *The Organisation of Prose and its Effects on Memory.* New York, NY: North Holland.

Pearson, P. D. and Johnson, D. D. (1978). *Teaching Reading Comprehension.* New York, NJ: Holt, Rinehart and Winston.

Sezer, A. (2002). *Prepare for the TOEFL.* Ankara: Hacettepe-Tas Kitapcılık Ltd. Sti.

Ur, P. (1996). *A Course in Language Teaching.* Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative Language Testing.* New York: Prentice Hall.

Weir, C. J. (1993). *Understanding & Developing Language Tests.* New York: Prentice Hall.

Salim RAZI is an Instructor of English in the ELT Department at Canakkale Onsekiz Mart University, Canakkale, Turkey. He holds an MA degree in ELT. He previously taught English for the Ministry of Education and worked as a Research Assistant during his MA studies at the ELT department in Canakkale. He has also presented and published papers at international conferences. He is interested in the effects of schema on reading comprehension, the reading process, reading activities and assessing reading. He can be

contacted at: Canakkale Onsekiz Mart University, Anafartalar Campus, C1 206, 17100 Canakkale, Turkey; or by e-mail: salimrazi@comu.edu.tr

Appendices

Appendix A

Below is an example of a cloze test constructed by deleting every sixth word in the abstract of this article. There occur no deletions in the first sentence.

After briefly discussing techniques such as ‘the cloze test’ and ‘gap-filling’, the main focus of the paper resides in the scoring process of ‘ordering tasks’, where students are asked to re-arrange the order of sentences given in incorrect order. 1) _____ the evaluation of such tasks 2) _____ quite complex, Reading Teachers rarely 3) _____ them. According to Alderson, Reading 4) _____ frequently tend to mark these 5) _____ either wholly right or totally 6) _____ since the partial marking process 7) _____ very time-consuming readers will be 8) _____ to a new approach, namely 9) _____ *Marking Protocol*, developed by the 10) _____ for ordering tasks in order 11) _____ achieve a fairer evaluation. This 12) _____ approach makes it possible for 13) _____ Teachers to reward their students 14) _____ all right answers and not 15) _____ them just for a single 16) _____.

Appendix B

Below is an example of a gap filling test which is constructed by deleting ten content words in the abstract of this article.

After briefly discussing techniques such as ‘the cloze test’ and ‘gap-filling’, the main 1) _____ of the paper resides in the scoring process of ‘ordering tasks’, where students are asked to re-arrange the order of sentences given in 2) _____ order. Since the evaluation of such tasks is quite 3) _____, Reading Teachers rarely use them. According to Alderson, Reading Teachers frequently tend to mark these 4) _____ either wholly right or totally wrong since the partial marking process is very time-consuming 5) _____ will be introduced to a new 6) _____, namely *Weighted Marking Protocol,* 7) _____ by the author for ordering tasks in order to achieve a 8) _____ evaluation. This new approach makes it 9) _____ for Reading Teachers to reward their students for all right answers and not 10) _____ them just for a single mistake.

Appendix C

Below is an example of a C-test which is constructed by deleting the second half of every second word in the first sentence of the abstract of this article.

After bri_____ discussing techn_____ such a_____ ‘the cl_____ test’ a_____ ‘gap-filling’, t_____ main fo_____ of t_____ paper res_____ in t_____ scoring pro_____ of ‘orde_____ tasks’, wh_____ students a_____ asked t_____ re-arrange t_____ order o_____ sentences gi_____ in inco_____ order.

Appendix D

Below is an example of a close elide test which is constructed by inserting three words into the first two sentences of the abstract of this article.

After briefly discussing techniques such as ‘the cloze test’ and ‘gap-filling’, the main focus of the paper resides in about the scoring process of ‘ordering tasks’, where students are asked to re-arrange and the order of sentences given in incorrect order. Since the evaluation of such tasks is quite complex, because Reading Teachers rarely use them.