The generalizability of summarization and free recall ratings in L2 reading assessment

Yasuyo SAWAKI

University of California, Los Angeles¹

1. Introduction

Using reading comprehension questions for assessing readers' comprehension of a text is a common practice in current educational testing. Discrete-point item formats such as multiple choice have been advocated in educational contexts in particular, primarily due to the focus on large-scale group testing of silent reading in society as well as the preference for efficiency and ease of test administration and scoring (Johnston). However, second-language (L2) reading and language assessment researchers have recently raised several concerns about the use of questions to assess reading comprehension. Some reasons include the possibility that the use of questions may change the reading comprehension activity to a "matching exercise" (Rusch and Stoddard 117, cited by Johnston), which may lead to a "fragmented, compartmentalized reading of a text" (Riley and Lee 174). Moreover, questions with extra text make interpretation of the test results difficult. Namely, it is hard to pinpoint exactly where—in the text for the question, the prompt text, or both—the reader has had a reading comprehension problem (Bernhardt, "Three Approaches"). In an attempt to provide a remedy to the limitations of the use of reading comprehension questions, several researchers in L2 reading and assessment have proposed the use of extended response task types such as free written recall (e.g., Bernhardt, "Three Approaches"). Others have proposed summarization, another extended-response task type, as a test method (e.g., Cohen, "The role of instructions"; Kobayashi). According to the advocates of extended-response task types, these types are considered as global measures of reading comprehension because the task requirements facilitate the reader's construction of an internal representation of a text as a whole (Bernhardt, "Three Approaches"; Bernhardt, Reading).

Although this search for a "purer" measure of reading comprehension sounds attractive, practical applications of extended-response tasks for reading assessment are a challenge. A primary reason is the lack of empirical evidence that explicates the nature of the constructs that extended-response task types assess. Another critical issue is the lack of a practical, yet reliable scoring method and empirical evidence of the reliability that measurement of extended-response task types can achieve in L2 assessment contexts. To shed light on this second issue, in particular, the present study takes a multivariate generalizability theory approach to compare two types of extended-response reading comprehension tasks—written summarization and free recall—in terms of the reliability of ratings awarded to protocols generated by learners

of Japanese as a foreign language (JFL).

2. Review of literature

Both summarization and free recall have been studied extensively in contexts other than L2 reading assessment. On the one hand, free written recall has a long history as a reading comprehension measure in reading research (e.g., Bernhardt, Reading; Lee; Carrell). In these studies, free written recall was conceptualized as a “purer” measure of reading comprehension compared to question-based reading comprehension tasks. This is generally consistent with the notion of recall as a natural product of text comprehension defined in Kintsch and van Dijk’s process model of reading. On the other hand, readers’ summarization behaviors and strategies have been the targets of investigation in first-language (L1) reading research for the past three decades due to the importance of summarization as a study skill and its close relationship with the use of metacognition while reading (e.g., Baker and Brown; Brown and Day; Paris, Wasik, and Turner; Rinehart, Stahl, and Erickson; Winograd). This line of research was replicated with L2 readers in a few studies (e.g., Cordero-Ponce; Johns and Mayes; Kim). Discussions in these studies centered primarily on summarizers’ use of various strategies for gist formation and the relationship of these strategies with the development of reading skills.

In language assessment, few studies have investigated the construct validity of free written recall and/or summarization so far. For example, in a study where summarization and free recall were compared based on the performance of university-level learners of French as a foreign language, Riley and Lee found that, compared to free recall, summarization led to recall of significantly more idea units and main ideas as well as qualitatively more coherent protocols. Cohen conducted an in-depth qualitative analysis of the process of completing summarization tasks by a small group of Brazilian learners of English as a foreign language (EFL) who had various language ability levels (Cohen, “English for academic purposes”). The students in Cohen’s study reported a wide variety of reading strategies, which suggested high interactiveness (Bachman and Palmer) of the task type.

Although these previous studies suggest the usefulness of summarization and free recall as measures of reading comprehension, a vexing issue is how to score summarization and free recall protocols. In previous nonassessment studies, various text analysis methods were employed for scoring summarization and recall protocols, while the use of other scoring methods more common in the assessment literature (e.g., rating scales) has been infrequent. The complexity and the labor-intensiveness of the previous scoring methods make their practical applications for testing unfeasible, however. For example, in Meyer’s idea unit analysis—a widely used method for scoring recall protocols—the time required for developing a scoring template is usually about twenty to fifty hours per 250-word protocol, while scoring itself normally takes

thirty to sixty minutes per protocol (Bernhardt 202-03).

A second issue is the paucity of evidence on the reliability of scoring assigned to summarization and free recall protocols in previous studies in general. Often available information about test reliability in these studies is based on rater agreement (Cohen, "The role of instruction"), while systematic analysis of score reliability beyond that has rarely been reported. Meanwhile, Cohen's (Cohen, "English for academic purposes") insight sheds light on a central issue in considering the reliability of summarization rating: considerable subjective judgments on the part of raters involved in scoring summarization protocols. By nature, integrating information for summarizing a text often leads to a summary writer's departure from the text. Assessing the appropriateness of these departures is a judgment call for raters because they may not be able to find a match between a summary writer's representation of original information and the scoring keys (which are often built to be parallel to a prompt passage in terms of the points to be covered and their order of presentation). Although evaluating the appropriateness of such departure from the original passage is involved in free recall scoring to some degree as well, the relative extensiveness of this judgment required of raters seems to distinguish the scoring requirement for summarization from that for free recall.

In order to explore these issues, the present study compared the generalizability of ratings awarded to summarization and free recall protocols written by university-level JFL learners. This researcher developed the analytic rating scales based on the criteria that were identified as critical for scoring summary and recall protocols in previous research (e.g., Riley and Lee; Winograd). The design of the analytic rating scales was also informed by the scoring rubrics for a language placement test developed at the University of California, the Language Ability Assessment System (Bachman, Lynch, and Mason), in which language ability was defined as multicomponential. The generalizability of the summarization and free recall ratings was investigated through a multivariate generalizability theory (G theory) analysis (Brennan, "Generalizability"; Cronbach, Gleser, Nanda, and Rajaratnam; Webb, Shavelson, and Maddahian). Multivariate G theory is a broad analytic framework that subsumes univariate G theory often used in language assessment to investigate test score reliability (e.g., Bolus, Hinofotis, and Bailey; Brown; Brown and Bailey; Bachman et al.; Lynch and McNamara). Multivariate G theory provides an adequate analytic framework when ability across multiple constructs is assessed simultaneously. This framework is particularly useful for the present study because recalling and summarization abilities are defined as different combinations of constructs associated with the test taker's L2 reading ability. Besides the information about generalizability of individual rating scales available in conventional univariate G theory analyses, a multivariate G theory analysis provides information about interrelationships among the analytic rating scales. This information is particularly valuable because it allows one to examine the

distinctiveness of the constructs tapped by individual rating scales (Brennan, “Generalizability”; Nussbaum; Webb, Shavelson, and Maddahian; Webb and Shavelson). The questions addressed in this study are:

1. Is the generalizability of ratings comparable between summarization and free recall?
2. To what extent are the constructs being measured by the summarization and free recall rating scales similar to or different from each other?

3. Method

3.1 Participants

The participants in this study were 188 volunteers among beginning and intermediate learners of Japanese enrolled in JFL courses at five large state universities in the United States during the 2002–2003 academic year. The participants were sampled from eight target Japanese language courses at these universities, including three first-year (elementary-level) Japanese language courses and five second-year (intermediate-level) courses. One of the first-year courses was designed for heritage speakers of Japanese, while all the other courses were offered as part of the regular Japanese language course sequences. At the time of their participation in this study, 66 of the learners (35.1%) were in the first-year JFL courses, while 122 of them (64.9%) were in the second-year courses. The entire sample represented JFL learners who were midway through the first-year elementary Japanese instruction to those who were at near completion of the second-year intermediate Japanese instruction at the universities at the time of their participation. Among them ninety-one (48.4%) were males, while 117 (51.6%) were females. A great majority of them (178, or 94.8%) were undergraduate students, while the rest included four graduate students as well as four high school students and community education program students attending Japanese courses. The participants represented a wide range of disciplinary areas including humanities, sciences, and arts, while 19 had not declared their majors at the time of their participation in this study. Ten different languages were reported as the participants’ primary languages: The large language groups were English (66.5%), Chinese (16.0%), Korean (3.7%), Indonesian (3.7%), and Cantonese (2.7%), while one to three speakers each (2.5% or less) of Malay, Tagalog, Thai, Spanish, and Swedish were also present in the sample.

3.2 Materials

The instruments used in this study were an information sheet and web-based summarization and free recall tasks. The information sheet was a one-page questionnaire in a short-answer format, in which the participants provided their gender, age, academic status, academic major, language background, Japanese language courses completed, and any other experience in the Japanese language.

Four web-based summary and recall tasks based on two prompt passages (Passages A and B) were developed for use in this study: Passage A Summary (Task 1), Passage B Summary (Task 2), Passage A Recall (Task 3), and Passage B Recall (Task 4). The two prompt passages were fictitious e-mail messages developed by the researcher that were 432–8 Japanese characters in length. The topics of Passages A and B were scheduling travel and apartment hunting, respectively, which were considered to be tasks familiar to the university-level JFL learners. Based on Meyer's classification of rhetorical organizations of texts, Passage A had features of a *collection* passage, while Passage B had those of *causation* and *comparison*. The linguistic characteristics were roughly parallel between the two passages in terms of orthography, syntactic structure, and vocabulary, the level of which were set at Grade 3 of the Japanese Language Proficiency Test (Japan Foundation and Association of International Education). This roughly corresponded to the exit level of an introductory Japanese course textbook, which is usually completed at the beginning or the midway through the second year of JFL instruction at the participating institutions (see Figure 1 below for a sample passage).

For both summarization and free recall, the participants were allowed to read the prompt passages as many times as they wanted and take notes, although referring back to the prompt passages or notes was not allowed once they moved on to the answer sheet. The directions specific to each task type were as follows. For the summarization tasks, the participants were instructed to read a prompt passage and summarize it in

!! Scroll down to see notes !!

Time spent so far: 0:03:29 (TESTTAKERID)%> (TASKID)%>

Context

Anna is an American student living in Hiroshima. Anna received an e-mail from a Japanese woman who Anna is planning to visit shortly. However, Anna does not completely understand what the e-mail says.

Instructions

Pretend that Anna asked you to write down briefly what the e-mail says in English. Read the e-mail to the right and summarize the e-mail in your own words for Anna. Your summary must be approximately 60 words in length. You may read the e-mail as many times as you like before you begin writing your summary. You can also take notes while reading. However, you CANNOT refer back to the e-mail or your notes once you begin writing your summary.

In your summary be sure to include the main points of the e-mail.

1. information that is useful for Anna and
2. information that Anna needs to understand for writing an adequate reply.

Reduce the information to main points only; avoid inclusion of redundant information and details. Try to condense the important information in the e-mail: your summary must provide a gist of the e-mail. Write your summary in connected sentences instead of a list of words or sentences. Your answer will be scored based on the amount and accuracy of main ideas and the effectiveness of condensation of information in your summary.

Please note that you CANNOT go back to the e-mail once you go on to the Answer Sheet.

アンナさん、

お手紙ありがとうございました。8月にこちらにいらっしゃるのが楽しみです。

広島から京都までの来かたはわかりますか。広島駅で京都へ行く新幹線にのれますよ。新幹線は高いですが、とてもはやくて、べんりです。広島から京都までは、2時間ぐらいです。

8月3日に広島を出る前に、何時ごろ京都に着くか、うちに電話してくだされませんか。その日は、朝からずっとうちにいるよいです。アンナさんが着く時間に、京都駅へむかえに行きます。

8月4日と5日は、アンナさんといっしょに姉のうちにあそびに行こうと思っていますが、どうですか。姉のところはわたしのうちから車で2時間ぐらいです。海のそばで、おいしい魚がたくさん食べられるし、近くにいいおんせん²もあります。

もしほかに行きたいところがあったら、知らせてくださいね。では、また。

Notes

1 新幹線 bullet train

Figure 1. Screen shot of Task 1 (Page 1).

English in order to help the recipient of the message understand what the message said. More detailed directions included that a summary must be approximately sixty words in length and cover information (1) that is useful for the recipient and (2) that the recipient needs to understand in order to write an adequate reply. For the free recall tasks, the participants were instructed to recall as many main ideas and details as possible without any length limitation. Each task was designed as a web-based reading exercise and was structured in the same way. In each task, the first screen displayed the directions and the reading passage, from where a user was navigated to the answer sheet on the second page by clicking a button when he/she was ready to provide a summary or recall protocol (as an example, see Figure 1).

3.3 Procedure

The research design is summarized in Figure 2. Each participant completed all four summarization and free recall tasks in two sixty-minute sessions with a two to three week interval in between. Each task was separately timed for thirty minutes. Each participant accessed to the project website from a convenient location for him/her (e.g., home, computer lab on campus) and completed all the tasks individually. Before Session 1, the participants were randomly assigned to Condition A or B. The order of presentation of the tasks was counterbalanced between the two conditions. See Figure 2. The two tasks within each session were presented to each participant in random order.

Session	Condition A	Condition B
1	Task 1 Passages A/Summary	Task 2 Passage B/Summary
	Task 4 Passage B/Free Recall	Task 3 Passage A/Recall
2	Task 2 Passage B/Summary	Task 1 Passages A/Summary
	Task 3 Passage A/Recall	Task 4 Passage B/Free Recall

Figure 2. Research design.

3.4 Scoring method

The summarization and free recall scoring rubrics are presented in Appendices A and B. The summarization scoring rubric consisted of three analytic rating scales: Main Idea Coverage, Main Idea Accuracy, and Integration. The free recall scoring rubric consisted of four analytic rating scales, where the first two, Main Idea Coverage and Main Idea Accuracy, were identical to those for summarization, while the other two, Detail Coverage and Detail Accuracy, were unique to free recall. All the rating scales

had five rating points representing: “No evidence (0),” “Poor (1),” “Moderate (2),” “Good (3),” and “Excellent (4).”

The raters were five doctoral students majoring in applied linguistics or Japanese linguistics with extensive experience in teaching university-level Japanese as a foreign language. They were all female native speakers of Japanese. This researcher conducted two rater training sessions, where the scoring rubrics were explained and sample responses of participants in a pilot study discussed.

Each summarization or a free recall protocol was rated by two independent raters. The participants’ summarization and free recall protocols were randomly assigned to the raters. Each rater rated from 250 to 340 summarization and recall protocols. Since two independent raters provided ratings on two tasks for each of the seven summarization and free recall rating scales, 28 ratings were obtained per participant. With regard to the agreement of two raters who rated the same protocol, exact and adjacent scores accounted for 60.7% (570) and 36.1% (339) of the entire 939 summary rating pairs obtained, respectively; however, the estimates were slightly better for the recall ratings with exact and adjacent scores accounting for 66.7% (854) and 31.5% (403) of the 1,280 total recall rating pairs obtained, respectively.

3.5 Analyses

After descriptive statistics were obtained for the scores assigned to the summarization and free recall protocols based on the scoring rubrics, analyses were made of outliers and score missingness. In addition, characteristics of the univariate and multivariate score distributions were analyzed although multivariate G theory does not make any distributional assumption (Brennan 295). This is because the present multivariate G study served as a preliminary analysis to inform the design of a structural equation modeling (SEM) of the summarization and free recall rating scales reported elsewhere (Sawaki), in which the multivariate normality of the distributions had direct implications to the selection of appropriate criteria for assessment of model fit.

In the present G-theory analyses, the generalizability of the summarization and free recall ratings were evaluated from two different perspectives: (1) similarities and differences in the generalizability coefficient (G coefficient) estimates at the individual rating scale level as well as at the composite score level for each task type; and (2) distinctiveness of the individual rating scales observed in the universe-score correlations among the rating scales, which are a G-theory analogue of disattenuated correlations. In order to address these two issues, three separate G studies were conducted. Each G study was a two-facet crossed design ($p \times s \times r$), where the persons (p) were modeled as the objects of measurement, and the prompt passages (s) and the ratings (r) were modeled as random facets of measurement. The individual summarization and free recall rating scales were treated as levels in the fixed facet. The

first G study of this two-facet crossed design was conducted for the three rating scales associated with the summarization task type only, while the second G study of the same design involved the four rating scales for the free recall task type only. The decision studies (D studies) corresponding to these two G studies for two passages and two ratings generated the univariate G-coefficient estimates for the individual summarization and free recall rating scales and the multivariate G-coefficient estimates for the composites for the two task types. The multivariate G coefficients for the two task types were obtained for the case where all the ratings awarded to a particular summarization or free recall protocol were averaged across all the associated ratings with equal weights. In order to investigate the distinctiveness of the individual rating scales, the universe-score correlations among the individual summarization and free recall rating scales were obtained from the D study for two passages and two ratings based on the third G study, in which all the summarization and free recall rating scales were analyzed simultaneously. *SPSS* 10.0 and *EQS* 6.0 (Bentler) were used for the preliminary analyses and *mGENOVA* (Brennan) for the multivariate G studies.

4. Results

4.1 Preliminary analyses

One hundred and fifty out of the 188 participants in this study completed all four summarization and free recall tasks. However, the intervals between the two sessions were longer (twenty-two to fifty-four days) than specified for 47 out of the 150 participants with complete data. In this study, too long an interval between the two sessions could be problematic because the participants' performance could change due to what they might have learned in the JFL courses in which they were enrolled. To ascertain the feasibility of including these 47 cases, a series of one-sample t-tests was conducted for each rating scale. The t-test results showed that the mean score increases for these 47 participants were not statistically significantly different from those of the students who had completed both sessions within twenty-one days. Thus the 47 cases were retained in the subsequent analyses.

Among the 150 cases in the sample, 19 cases had missing scores for at least one of the four tasks due to various technical problems experienced while participants completed the tasks and/or database errors. For fourteen out of the nineteen participants, scores for only one of the tasks were missing. EM imputation (Dempster, Laird and Rubin) was applied to estimate these missing scores. The remaining 5 cases, which had missing scores for two or more tasks, were deleted listwise. None of the remaining 145 cases was identified as a univariate outlier on any of the twenty-eight ratings by using a z-score of 3.3 as the criterion. Eleven of the 145 cases were identified as multivariate outliers based on Mahalanobis distance (see Ishimura for a brief overview of Mahalanobis distance), however. A listwise deletion of these multivariate outliers resulted in the final sample size of 134. The normality of the

twenty-eight univariate score distributions was assessed based on a review of histograms as well as standardized skewness and kurtosis values. All the score distributions were univariate normal, using the absolute skewness and kurtosis values of 3.3 as the criteria. Multivariate normality of the distributions was assessed by using normalized Mardia's coefficient (Mardia, "Measures"; Mardia, "Applications"), which can be interpreted as a z-score. The value for the twenty-eight measures was 2.25. Considering the large number of measured variables in this dataset, this value does not indicate considerable departure from multivariate normality.

The descriptive statistics for the summary and recall rating scales (means, standard deviations, skewness, kurtosis, minimum and maximum scores) for the final sample are summarized in Appendix C. The mean ratings for the individual rating scales are graphically represented in Figure 3. There are two notable patterns in the mean ratings across the rating scales shown in Figure 3. First, the similarity of the means across the two independent ratings provided across the rating scales suggests that the rating effect was minimal on the summarization and free recall ratings. Second, the noticeable mean rating differences observed between the prompt passages indicate presence of a passage effect. On average, the participants tended to do considerably better on the summarization and free recall tasks based on Passage A (Tasks 1 and 3, respectively) for all the summarization rating scales and the main idea-related free recall rating scales. This scoring pattern was reversed for the Detail Coverage and Detail Accuracy rating scales, however. On these rating scales, the participants tended to do better for the free recall task based on Passage B (Task 4) compared to the other based on Passage A (Task 3), although the across-passage mean rating differences were relatively smaller compared to those for the other rating scales.

Figure 3: Summarization and free recall rating means.

4.2 Multivariate generalizability study analysis

4.2.1 Generalizability of individual rating scales

The obtained G coefficients for the individual summarization and free recall rating scales obtained from the two separate D studies for the two task types are presented in Table 1. As can be seen at the bottom of the table, the G-coefficient estimates for the individual summarization rating scales ranged from .69 (Main Idea Coverage) to .73 (Integration), while those for the free recall rating scales were higher, ranging from .73 (Main Idea Accuracy) to .82 (Detail Coverage and Detail Accuracy).

Relative contributions of sources of score variation in terms of variance component estimates (presented in Table 1) provide us with insights into the observed differences between the task types in the generalizability of ratings. With regard to the three summarization rating scales, the variance component estimates for the persons suggest that 64.9% to 66.6% of the variances on these three rating scales were explained by the

Table 1. Estimated D-Study Variance Components for Individual Rating Scales (2 passages x 2 ratings)

Source of variation	Estimated variance component						
	Summary			Recall			
	Main idea coverage	Main idea accuracy	Integration	Main idea coverage	Main idea accuracy	Detail coverage	Detail accuracy
<i>Person (p)</i>	0.45 64.9%	0.49 66.6%	0.66 66.0%	0.78 66.6%	0.76 65.4%	1.09 80.4%	1.08 81.8%
<i>Passage (s)</i>	0.04 5.4%	0.05 6.2%	0.09 9.1%	0.14 11.9%	0.12 10.2%	0.01 0.6%	0.00 ^a 0.0%
<i>Rating (r)</i>	0.00 ^a 0.0%	0.00 ^a 0.0%	0.00 0.1%	0.00 0.0%	0.00 ^a 0.0%	0.00 ^a 0.0%	0.00 ^a 0.0%
<i>p x s</i>	0.14 20.1%	0.14 19.5%	0.18 18.0%	0.21 17.6%	0.24 20.4%	0.20 15.0%	0.19 14.6%
<i>p x r</i>	0.01 2.0%	0.00 ^a 0.0%	0.00 ^a 0.0%	0.00 ^a 0.0%	0.01 0.7%	0.00 ^a 0.0%	0.00 0.2%
<i>s x r</i>	0.01 1.0%	0.01 0.8%	0.00 0.0%	0.00 ^a 0.0%	0.00 ^a 0.0%	0.00 0.1%	0.00 0.2%
<i>psr,e</i>	0.05 6.6%	0.05 6.9%	0.07 6.8%	0.05 3.9%	0.04 3.3%	0.05 3.9%	0.04 3.3%
Total	0.69	0.74	1.00	1.18	1.16	1.35	1.32
G-coefficient	0.69	0.72	0.73	0.76	0.73	0.82	0.82

^a Negative variance component fixed to zero after calculation.

participants' true differences in the abilities being assessed on the rating scales. However, relatively large portions of the score variation were accounted for by the passage(s), person-by-passage ($p \times s$), and residual (psr, e) variance components. Particularly, the relatively high contributions of the passage (p) main effect and the person-by-passage ($p \times s$) interaction reflected the passage effect shown in the preliminary analyses. The passage (s) variance component estimates accounted for 5.4% to 9.1% of the score variances on these three rating scales, suggesting that the participants scored better on these rating scales, in average, when the summarization task was based on Passage A. The person-by-passage ($p \times s$) variance component estimates explained as much as 18.0% to 20.1% of the total score variances, indicating that the participants' rank-ordering on these rating scales depended on the passage on which the task was based. Finally, the residual (psr, e) variance component estimates indicated that 6.6% to 6.9% of the score variances were accounted for by the confounding effect of the person-by-passage-by-rating interaction and/or sources of systematic and nonsystematic error variability that were not captured by the present two-facet D study.

The variance component estimates for the free recall rating scales can be found in the right four columns of Table 1. The pattern of the relative sizes of the variance component estimates across the different sources of variation for the Main Idea Coverage and Main Idea Accuracy rating scales were similar to those for the summarization rating scales. For these two rating scales, the person (p) variance components accounted for only 65.4% to 66.6% of the total score variances, while the passage (s) and person-by-passage ($p \times s$) variance components explained as much as 10.2% to 11.9% and 17.6% to 20.4% of the score variation on these rating scales, respectively. The pattern observed for the Detail Coverage and Detail Accuracy rating scales was different, however. The true ability differences among the participants accounted for as much as 80.4% to 81.8% of the score variances on these rating scales. Moreover, the passage (p) variance component estimates were minimal (0.6% to 0.0%), while those for the person-by-passage ($p \times s$) interaction (14.6% to 15.0%) were relatively small and yet still substantial. Unlike the summarization rating scales, however, the residual (psr, e) variance components were relatively small across all the free recall rating scales, although they were still nonzero.

A common finding across the two task types in the D-study results above was the minimal variance component estimates for the sources of variation involving the ratings (the rating main effect as well as the person-by-rating and passage-by-rating interactions). One potential explanation for this is that counter-balancing of the individual raters in the rating design (i.e., one rater provided the first ratings for a batch of protocols, while the same rater provided the second ratings for another) and the modeling of these counter-balanced ratings rather than the individual raters in these D studies washed out the effects of potential severity differences among the individual

raters on the rating variability.

The composite universe score and relative error variances as well as the multivariate generalizability coefficients obtained from the separate analyses of the rating scales associated with each task type in the first two D studies (for two passages and two ratings) are presented in Table 2.² A multivariate G coefficient is defined as “the ratio of composite universe score variance to itself plus composite relative error variance” (Brennan 306). Thus, the multivariate G coefficient for summarization (0.74) was obtained, for example, by calculating the ratio of the composite universe score variance (0.53) to the sum of itself and the composite relative error variance ($0.53+0.19=0.72$). The obtained multivariate G coefficient was again higher for free recall compared to that for summarization, suggesting the higher generalizability of free recall rating at the composite score level as well.

Table 2. Composite universe score and relative error variances and multivariate G-coefficient estimates (D-study for 2 passages x 2 ratings)

	Summarization	Free recall
Composite universe-score variance	0.53	0.92
Composite relative error variance	0.19	0.19
Multivariate G coefficient	0.74	0.83

4.2.1 Interrelationships among the rating scales

In the third G study, where all the summarization and free recall rating scales were analyzed simultaneously, the universe-score correlations among all the summarization and free recall rating scales were obtained. The estimated universe-score correlations across the summarization and free recall rating scales are presented in Table 3.³ The estimated values of approximately 1.0 common across the rating scales indicate that, if it were possible to measure the abilities as defined by the summarization and recall rating scales without error, the ratings given on these seven rating scales correlate perfectly. The unit intercorrelations across the rating scales also suggest that the ratings on these rating scales as they stand now are not distinguishable from one another. A comparison of these universe-score correlations with the observed Pearson correlations among the rating scales signifies the impact of the various sources of error on the observed correlations. As can be seen in Table 4, the convergence among the different rating scales within the same task type in terms of the Pearson correlation coefficients was extremely high, ranging from .96 to .98 for the summary tasks, and from .90 to .99 for the recall tasks. However, the correlations among the rating scales were not perfect; and the correlations between the pairs of ratings across the task types (.79 to .83) were uniformly lower than those pairs of ratings within the same task type. The dramatic difference in the patterns of the interrelationships among the rating scales in the universe-score correlations and the observed correlations is primarily due to the

presence of the sizable variance-covariance component estimates due to the different facets of measurement modeled in the present D study, which “masked” the true relationships among the strength of associations among the rating scales.

Table 3: Universe-score Correlations among the Summary and Recall Rating Scales (2 passages x 2 ratings)

Rating scale	Summary			Recall			
	Main Cov.	Main Acc.	Int.	Main Cov	Main Acc.	Det. Cov.	Det. Acc.
S Main Idea Coverage	1.00						
S Main Idea Accuracy	1.00 ^a	1.00					
S Integration	1.00 ^a	1.00	1.00				
R Main Idea Coverage	1.00 ^a	1.00 ^a	1.00 ^a	1.00			
R Main Idea Coverage	1.00 ^a	1.00 ^a	1.00 ^a	1.00 ^a	1.00		
R Detail Coverage	1.00 ^a	1.00 ^a	1.00 ^a	1.00	1.00 ^a	1.00	
R Detail Accuracy	1.00 ^a	1.00 ^a	1.00 ^a	1.00	1.00 ^a	1.00	1.00

^a Universe-score correlation estimates of over 1 adjusted to 1.00 after calculation.

Table 4: Observed Pearson Correlations among the Summary and Recall Rating Scales

Rating scale	Summary			Recall			
	Main Cov.	Main Acc.	Int.	Main Cov	Main Acc.	Det. Cov.	Det. Acc.
S Main Idea Coverage	1.00						
S Main Idea Accuracy	.98	1.00					
S Integration	.96	.97	1.00				
R Main Idea Coverage	.80	.81	.81	1.00			
R Main Idea Coverage	.81	.83	.82	.99	1.00		
R Detail Coverage	.80	.81	.79	.90	.91	1.00	
R Detail Accuracy	.80	.80	.80	.91	.91	.98	1.00

5. Discussion

In the present study, multivariate G theory analyses were conducted for the summarization and free recall ratings. The G coefficients obtained for the individual summary and free recall rating scales as well as for the composite scores suggested that

the recall tasks tended to be rated more reliably, although the generalizability of both task types were moderate. The differences in the relative contributions of the different sources of score variation to the total score variability between summarization and free recall were observed at three places. The first is the sizable proportion of variances accounted for by the relatively large passage and person-by-passage variance component estimates for all the summarization ratings and the main idea-related free recall rating scales. The second is the absence of the mean differences as well as the relatively small person rank-ordering differences across the passages for the detail-related free recall rating scales. The third is the nonzero but relatively small contribution of the residual variance components to the variability of free recall ratings. Among these, the relatively larger contribution of the person-by-passage interaction and residual variance components to the variability of the summarization ratings is attributable to the lower G coefficient estimates for the summarization rating scales compared to those for the free recall rating scales. This is because these two sources of variation contribute to the measurement error in the calculation of the G coefficients.

The difference observed between the two task types in the generalizability of the ratings in the present multivariate G-theory analysis raises two issues that require further investigation: the participants' performance differences across the two passages on the summarization and free recall rating scales, or the passage effect, and the interpretation of the residual variance.

First, a potential explanation for the observed passage effect may be some insufficiently controlled task variables such as the text structures of the prompt passages. For example, the straightforwardness of the collection structure in Passage A as opposed to the complexity of that in Passage B, which involved causation and comparison, might have impacted the accessibility of the main ideas, leading to the participants' superior performance on the summarization and the main idea-related free recall rating scales for the Passage A-based tasks. Conversely, the same text structure might have acted differently for recall of the details. The embedding of details in the comparison structures in Passage B might have been the reason for the good level of recall of the particular details, while the loosely-structured nature of Passage A might have made retrieval of the associated details ineffective (cf. see Meyer for the relationship between rhetorical organization and effectiveness of information retrieval).

Although such a speculation may sound reasonable to account for the seeming presence of the passage effect, the large contributions of the passage and person-by-passage variance components to the variability of the summarization and free recall rating observed in this study must be interpreted carefully. This finding resonates with those in the performance assessment literature, where the sampling of tasks has often been found to be a major source of score variability. However, recent performance assessment research has suggested that the seemingly large task-sampling variability may in fact be a reflection of a "hidden" facet,—the testing

occasion—which is often confounded with tasks (Cronbach, Linn, Brennan, and Haertel; McBee and Barnes; Shavelson, Ruiz-Primo, and Wiley; Webb, Schlackman, and Sugrue). This issue is relevant to the present study because each participant saw different prompt passages for the same task type on two different occasions. The effects of the testing occasion and its confounding with the prompt passages could be adequately addressed by conducting a G study that explicitly models the testing occasion as a facet of measurement (e.g., a fully-crossed G study for a person $\times$ rater $\times$ passage $\times$ occasion design). However, such an analysis was not possible in the present research design. The “clean” occasion facet could not be separated from the task type because of the different combinations of the prompt passages with the task types used across the two testing occasions in this study. In the future, employing research designs where each participant encounters identical tasks—based on the same prompt passage and the same task type—in two occasions may explicate the potential effects of the testing occasion and its interactions with other facets of measurement on observed ratings.

The second issue, the relatively large residual variance component estimates for the summarization rating scales compared to those for the free recall rating scales, seems to support Cohen’s point about the susceptibility of summarization ratings to subjective judgments of raters discussed in the literature review above. This turned out to be a challenging issue in this study as well. An analysis of the audio recording of the verbal protocols that these raters produced during the rating sessions, which was reported in detail elsewhere (Sawaki), revealed the raters’ increased uncertainties about what score to assign to summarization protocols, especially when the reproduction and integration of information used in summarization departed from the expressions used in the prompt passages. Considering the relative absence of this extra thinking process for free recall reported in the verbal protocols, the increased rater subjectivity may have manifested in the residual variance components for the summarization rating scales, which subsume both systematic and nonsystematic error variances not captured by the present two-facet crossed G study design.

Another major finding in this study was the indistinguishable nature of the individual rating scales suggested by the extremely high universe-score correlations. There are a few possible explanations to this finding. One possibility is the presence of a closely overlapping common underlying dimension across the rating scales at the construct level. This may be accounted for by the nature of the overlap in the construct definitions across the two task types, where both summarization and free recall were hypothesized to be tapping into understanding of main ideas of prompt passages. For this reason, the Main Idea Coverage and Main Idea Accuracy rating scales were shared across the task types. Although the raters’ verbal protocol study cited above (Sawaki) revealed that the raters partially adopted differential rating criteria for these rating scales across the task types, a considerable overlap of score variance among these

rating scales was unavoidable to some extent. A second possibility is that the raters did not sufficiently distinguish among the rating scales, although the constructs themselves may be theoretically and empirically distinguishable. The possibility of this explanation cannot be ruled out in this case because of the nature of the present rating design, where individual raters were allowed to provide ratings for all the rating scales associated with a task type at once rather than one rating at a time. This could lead to a halo effect, where the first impression that a rater developed on a particular protocol might have carried over to the ratings on the other rating scales. A similar finding was reported by a previous multivariate G study of performance assessment in painting (Nussbaum). Thus comparing results across different rating designs may allow one to gauge the extent to which a halo effect accounts for the indistinguishableness of the rating scales.

6. Conclusions and directions for further research

The results of the present G study of summarization and free recall ratings indicated that with rater training a sufficient level of generalizability could be achieved for ratings awarded to summarization and free recall protocols based on analytic rating scales. In this study, summarization ratings were found to be scored relatively less reliably and to be more susceptible to method effect compared to free recall. However, it is premature to draw a conclusion about the superiority of one task type over the other in terms of reliability. In future research, various issues must be investigated closely in order to evaluate the plausible alternative explanations for the observed relative susceptibility of summarization ratings to measurement error compared to free recall ratings. Some suggested directions for future research to achieve this goal include (1) tighter control for variables associated with the prompt passages and use of a study design that allows explicit modeling of testing occasions for further comparisons of the generalizability of summarization and free recall ratings; and (2) refinement of the operational definitions of the rating scales and comparisons of rating designs that may increase the distinctiveness of the constructs tapped into by analytic rating scales.

Acknowledgements

This article is based on the author's doctoral dissertation completed at the University of California, Los Angeles in 2003. The author is indebted to the dissertation committee, Professor Lyle F. Bachman, Professor Eva L. Baker, Professor Charles Goodwin, Professor John H. Schumann, and Professor Noreen Webb, for their continuous guidance and support throughout the completion of this project; and to Kim Fryer, Yong-Won Lee and Xiaoming Xi for their thorough review of an earlier version of this manuscript. The author is also grateful to the Japan Language Testing Association and the TOEFL Program at Educational Testing Service for their generous financial support

of this project.

Notes

¹ The author is currently at Educational Testing Service.

^{2,3} The data presented in Tables 2 and 3 were calculated based on the variance-covariance component matrices for the different sources of variation obtained in the three D studies (for two passages and two ratings) reported in this article. The matrices are available from the author: Yasuyo Sawaki, Center for Validity Research, MS 10-R, Educational Testing Service, Rosedale Road, Princeton, NJ 08541, U.S.A. (ysawaki@ets.org).

Works Cited

- Bachman, Lyle F., Brian Lynch, and Maureen Mason. "Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Speaking." Language Testing 12 (1995): 238–57.
- Bachman, Lyle F., and Palmer, Adrian. Language Testing in Practice. Oxford: Oxford University Press, 1996.
- Baker, L., and Ann L. Brown. "Metacognitive Skills and Reading." Handbook of Reading Research. Ed. P. David Pearson. New York: Longman, 1984. 352–594.
- Bentler, Peter M. EQS for Windows. Vers. 6.0, Beta Build 94. Encino, CA: Multivariate Software, Inc., 1985–2002.
- Bernhardt, Elizabeth B. "Three Approaches to Reading Comprehension in Intermediate German." Modern Language Journal 67.2 (1983): 111–15.
- . Reading Development in a Second Language: Theoretical, Empirical, and Classroom Perspectives. Norwood, New Jersey: Ablex, 1991.
- Bolus, R. E., Frances B. Hinofotis, and Kathleen M. Bailey. "An Introduction to Generalizability Theory in Second Language Research." Language Learning 32 (1982): 245–58.
- Brennan, Robert L. mGENOVA. Vers. 2.0. Iowa City, IA: University of Iowa, 1999.
- . Generalizability Theory. New York: Springer-Verlag, 2001.
- Brown, James Dean. "The Relative Importance of Persons, Items, Subtests and Languages to TOEFL Test Variance." Language Testing 16.2 (1999): 217–38.
- Brown, James Dean, and Kathleen M. Bailey. "A Categorical Instrument for Scoring Second Language Writing Skills." Language Learning 34.4 (1984): 21–42.
- Brown, Ann L., and Jeanne D. Day. "Macrorules for Summarizing Texts: The Development of Expertise." Journal of Verbal Learning and Verbal Behavior 22 (1983): 1–14.
- Carrell, Patricia L. "Content and Formal Schemata in ESL Reading." TESOL Quarterly 21.3 (1987): 461–81.
- Cohen, Andrew D. "The role of instructions in testing summarizing ability." A New

- Decade of Language Testing Research. Eds. Dan Douglas and Carole Chapelle. Alexandria, VA: Teachers of English to Speakers of Other Languages.
- . "English for Academic Purposes in Brazil: The Use of Summary Tasks." From Testing to Assessment: English as an International Language. Eds. Clifford Hill and Kate Parry. London: Longman, 1994.
- Cronbach, Lee J., G. C. Gleser, H. Nanda, and N. Rajaratnam. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.
- Cordero-Ponce Wanda L. "Summarization Instruction: Effects on Foreign Language Comprehension and Summarization of Expository Texts." Reading Research and Instruction 29.4 (2000): 329-50.
- Dempster, Arthur P., Laird, Nan M., and Donald B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society 39.1 (1977):1-38.
- Ishimura, Sadao (石村貞夫). 「すぐわかる多変量解析」東京：東京図書株式会社. Japan Foundation and Association of International Education, eds. Japanese Language Proficiency Test: Test Content Specifications. 2nd ed. Tokyo: Bonjinsha, 1999.
- Johns, Ann M., and Patricia Mayes. "An Analysis of Summary Protocols of University ESL Students." Applied Linguistics 11.3 (1990): 253-71.
- Johnston, Peter H. "Assessment in Reading." Handbook of Reading Research. Ed. P. David Pearson. New York: Longman, 1984. 147-82
- Kim, Sung-Ae. "Characteristics of EFL Readers' Summary Writing: A Study with Korean University Students." Foreign Language Annals 34.6 (2001): 569-81.
- Kintsch, Walter, and Teun A. van Dijk. "Toward a Model of Text Comprehension and Production." Psychological Review 85.5 (1978): 363-94.
- Kobayashi, Miyoko. "Method Effects on Reading Comprehension Test Performance: Text Organization and Response Format." Language Testing 19.2 (2002): 193-220.
- Lee, James. F. "Background Knowledge and L2 Reading." Modern Language Journal 70.4 (1986): 350-54.
- Lynch, Brian, and Tim. McNamara. "Using G-theory and Many-facet Rasch Measurement in the Development of Performance Assessments of the ESL Speaking Skills of Immigrants." Language Testing 15.2 (1998): 158-80.
- Mardia, K.V. "Measures of Multivariate Skewness and kurtosis with applications." Biometrika 57 (1970):519-530.
- . "Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya, B36 (1974): 115-128.
- McBee, Maridyth M., and Laura L. B. Barnes. "The Generalizability of a Performance Assessment Measuring Achievement in Eighth-Grade Mathematics." Applied Measurement in Education 11.2 (1998): 179-94.

- Meyer, Bonnie J. F. The Organization of Prose and Its Effects on Memory. Amsterdam: North-Holland Publishing Co., 1975.
- Nussbaum, Albert. "Multivariate Generalizability Theory in Educational Measurement: An Empirical Study." Applied Psychological Measurement 8.2 (1984): 219–30.
- Paris, Scott. G., Barbara A. Wasik, and Julianne. C. Turner. "The Development of Strategic Readers." Handbook of Reading Research. Eds. Rebecca Barr, Michael L. Kamil, Peter B. Mosenthal, and P. David Pearson. Vol. II.. London: Longman, 1991. 609–40.
- Riley, Gail L., and James F. Lee. "A Comparison of Recall and Summary Protocols as Measures of Second Language Reading Comprehension." Language Testing 13.2 (1996): 173–90.
- Rinehart, Steven D., Steven A. Stahl, and Lawrence G. Erickson. "Some Effects of Summarization Training on Reading." Reading Research Quarterly 21.4 (1986): 422–35.
- Rusch, G.M., and Stoddard, G.D. Tests and Measurements in High School Instruction. Yonkers, N.Y.: World Book, 1927.
- Sawaki, Yasuyo. "A Comparison of Summarization and Free Recall as Reading Comprehension Tasks in Web-Based Assessment of Japanese as a Foreign Language." Diss.: UCLA, 2003. Ann Arbor, MI: UMI, 2004. AAT 3104336
- Shavelson, Richard J., Ruiz-Primo, Maria A., & Wiley, Edward W. "Note on sources of sampling variability in science performance assessments." Journal of Educational Measurement 36.1 (1999): 61-67.
- SPSS. Vers. 10.0. Chicago: SPSS Inc., 1999.
- Webb, Noreen M., and Richard J. Shavelson. "Multivariate Generalizability of General Educational Development Ratings." Journal of Educational Measurement 18.1 (1981): 13–22.
- Webb, Noreen M., Richard J. Shavelson, and E. Maddahian. "Multivariate Generalizability Theory." Generalizability Theory: Inferences and Practical Applications. Ed. Leslie J. Fyans, Jr. San Francisco, CA: Jossey-Bass, 1983. 67–81.
- Webb, Noreen M., Jonah Schlackman, and Brenda Sugrue "The Dependability and Interchangeability of Assessment Methods in Science." Applied Measurement in Education. 13 (2000): 277–302.
- Winograd, Peter N. "Strategic Difficulties in Summarizing Texts." Reading Research Quarterly 19.4 (1984): 404–25.

Appendix A: Summarization scoring rubric

A. MAIN IDEA COVERAGE: The extent to which main ideas are covered.

- 4 EXCELLENT: Complete coverage of main ideas
- 3 GOOD: Coverage of most main ideas
- 2 MODERATE: Moderate coverage of main ideas
- 1 POOR: Some coverage of main ideas
- 0 NO: No coverage of main ideas

B. MAIN IDEA ACCURACY: The extent to which representation of main ideas is accurate. Inaccurate information refers to extraneous ideas, which consist of information that was not present in the original passage including distortions and intrusions.

Distortion: a misleading statement, an incomplete item, or a confusion of facts.

Intrusion: the reader's prior knowledge or opinion about the content of the text rather than the information given in the text.

- 4 EXCELLENT: Accurate representation of main ideas throughout with no distortions
- 3 GOOD: Accurate representation of main ideas most of the time, possibly with a few minor distortions
- 2 MODERATE: Accurate representation of main ideas half of the time,

possibly with some notable distortions

- 1 POOR: Accurate representation of main ideas less than half of the time, possibly with major distortions
- 0 NO: No evidence of accurate representation of main ideas

C. INTEGRATION: The extent to which the information in the text is presented succinctly by using strategies such as deleting unnecessary information, combining information across idea units, reordering information in text, and by using topic sentences. Effective integration refers to the use of integration (1) at right places and (2) without distortion of the original meaning.

- 4 EXCELLENT: Effective and appropriate integration throughout
- 3 GOOD: Effective and appropriate integration most of the time
- 2 MODERATE: Some effective integration, possibly with some instances of inappropriate use
- 1 POOR: Some integration, but they are mostly inappropriate or misleading OR little evidence of the use of integration
- 0 NO: No evidence of the use of integration

Appendix B: Free recall scoring rubric

A. MAIN IDEA COVERAGE: The extent to which main ideas are covered.

- 4 EXCELLENT: Complete coverage of main ideas
- 3 GOOD: Coverage of most main ideas
- 2 MODERATE: Moderate coverage of main ideas
- 1 POOR: Some coverage of main ideas
- 0 NO: No coverage of main ideas

B. MAIN IDEA ACCURACY: The extent to which representation of main ideas is accurate. Inaccurate information refers to extraneous ideas, which consist of information that was not present in the original passage including distortions and intrusions.

Distortion: a misleading statement, an incomplete item, or a confusion of facts.

Intrusion: the reader's prior knowledge or opinion about the content of the text rather than the information given in the text.

- 4 EXCELLENT: Accurate representation of main ideas throughout with no distortions
- 3 GOOD: Accurate representation of main ideas most of the time, possibly with a few minor distortions
- 2 MODERATE: Accurate representation of main ideas half of the time, possibly with some notable distortions
- 1 POOR: Accurate representation of main ideas less than half of the time, possibly with major distortions
- 0 NO: No evidence of accurate representation of main ideas

C. DETAIL COVERAGE: The extent to which details are covered.

- 4 EXCELLENT: Complete coverage of details
- 3 GOOD: Coverage of most details
- 2 MODERATE: Moderate coverage of details
- 1 POOR: Some coverage of details
- 0 NO: No coverage of details

D. DETAIL ACCURACY: The extent to which representation of details is accurate. Inaccurate information refers to extraneous ideas, which consists of information that was not present in the original passage including distortions and intrusions.

Distortion: a misleading statement, an incomplete item, or a confusion of facts.

Intrusion: the reader's prior knowledge or opinion about the content of the text rather than the information given in the text.

- 4 EXCELLENT: Accurate representation of details throughout with no distortions
- 3 GOOD: Accurate representation of details most of the time, possibly with a few minor distortions
- 2 MODERATE: Accurate representation of details half of the time, possibly with some notable distortions
- 1 POOR: Accurate representation of details less than half of the time, possibly with major distortions
- 0 NO: No evidence of accurate representation of details

Appendix C: Descriptive statistics for summarization and free recall ratings

Rating scale	Prompt	Rater	N	Mean	STD	Skewness	Kurtosis	Min	Max
<u>Summary task</u>									
Main Idea	Passage A	1	134	2.02	0.94	0.38	-2.25	0	4
Coverage		2	134	1.90	0.92	1.00	-1.79	0	4
	Passage B	1	134	1.41	0.99	0.87	-2.36	0	3
		2	134	1.64	1.01	0.25	-2.37	0	4
Main Idea	Passage A	1	134	1.90	0.94	0.41	-2.26	0	4
Accuracy		2	134	1.75	0.98	1.28	-2.13	0	4
	Passage B	1	134	1.27	0.99	1.37	-2.26	0	3
		2	134	1.46	1.05	0.77	-2.38	0	4
Integration	Passage A	1	134	1.84	1.18	0.28	-2.36	0	4
		2	134	1.72	1.09	1.10	-1.97	0	4
	Passage B	1	134	1.19	1.11	2.19	-2.46	0	4
		2	134	1.16	1.16	2.06	-2.88	0	4
<u>Recall task</u>									
Main Idea	Passage A	1	134	2.51	1.05	-2.07	-1.73	0	4
Coverage		2	134	2.49	1.06	-1.56	-2.22	0	4
	Passage B	1	134	1.77	1.26	0.61	-2.80	0	4
		2	134	1.72	1.29	0.49	-2.98	0	4
Main Idea	Passage A	1	134	2.34	1.08	-1.04	-2.26	0	4
Accuracy		2	134	2.39	1.11	-1.03	-2.33	0	4
	Passage B	1	134	1.68	1.23	0.81	-2.73	0	4
		2	134	1.66	1.29	0.47	-3.12	0	4
Detail	Passage A	1	134	2.19	1.24	-0.12	-2.55	0	4
Coverage		2	134	2.10	1.26	0.21	-2.69	0	4
	Passage B	1	134	2.33	1.38	-1.16	-3.03	0	4
		2	134	2.36	1.31	-1.17	-2.91	0	4
Detail	Passage A	1	134	2.05	1.26	-0.30	-2.58	0	4
Accuracy		2	134	1.93	1.21	0.57	-2.38	0	4
	Passage B	1	134	2.07	1.34	-0.14	-2.96	0	4
		2	134	2.14	1.32	-0.89	-2.86	0	4

Appendix C: Descriptive statistics for summarization and free recall ratings

Rating scale	Prompt	Rater	N	Mean	STD	Skewness	Kurtosis	Min	Max
<u>Summary task</u>									
Main Idea	Passage A	1	134	2.02	0.94	0.38	-2.25	0	4
Coverage		2	134	1.90	0.92	1.00	-1.79	0	4
	Passage B	1	134	1.41	0.99	0.87	-2.36	0	3
		2	134	1.64	1.01	0.25	-2.37	0	4
Main Idea	Passage A	1	134	1.90	0.94	0.41	-2.26	0	4
Accuracy		2	134	1.75	0.98	1.28	-2.13	0	4
	Passage B	1	134	1.27	0.99	1.37	-2.26	0	3
		2	134	1.46	1.05	0.77	-2.38	0	4
Integration	Passage A	1	134	1.84	1.18	0.28	-2.36	0	4
		2	134	1.72	1.09	1.10	-1.97	0	4
	Passage B	1	134	1.19	1.11	2.19	-2.46	0	4
		2	134	1.16	1.16	2.06	-2.88	0	4
<u>Recall task</u>									
Main Idea	Passage A	1	134	2.51	1.05	-2.07	-1.73	0	4
Coverage		2	134	2.49	1.06	-1.56	-2.22	0	4
	Passage B	1	134	1.77	1.26	0.61	-2.80	0	4
		2	134	1.72	1.29	0.49	-2.98	0	4
Main Idea	Passage A	1	134	2.34	1.08	-1.04	-2.26	0	4
Accuracy		2	134	2.39	1.11	-1.03	-2.33	0	4
	Passage B	1	134	1.68	1.23	0.81	-2.73	0	4
		2	134	1.66	1.29	0.47	-3.12	0	4
Detail	Passage A	1	134	2.19	1.24	-0.12	-2.55	0	4
Coverage		2	134	2.10	1.26	0.21	-2.69	0	4
	Passage B	1	134	2.33	1.38	-1.16	-3.03	0	4
		2	134	2.36	1.31	-1.17	-2.91	0	4
Detail	Passage A	1	134	2.05	1.26	-0.30	-2.58	0	4
Accuracy		2	134	1.93	1.21	0.57	-2.38	0	4
	Passage B	1	134	2.07	1.34	-0.14	-2.96	0	4
		2	134	2.14	1.32	-0.89	-2.86	0	4

Appendix C: Descriptive statistics for summarization and free recall ratings

Rating scale	Prompt	Rater	N	Mean	STD	Skewness	Kurtosis	Min	Max
<u>Summary task</u>									
Main Idea	Passage A	1	134	2.02	0.94	0.38	-2.25	0	4
Coverage		2	134	1.90	0.92	1.00	-1.79	0	4
	Passage B	1	134	1.41	0.99	0.87	-2.36	0	3
		2	134	1.64	1.01	0.25	-2.37	0	4
Main Idea	Passage A	1	134	1.90	0.94	0.41	-2.26	0	4
Accuracy		2	134	1.75	0.98	1.28	-2.13	0	4
	Passage B	1	134	1.27	0.99	1.37	-2.26	0	3
		2	134	1.46	1.05	0.77	-2.38	0	4
Integration	Passage A	1	134	1.84	1.18	0.28	-2.36	0	4
		2	134	1.72	1.09	1.10	-1.97	0	4
	Passage B	1	134	1.19	1.11	2.19	-2.46	0	4
		2	134	1.16	1.16	2.06	-2.88	0	4
<u>Recall task</u>									
Main Idea	Passage A	1	134	2.51	1.05	-2.07	-1.73	0	4
Coverage		2	134	2.49	1.06	-1.56	-2.22	0	4
	Passage B	1	134	1.77	1.26	0.61	-2.80	0	4
		2	134	1.72	1.29	0.49	-2.98	0	4
Main Idea	Passage A	1	134	2.34	1.08	-1.04	-2.26	0	4
Accuracy		2	134	2.39	1.11	-1.03	-2.33	0	4
	Passage B	1	134	1.68	1.23	0.81	-2.73	0	4
		2	134	1.66	1.29	0.47	-3.12	0	4
Detail	Passage A	1	134	2.19	1.24	-0.12	-2.55	0	4
Coverage		2	134	2.10	1.26	0.21	-2.69	0	4
	Passage B	1	134	2.33	1.38	-1.16	-3.03	0	4
		2	134	2.36	1.31	-1.17	-2.91	0	4
Detail	Passage A	1	134	2.05	1.26	-0.30	-2.58	0	4
Accuracy		2	134	1.93	1.21	0.57	-2.38	0	4
	Passage B	1	134	2.07	1.34	-0.14	-2.96	0	4
		2	134	2.14	1.32	-0.89	-2.86	0	4