VNU Journal of Science: Foreign Studies, Vol. 32, No. 2 (2016) 75-84

Rater Consistency in Rating L2 Learners' Writing Task

Nguyen Thi Quynh Yen*

Center for Language Testing and Assessment, VNU University of Languages and International Studies
Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 18 March 2015
Revised 27 November 2015; Accepted 18 May 2016

Abstract: Rater consistency plays a critical part in the rating procedure. Test scores will be unreliable if examiners are inconsistent in their rating and fail to agree with other raters on the relative merits of rating scale, severity and leniency and so on. Despite the difficulty in matching the standard, writing paper is widely used in various kinds of language tests because it can provide not only a high motivation for writing, but also an excellent backwash effect on teaching. For this reason, it is necessary to establish high consistency in the scores given by one rater (intra-rater reliability) and by different raters (inter-rater reliability). This article discusses rater consistency in essay evaluation conducted by some randomly chosen raters in the Faculty of English, the University of Languages and International Studies, VNU and from that, some suggestions are made to improve the reliability in rating L2 learners' essay writing.

Keywords: Rater consistency, intra-rater reliability, inter-rater reliability, holistic scoring, analytical scoring.

1. Introduction

Within the past few decades, writing assessment has been a constant concern. There has been much research on the validity and reliability of scores given to written products. According to McNamara [1], rating always contains a significant degree of chance, associated with the rater and other factors. Any malfunctioning in the writing assessment might raise a basic but critical question about the rating procedure. Weir [2] identifies a number of factors that can threaten the reliability or scoring validity of writing tests. Among these factors, rater reliability has been a matter of longstanding concern for many large-scale testing agencies. In test scores that are subjectively obtained such as ratings of essays, it is necessary to minimize the inconsistency in the scores given by one rater (intra-rater reliability) and by different raters (inter-rater reliability). The test scores will be unreliable if raters are inconsistent in their own rating and fail to agree with other raters on the relative merits of rating scale, severity and leniency and so on. That is the reasons why whether or not such subjective testing items as essays should be utilized in the high-stake tests has been

* Tel.: 84-978931058
Email: yenntq@vnu.edu.vn

always in dispute. In spite of all possible risks of unreliability, essays are still widely used in a variety of language tests merely because they can help to measure critical thinking skills, understanding of course materials and writing skills. Therefore, more studies need conducting on the rater consistency and ways to improve it.

2. Rater consistency

Rater consistency refers to the extent to which the scores given by raters are stable, consistent and free from errors. Rater consistency can be viewed as rater reliability or rater agreement.

2.1. Intra-rater reliability

Intra-rater reliability refers to the degree of agreement among multiple repetitions performed by one rater. According to Bachman [3], when an examiner rates a given sample of language performance, whether it is written or spoken, that rating will be based on a set of criteria. If the rater applies the same set of criteria consistently in rating the language performance of different test takers, this will bring about a reliable set of ratings.

Bachman [3] describes three cases in which intra-rater consistency is affected. Firstly, the score given to a written paper may be affected by the rating criteria themselves, the consequence of rating or the contrast between the previous papers and the following ones. Secondly, the intra-rater consistency in applying rating criteria can also be affected by the sequence of scoring. Thirdly, the intra-rater consistency can be affected by the contrast between the quality of the previous and the following essays.

2.2. Inter-rater reliability

Inter-rater reliability refers to the degree of similarity between different raters in scoring the same set of writing without influencing one another. McNamara [1] says that even trained raters differ in their handling of the allocation of individual performances in borderline cases. According to Bachman [3], ratings given by different raters can also vary as a function of inconsistencies in the criteria used to rate and in the way in which these criteria are applied.

For example, if a group of five essays is given to five different raters, we would be likely to obtain very different results from the different raters. And even if these raters are asked to rate on the same component, say, organization, there are likely to be differences in the way the raters interpret this.

Because of the problems of subjectivity mentioned above with raters in assessment, rater-mediated assessment used to be discouraged in the 1950s and 1960s. To avoid direct testing, writings skills used to be assessed indirectly through examination of control over the grammatical system and knowledge of vocabulary such as in Toelf test. However, this restriction on the scope of assessment led to more losses than gains. Thus, the problem of subjectivity was something that had to be faced and managed in the removal or at least, reduction of rater inconsistency, which can also be affected by the scoring methods.

3. Scoring methods

Essays may be scored according to two different criteria: the holistic scoring and the analytic scoring.

3.1. Holistic Scoring

Holistic scoring is a method by which trained raters evaluate a piece of writing for its overall quality. Holistic scoring is often used in large-scale assessments, such as college placement tests. According to Babin and Harrison [4], an advantage to holistic scoring is that raters can evaluate many papers in a short span of time because they do not comment on or correct the student's work. Experienced scorers can judge a one-page of writing in just several minutes or even less.

However, critics of the method have questioned its validity and reliability. Different raters may choose to focus on different aspects of the written product and they may be swayed by superficial factors such as length and appearance of an essay. And as it is possible for each writing product to appear just to a certain rater but not others, the examiner's mark may be a highly subjective one.

Because the inherent unreliability in holistic marking of essays, it is essential to use another method of scoring: Analytic scoring.

3.2. Analytic scoring

Analytic scoring is a method that requires a separate score for each of a number of aspects of a task, such as task response, coherence and cohesion, lexical resource and grammatical range and accuracy. This method has several disadvantages. It takes more time than holistic scoring and according to Hughes [5], concentration on the different aspects may divert attention from the overall effect of the piece of writing. However, analytic scoring disposes of the problem of uneven development of subskills in individuals. The raters are compelled to consider aspects of performance which they might otherwise ignore and the fact that the rater has to give a number of scores will tend to make the scoring more reliable.

In universities and colleges, essay writing is a compulsory task in nearly every important language test (achievement and proficiency tests). In the faculty of English, University of Languages and International Studies, Vietnam National University, essay writing is required in the final examination of English semester 4 and 5. In the writing paper, students are required to complete two writing tasks, which are letter writing and essay writing. For evaluating the essays, the raters apply the public version of IELTS nine-band descriptors with four subcriteria (Task response, Coherence and cohesion, Lexical resource and Grammatical range and accuracy). These subcriteria are equally weighed. The details of the analytic scales for rating essays are presented in the appendix.

4. Data analysis

4.1. Methodology

This study applied both quantitative and qualitative methods. It selected and analysed the data about the scores the raters gave to different essays to investigate rater consistency. The consistency of each rater was measured by the deviation from the mean score. The closer to the mean scores the scores the raters gave to the essays were, the more reliable the raters were. The sums of deviation were also calculated to investigate the inter-rater and intra-rater reliability. The information about rater consistency will help qualify raters for important tests and decide which raters need retraining.

Besides, in order to investigate the factors that may affect the raters while rating essays such as the range of approaches used by the raters and the elements the raters focused on

while rating those essays, this study collected the comments the raters noted down while rating and employed the introspective verbal reports ([6]-[9]). All the raters involved in the study were asked three questions after they had finished their ratings: (1) Why do you give essay 1/2/3/4/5 score 6.5 (for example)? (2) Are there any factors that made you confused when scoring the essays? (3) What makes the scores you gave to different essays differ? The answers were then analysed to investigate the factors affecting the raters during the rating process so that these factors could be taken into consideration to ensure the scoring validity of any actual tests.

4.2. Procedure

A group of 10 raters were selected from the Faculty of English, ULIS-VNU. They are the ones aged 28-35, who are involved in scoring students' essays and have attended at least one rater training workshop. These raters were given 5 answer papers with both task 1 and task 2 for scoring independently. They were also provided with the public version of IELTS writing band descriptors and asked to evaluate each candidate's writing task according to 9 given bands. The writing papers were coded from 01 to 05. However, due to the limited scope of the study, this study focuses only on the analysis of the scores, comments and verbal reports given to the essays in task 2 by different raters.

4.3. Data analysis

A set of 5 writing papers were given to 10 raters and their rating scores, comments and verbal reports were collected after their marking had been finished. The study uses only the scores the raters gave to the essays for analyzing rater consistency (Table 1 and Table 2).

4.3.1. Deviation of mean score

Table 1 shows the scores given to each essay by 10 different raters and the mean scores for each essay.

Table 1. Scores given to essays

Essay	R 1	R 2	R 3	R 4	R 5	R 6	R 7	R 8	R 9	R 10	Mean score
01	7.5	7.5	8.0	6.5	6.5	7.0	7.0	7.0	7.5	7.0	7.15
02	5.5	6.0	6.5	6.0	5.5	5.5	6.0	6.0	6.0	5.5	5.85
03	7.5	6.5	7.0	6.5	6.5	6.5	6.0	6.0	6.5	6.0	6.50
04	6.5	6.0	6.5	6.5	6.0	5.5	6.0	6.5	6.5	6.5	6.25
05	5.5	5.5	6.0	5.0	5.0	5.5	5.0	5.0	5.0	5.5	5.30

Table 2 shows the deviation of mean for each rater.

Table 2. Deviation

Standard deviation	R 1	R 2	R 3	R 4	R 5	R 6	R 7	R 8	R 9	R 10	Sum
Essay 01	0.35	0.35	0.85	0.65	0.65	0.15	0.15	0.15	0.35	0.15	3.80
Essay 02	0.35	0.15	0.65	0.15	0.35	0.35	0.15	0.15	0.15	0.35	2.80
Essay 03	1.00	0.00	0.50	0.00	0.00	0.00	0.50	0.50	0.00	0.50	3.00
Essay 04	0.25	0.25	0.25	0.25	0.25	0.75	0.25	0.25	0.25	0.25	3.00
Essay 05	0.20	0.20	0.70	0.30	0.30	0.20	0.30	0.30	0.30	0.20	3.00
Sum of deviation	2.15	0.95	2.95	1.35	1.55	1.45	1.35	1.35	1.05	1.45	15.60

As can be seen from the two tables, the deviation ranges from 0.00 to 1.00 which is acceptable in subjective evaluation such as writing skill. In general, the inter-rater reliability for essay 02 is rather high and that for essay 03-05 is acceptable. The greatest difference between the scores given to these essays by 10 raters is 1.0. However, the inter-rater consistency for essay 01 is rather low with the greatest difference of 1.5.

In terms of intra-rater reliability, rater 2 and 9 are the most reliable in their ratings. Rater 4, 5, 6, 7, 8 and 10 also demonstrate their consistency in their scores. In contrast, rater 1 and 3 seem less reliable in their evaluation.

The statistics in table 1 and 2 give information about the inter-rater consistency and intra-rater reliability, which are very important in training and choosing reliable raters for evaluating essays.

4.3.2 Raters' comments

Analysis of the comments the raters wrote down while scoring and interpretation of verbal reports conducted after the raters had completed their task show that factors that affect to some extent the raters' evaluation are the raters' reading styles, raters' scoring method and contrast between the previous essay and the later essay.

4.3.2.1 Reading styles and scoring methods

The reading styles of raters determine what occupies their attention while they read the essays and how the final score is assigned to each essay. This may lead to the inconsistency among raters. In spite of scoring according to the band descriptors, only 4 out of 10 raters paid equal attention to all the criteria while six of them seemed to focus more on one or two criteria. Among the criteria, grammatical range and accuracy was the criterion affecting the raters most, followed by organization and content of the essay.

When asked why band 5.0 was given to essay 5, rater 8 said that this essay used only a limited range of structures, made frequent grammatical mistakes and that the candidate used a limited range of vocabulary and made noticeable errors in spelling. This means the candidate accidentally ignored the criteria of task response and coherence and cohesion of the essay. Similarly, rater 3 was asked why band 8.0 was given to essay 1. This rater said that the grammatical structure used in the essay was very impressive and this essay hardly contained any errors. The idea development of the essay was very good. However, this rater did not pay much attention to the lexical range used in the essay which deserved the band between 6.0 and 7.0.

One interesting thing is that nearly all of the raters felt confused about the penalty for lack of words, bad handwriting and inappropriate layout of the essays, which is not stated in the band descriptors. Most of them confessed that these kinds of deficiency often lead to a bad impression on the essay.

4.3.2.2. Contrast

The data analysis has proved that the order in which essays were read could have some certain effects on the raters' judgment. One of the rater when interviewed said that in comparison with essay 1, essay 3 could not get the same score since the test taker used less complex structures. This helps strengthen the findings of the research carried out by Daly and Dickson-Markman [10] and Hughes et al [11] that the raters' evaluations of an essay may differ depending on how each rater perceived its quality relative to the preceding ones.

5. Suggestions

There has been much research into the source of rater inconsistency and ways to establish high rater consistency in writing assessment. Researchers have pointed out that lack of appropriate rating scales ([3], [12], [13], [14], [15], [16]), lack of training and inappropriate procedures ([17]) can lead to the unreliability of assessment.

As can be seen from the data analysis, to some extent, inconsistency still exists among raters in evaluating written products. Despite the difficulty in matching the standard, all the effort in making a good test will be in vain if the test takers find the scores raters give to their test papers unreliable. For this reason, it is necessary to establish high consistency in written product evaluation among the raters so that test-takers can place their total confidence in the scores they receive.

5.1. Employing appropriate rating scales

According to Alderson et al [12] and McNamara [13], the choice of appropriate rating criteria and the consistent application of rating scales by trained examiners are regarded as key factors in the valid assessment of second language performance. For this reason, firstly, rating scales needs to be appropriately defined and represented with band extremities which determine features that constitute the end of one band and beginning of the next. Secondly, any awkward descriptors should be rewritten so that they do not cause difficulties in raters' interpretation. Thirdly, the expressions in the descriptors should be clear and straightforward. Any evaluative expressions such as "unsatisfactory", "adequate" or "good" should be avoided. Fourthly, the penalty for lack of words, bad handwriting and inappropriate layout of the essays should also be stated in the rating scale or at least discussed for consensus before the rating process.

5.2. Training raters

Alderson et al [12] argue that it is widely accepted in second language writing assessment circles that the training of raters is crucial to validity in testing language performance and emphasize the vital role training has to play in the removal (or at least the reduction) of rater inconsistency.

In order to maximize the rater reliability in evaluating writing papers, raters need to be trained before official marking. During the training, raters get familiarized with the rating scales and the procedure of a real rating. Benchmark scripts must also be explained for each score band before trial rating. The individuals whose scorings deviate markedly and inconsistently from the norm should not be used and they should be retrained before being used to rate writing papers.

Besides, a good rater does not only need to meet the demand of reliability but he/she is supposed to satisfy the requirement of rating speed. For this reason, the record of each rater for every marking session should be monitored and analyzed so that rater's quality is always in control.

5.3. Sample marking

After the test has been administered, the chief examiner or team leader should select from 5 to 10 writing papers randomly for sample marking. This process aims at setting the specific standards before the real scoring. All of raters should be given copies of the scripts selected by the chief examiner or team leader, in random order, and each member

should mark all of these scripts before setting the standards.

During the sample marking, scripts which represent "adequate" and "inadequate" performances should be extracted and problems which examiners are often faced with but which are rarely described in rating scales should also be discussed, such as bad handwriting, excessively short or long responses, responses which indicate that the candidates misunderstood the task etc.

5.4. Double marking

In order to ensure the reliability in evaluation, every writing paper should be marked by at least two different raters. Each rater will work independently. The score that the candidate receives for a piece of writing is the mean of the scores given by the two raters. However, if the difference between the scores given by two raters is too big, the third rater should be invited. In this case, the third rater should be the team leader, who will decide what score should be given to that piece of writing.

Finally, we conclude that in order to enhance the validity and reliability of the scores given to written products in any examinations (progress tests, achievement tests or proficiency tests), there is a very real need for more studies focusing on raters. In cases where nearly every teacher is involved in the rating, rater training should be periodically conducted.

References

[1] McNamara, T. (2000). Language Testing. Oxford: Oxford University Press.

[2] Weir, C J. (2005). Language Testing and Validation. Palgrave Macmillan.

[3] Bachman, L. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

[4] Babin, E and Harrison, K. (1999). Contemporary Composition Studies: A Guide to Theorists and Terms. Greenwood Press.

[5] Hughes, A. (1989). Testing for Language teachers. Cambridge: Cambridge University Press.

[6] Milanovic, M, Saville, N and Shuhong, S. (1996). A Study of the Decision – making Behaviour of Composition Markers, in Milanovic, M and Saville, N (Eds) Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium and Arnhem, Studies in Language Testing 3. Cambridge: UCLES/Cambridge University Press.

[7] Orr, M. (2002). The FCE Speaking Test: Using Rater Reports to Help Interpret Test Scores. System, 30.

[8] O'Donnell, D, Thomas, G and Park, S. (2006). Revisiting Assessment Criteria in a Speaking test. Paper Presented at JALT2006 Annual Conference, Kitakyushu, Japan.

[9] Brown, A, Iwashita, N and McNamara, T. (2005). An Examination of Rater Orientations and Test-taker Performance on English-for-Academic-Purposes Speaking Tasks. Princeton, NJ: ETS.

[10] Daly, J. A and Dickson- Markman, F. (1982). Contrast effects in evaluating essays. Journal of Educational Measurement 19: 309-316

[11] Hughes, D.C., Keeling, B. and Tuck, B.F. (1980). The influence of context position and scoring method on essay scoring. Journal of Educational Measurement 17: 131 -135.

[12] Alderson, J C, Clapham, C and Wall, D. (1995). Language test construction and evaluation. Cambridge: Cambridge University Press.

[13] McNamara, T. (1996). Measuring Second Language Performance. London: Longman.

[14] Hamp-Lyons, L. (1991). Scoring Procedures for ESL Contexts. In Hamp-Lyons: 241-278.

[15] Hout, B. (1996). Toward a new theory of writing assessment. College Composition and Communication 47.

[16] Weir, C J. (1990). Communicative Language Testing. Englewood Cliffs, NJ: Prentice Hall.

[17] Taylor, L and Falvey, P (Eds). (2007). IELTS Collected Papers: Research in speaking and writing assessment. Studies in Language Testing 19. Cambridge: Cambridge University Press.

Đảm bảo độ tin cậy trong việc chấm bài luận

Nguyễn Thị Quỳnh Yến

Trung tâm Khảo thí, Trường Đại học Ngoại ngữ, ĐHQGHN,
Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Tóm tắt: Độ tin cậy đóng một vai trò rất quan trọng trong quá trình chấm các bài kiểm tra. Điểm thi sẽ không tin cậy nếu cán bộ chấm không nhất quán trong quá trình đánh giá của chính mình và không thống nhất với những cán bộ chấm thi khác xét về các tiêu chí chấm, độ nghiêm khắc và nhẹ tay trong quá trình chấm, v.v.. Mặc dù thực tế vẫn tồn tại những khó khăn trong việc đảm bảo việc đánh giá chính xác các bài thi mang tính chủ quan như môn viết, các bài kiểm tra kỹ năng viết vẫn được sử dụng rộng rãi trong các kỳ thi ngôn ngữ khác nhau vì nó giúp tạo động cơ cho người học và có tác động ngược lại với quá trình giảng dạy. Chính vì lý do này, việc nâng cao độ tin cậy trong việc đánh giá các bài kiểm tra viết rất quan trọng. Bài viết này bước đầu khảo sát độ tin cậy trong việc đánh giá các bài luận tại khoa Tiếng Anh, Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội và từ đấy đề xuất một số giải pháp giúp nâng cao độ tin cậy trong việc chấm thi.

Từ khóa: Độ tin cậy trong việc đánh giá của một cán bộ chấm thi, độ tin cậy trong việc đánh giá giữa các giám khảo chấm thi, chấm thi tổng quát, chấm thi theo các tiêu chí.

Appendix

IELTS public band descriptors for writing task 2

Fig.1. Analytic scales for Task Response

Band	Task Response
9	Fully addresses all parts of the task; presents a fully developed position in answer to the question with relevant, fully extended and well supported ideas.
8	Sufficiently addresses all parts of the task; presents a well-developed response to the question with relevant, extended and supported ideas.
7	Addresses all parts of the task; presents a clear position throughout the response; presents, extends and supports main ideas, but there may be a tendency to overgeneralise and/or supporting ideas may lack focus.
6	Addresses all parts of the task although some parts may be more fully covered than others; presents a relevant position although the conclusions may become unclear or repetitive; present relevant main ideas but some may be inadequately developed/unclear.
5	Addresses the task only partially; the format may be inappropriate in places; expresses a position but the development is not always clear and there may be no conclusions drawn; presents some main ideas but these are limited and not sufficiently developed; there may be irrelevant detail.
4	Responds to the task only in a minimal way or the answer is tangential; the format may be inappropriate; presents a position but this is unclear; presents some main ideas but these are difficult to identify and may be repetitive; irrelevant or not well-supported.
3	Does not adequately address any part of the task; does not express a clear position; present few ideas, which are largely undeveloped or irrelevant.
2	Barely responds to the task; does not express a position; may attempt to present one or two ideas but there is no development.
1	Answer is completely unrelated to the task.
0	Does not attend; does not attempt the task in any way; write a totally memorized response.

Fig.2. Analytic scales for Coherence and Cohesion

Band	Coherence and cohesion
9	Uses cohesion in such a way that it attracts no attention; skillfully manages paragraphing.
8	Sequences information and ideas logically; manages all aspects of cohesion well; uses paragraphing sufficiently and appropriately.
7	Logically organises information and ideas; there is clear progression throughout; uses a range of cohesive devices appropriately although there may be some over-/under-use; present a clear central topic with each paragraph.
6	Arranges information and ideas coherently and there is a clear overall progression; uses cohesive devices effectively, but cohesion within and/or between sentences may be faulty or mechanical; uses paragraphing, but not always logical.
5	Presents information with some organization but there may be a lack of overall progression; makes inadequate, inaccurate or over-use of cohesive devices; may be repetitive because of lack of referencing and substitution.
4	Presents information and ideas but these are not arranged coherently and there is no clear progression in the response; uses some basic cohesive devices but these may be inaccurate or repetitive; may not write in paragraphs or their use may be confusing.
3	Does not organize ideas logically; may use a very limited range of cohesive devices, and those used may not indicate a logical relationship between ideas.
2	Has very little control of organization features.
1	Fails to communicate any message.
0	Does not attend; does not attempt the task in any way; write a totally memorized response.

Fig.3. Analytic scales for Lexical Resource

Band	Lexical Resource
9	Uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as slips.
8	Uses a wide range of vocabulary fluently and flexibly to convey precise meanings; skillfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation; produces rare errors in spelling and/or word formation.
7	Uses a sufficient range of vocabulary to allow some flexibility and precision; uses less common lexical items with some awareness of style and collocation; may produce occasional errors in word choice, spelling and/or word formation.
6	Uses an adequate range of vocabulary for the task; attempts to use less common vocabulary but with some accuracy; makes some errors in spelling and/or word formation, but they do not impede communication.
5	Uses a limited range of vocabulary, but this is minimally adequate for the task; may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader.
4	Uses only basic vocabulary which may be used repetitively or which may be inappropriate for the task; has limited control of word formation and/or spelling; errors may cause strain for the reader.
3	Uses only very limited range of words and expressions with very limited control of word formation and/or spelling; errors many severely distort the message.
2	Uses an extremely limited range of vocabulary; essentially no control of word formation and/or spelling.
1	Can only use a few isolated words
0	Does not attend; does not attempt the task in any way; write a totally memorized response.

Fig.4. Analytic scales for Grammatical Range and Accuracy

Band	Grammatical range and accuracy
9	Uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as slips.
8	Uses a wide range of structures; the majority of sentences are error-free; make only very occasional errors or inappropriacies.
7	Use a variety of complex structures; produces frequent error-free sentences; has good control of grammar and punctuation but may make a few errors.
6	Uses a mix of simple and complex sentences forms; makes some errors in grammar and punctuation but they rarely reduce communication.
5	Uses only a limited range of structures; attempts complex sentences but these tend to be less accurate than simple sentences; may make frequent grammatical errors and punctuation may be faulty; errors can cause some difficulty for the reader.
4	Uses only a very limited range of structures with only rare use of subordinate clauses; some structures are accurate but errors predominate, and punctuation is often faulty.
3	Attempts sentences forms but errors in grammar and punctuation predominate and distort the meaning.
2	Cannot use sentence forms except in memorized phrases.
1	Cannot use sentence forms at all.
0	Does not attend; does not attempt the task in any way; write a totally memorized response.