Note-taking quality and performance on an L2 academic listening test
Min-Young Song
Language Testing published online 24 October 2011
DOI: 10.1177/0265532211415379

The online version of this article can be found at:
http://ltj.sagepub.com/content/early/2011/06/21/0265532211415379

A more recent version of this article was published on - Jan 25, 2012

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for Language Testing can be found at:

Email Alerts: http://ltj.sagepub.com/cgi/alerts

Subscriptions: http://ltj.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Version of Record - Jan 25, 2012

>> OnlineFirst Version of Record - Oct 24, 2011

What is This?
Note-taking quality and performance on an L2 academic listening test

Min-Young Song
Korea Institute for Curriculum & Evaluation, Korea

Abstract
This study investigated the relationships among the quality of L2 test takers’ notes evaluated in terms of different levels of information and test takers’ performance on open-ended listening tasks tapping into different comprehension subskills. In addition, this study examined the invariance of the structural relationships among the variables across two different note-taking formats, that is, a blank format and an outline format, by employing a multi-group structural equation modeling (SEM) approach. The results indicated that note quality measures, in particular the number of topical ideas found in the notes and the organization of these notes, may be good indicators of test takers’ second language academic listening proficiency. It was also found that despite the invariance of structural relationships among variables across the two note-taking formats, the associations between the open-ended listening measures and note quality measures were slightly stronger in the outline format than in the blank format. The implications of these results for L2 academic listening assessment are considered.

Keywords
academic listening test, comprehension subskills, note-taking, note-taking format, structural equation modeling

Taking notes while listening to a lecture is an important strategy that is widely used for increasing attention to the lecture and retention of its content (Dunkel & Davy, 1989). Since note-taking plays such crucial roles in lecture listening, not allowing note-taking during lectures is likely to make most students feel uncomfortable, especially when they are not familiar with the content of lectures (Carrier, 1983). In light of the importance of note-taking for academic success, a large number of studies have examined L1 college students’ note-taking practice (e.g. Meter, Yokoi, & Pressley, 1994; Trafton & Trickett, 2001) and the role that L2 college students’ notes play in lecture learning (e.g. Chaudron, Loschky, & Cook, 1994; Dunkel & Davy, 1989; Dunkel, Mishra, & Berliner, 1989).

Corresponding author:
Min-Young Song, Jeongdong Building, 15-5, Jeong-dong, Jung-gu, Seoul, Korea, 100-784.
Email: mysong88@kice.re.kr
Given that most academic listening tests allow test takers to take notes as a between-stage in answering listening test items (e.g. TOEFL, IELTS) or use note-taking itself as a measure of listening ability (e.g. the Occupational English Test), it has been considered important to explore the relationship between L2 learners’ note-taking and their subsequent listening test performance (e.g. Carrell, 2007; Carrell, Dunkel, & Mollaun, 2004; Hale & Courtney, 1994). The current study is premised on the notion that examining both the process of L2 test takers’ note-taking and the notes themselves may be one of the best ways to investigate these learners’ academic listening proficiency in the target language, since the note-taking occurs either simultaneously with or immediately after the listening activity.

From a theoretical perspective, a number of researchers into L2 academic lecture comprehension have investigated the specific skills that are necessary for effective lecture listening. Although what constitutes the content of academic listening and what levels of performance are necessary for academic success are controversial (Fulcher, 1999), students are supposed to process different levels of information presented in a lecture, and thus it is common for test developers to define the construct of academic listening in terms of different comprehension subskills or levels of understanding, such as ‘understanding the major ideas’, ‘understanding specific details’, and ‘making inferences’ (e.g. Alderson, 2000; Buck, 2001; Song, 2008a, b). Accordingly, in investigating second language learners’ note-taking, it is also common to evaluate the quality of notes in terms of the total numbers of different levels of information they contain, such as main ideas, supporting ideas, and specific details (e.g. Austin, Lee, & Carr, 2004; Clerehan, 1995; Kiewra & Fletcher, 1984). Furthermore, given that previous research (e.g. Dunkel & Davy, 1989; Dunkel et al., 1989) has shown that the quality of L2 learners’ notes is significantly affected by their target language listening proficiency, it seems reasonable to expect that these different indices of L2 note-taking quality may be linked to performance on the listening subskills that particular test items are designed to measure. To date, however, the potential link between the quality of L2 test takers’ notes measured in terms of different levels of information and their performance on different listening subskills has not been explored, although several studies have investigated the relationship between the overall quality of L2 learners’ notes and their performance on different types of listening tasks, such as recall, multiple-choice, or gap-fill (e.g. Chaudron, Loschky, & Cook, 1994; Dunkel et al., 1989; Tsai, 2004).

From a practical perspective, concerns for the relationship between lecture listening and note-taking have led test developers to consider what type of note-taking format should be adopted in order to elicit test takers’ best performance in designing academic listening tests. A case in point is the English as a Second Language Placement Exam (ESLPE) developed at the University of California in Los Angeles (UCLA) for the purpose of placing incoming international students into ESL courses. Since a computer-based version of the test was developed to replace the previous paper & pencil version in 2006, blank note-taking sheets have been provided for test takers in lieu of the partially outlined note-taking sheets that were previously provided. This change of format was mainly based on the assumption that note-taking without the aid of an outline is more authentic, that is, more similar to students’ real life practice of note-taking in a lecture. However, this decision appears not to be well grounded theoretically or
empirically. According to previous research, a lecture based on teachers’ guided notes or handouts (sometimes accompanied by visual aids such as PowerPoint slides) is one of the most common modes of presentation (e.g. DeCarrico & Nattinger, 1988; Neef, McCord, & Ferreri, 2006; Sweeney et al., 1999). Also, many researchers have advocated the use of guided notes to foster students’ quality notes and better performance on post-lecture tests (e.g. Austin, Lee, Thibeault, Carr, & Bailey, 2002; Hamilton, Seibert, Gardner, & Talbert, 2000), and the majority of students prefer partial handouts or skeletal outlines to other note-taking formats (e.g. Narjaikaew, Emarat, & Cowie, 2009; Russell, Caris, Harris, & Hendricson, 1983). It cannot, therefore, be assumed that a blank format is more common than an outline format in real life lectures. Furthermore, an outline format might work better than a blank format as a learning facilitator, although most research on this issue to date has been conducted with L1 learners. With a better understanding of the effect of the note-taking format on L2 learners’ note-taking, it would be possible to make a more informed decision on the note-taking format to be adopted in L2 academic listening tests.

To address the theoretical and practical issues discussed so far, the current study investigated the structural relationship among the quality indices of L2 test takers’ notes measured in terms of different levels of information and their performance on open-ended listening items which tap into different listening subskills. Furthermore, to investigate the effect of the note-taking format on that structural relationship, this study examined the invariance of the structural relationship across two note-taking formats, that is, a blank format and an outline format, by employing a multi-group structural equation modeling (SEM) approach.

The research questions addressed in this study are as follows:

1. To what extent is the quality of test takers’ notes and their test performance affected by their second language academic listening proficiency? Does this relationship vary depending on the note-taking format?
2. To what extent is the quality of second language test takers’ test performance related to their performance on the subsequent listening test? Does this relationship vary depending on the note-taking format?

**Previous studies**

**The effect of the note-taking format**

With regard to the effect of the note-taking format, most research conducted with L1 students demonstrated that note-taking in an outline format can be a more effective strategy in lecture learning than other note-taking formats. Some studies used guided or partial notes, in which instructors provide advance organizers, such as skeletal outlines or matrices, in which students could record key points (e.g. Austin et al., 2004; Barbetta & Skaruppa, 1995; Neef, McCord, & Ferreri, 2006). In Annis’ (1981) study, partial notes were better than personal notes and instructors’ full notes in promoting performance on multiple-choice tests. Similarly, Kiewra, Benton, Kim, Risch, and Christensen (1995) found that note-taking on an outline framework increased the completeness of students’
notes and fostered more internal connections among ideas. Sweeney et al. (1999) compared the notes taken with and without the use of guided notes and found that students correctly recorded more concepts and got better quiz scores under guided notes conditions. In Narjaikaew, Emarat, and Cowie’s recent study (2009), students involved in the guided note-taking approach performed better on a conceptual test than students who were not involved in that approach. From interviews with the participants, the researchers found that students viewed the guided note-taking approach as a supportive tool that helped them concentrate on the lecture. This is consistent with Armbruster’s (2000) statement that outlines ‘serve as advance organizers, focus student attention, provide guides for note-taking and give retrieval cues’ (p. 194).

The literature reviewed so far suggests that note-taking in an outline format may have a facilitating effect on the quality of students’ notes and their performance on the following comprehension tasks. In a study of the contents of L2 test takers’ notes, however, Cushing (1991) reported that high proficiency learners made better use of the given outlines than did low proficiency learners, and that 40% of the low proficiency students had written wrong or incomplete information in the blanks of the outlines. This suggests that the effect of the outline format might interact with L2 learners’ target language proficiency.

Criteria for evaluating the quality of notes

Researchers of effective note-taking have suggested a range of different indices for note quality, including number of information units in notes (Dunkel, 1988; Faraco, Barbier, & Piolat, 2002), efficiency of notes defined as the ratio between the number of information units to the number of words in the notes (e.g. Aiken, Thomas, & Shennum, 1975; Kiewra, 1984), and test answerability defined as the extent to which test related information is included in notes (e.g. Dunkel, 1988; Tsai, 2004), although the literature is mixed on the efficacy of those features (e.g. Aiken et al., 1975; Kiewra, 1984). Also, some researchers have paid attention to the levels of information, that is, the hierarchical propositional structure of notes (e.g. Kiewra, Benton, & Lewis, 1987; Clerehan, 1995; Tsai, 2004). Kiewra et al. (1987) categorized lecture points in four levels, that is, Level 1 main points and Levels 2 through 4 subordinate points, and reported that the records of Level 1 points were not predictive of test performance, while the records of Levels 2 and 3 intermediate level points have predictive value for immediate test performance, and Level 4 points were related to scores on the final course examination. Thus, it seems that the value of detailed notes increases with greater time between note-taking and assessment.

Finally, effective notes are characterized by clear organization, that is, hierarchical delineation between main and subordinate ideas (William & Eggert, 2002). Cushing (1991) reports that low proficiency ESL students did not distinguish between relevant and irrelevant information, indicating that such learners may have difficulty in organizing their notes in order to show the hierarchical relationship among the ideas. Also, Kiewra et al. (1995) suggest that notes in an outline format, that is, an organized format, may be positively correlated with test performance. Similarly, Tsai (2004) found that the organization of notes is one of the significant predictors of L2 test takers’ overall listening performance.

Based on the discussion so far, it seems that the literature supports the use of levels of information and organization of notes as good indicators of note quality. Accordingly, the
current study selected three measures of levels of information, that is, main/topical ideas, supporting details, and minor details, and organization of notes as criteria for measuring the quality of test takers’ notes.

**Method**

**Participants**

The participants in this study were 257 test takers who took the English as a Second Language Placement Exam (ESLPE) at UCLA in Fall, 2006, and in Winter and Spring, 2007. The majority of the participants were incoming international graduate students at UCLA, and accordingly assumed to have reached a certain level of academic English proficiency. Approximately 20% of them were enrolled in the extension program at UCLA and generally considered lower level English learners than UCLA graduate students. The majority of the participants were from Asian countries, including China, Japan, Korea, India, and some southwestern Asian countries, and the number of Chinese was the largest for a single country. As for the participants’ academic backgrounds, approximately half of the participants were from the departments of science or engineering, and the other half from the departments of liberal arts or humanities.

**Materials**

**Listening test.** The current listening section of the ESLPE was developed in two versions using two different lectures that are considered to be comparable to each other in terms of the type, length, organization, and academic field. Each test includes a video clip of approximately eight minutes recorded from an introductory level authentic lecture in the department of Sociology, with 18 open-ended questions tapping into three kinds of comprehension subskills, that is, ‘understanding main and topical ideas’, ‘understanding supporting and specific details’, and ‘making inferences from explicitly stated information’. The topic of the lecture for the version used for the current study was one believed to be unfamiliar to most international students, namely: how and why the number of interest groups increased in the USA during the 1940s and 1950s. Test takers take the test at individual stations in a computer laboratory. They listen to and watch the video clip only once, and are informed in advance that they would be wise to take notes while listening as the input is a long stretch of talk. After listening to the entire input, test takers can take the time to complete or review their notes before clicking the NEXT button, after which they move on to answering the test questions. Test takers are not allowed to return to previous questions after entering their answers. In this way, test takers can control the time taken for each question within the time limit for the entire listening section, that is, 30 minutes. Five of the 18 items were scored for partial credit and the others were dichotomously scored, and the maximum possible score was 27 points.

**Test takers’ notes.** Before taking the listening test, all test takers were provided with a sheet of note-taking paper. Among the 257 participants of the current study, 129 participants
were provided with blank sheets of paper and 128 participants were provided with sheets of paper containing skeletal linear outlines of the listening passage. (See Appendix.) All the participants were encouraged, but not forced, to take notes on the provided sheets while listening to the input.

Creating dependent variables

Hierarchical structure of the lecture. In order to categorize the listening items and obtain the scoring rubric for note quality, the hierarchical structure of the lecture used in this study was analyzed. As illustrated in Figure 1, the information in the listening passage was decomposed into units and classified into five levels of information. That is, all the information units in the listening passage were categorized as Level 1 main idea, Level 2 major topical ideas, Level 3 subtopic ideas, Level 4 supporting details, and Level 5 minor details. The current lecture contained one Level 1 main idea, two Level 2 major topics, two Level 3 subtopics supporting Major topic 1, four subtopics supporting Major topic 2, and many Level 4 and 5 supporting and minor details elaborating each subtopic, though not all illustrated here due to space constraints.

Categorizing and combining the listening items. On the basis of a content analysis, the 18 open-ended listening items were classified into three subskill categories as follows:

- **LTOP**: understanding main/topical ideas;
- **LDET**: understanding supporting and minor details;
- **LINF**: making inferences based on the explicitly stated information.

![Figure 1. Textual organization and levels of information](image-url)
Those items that directly elicit Level 1 to 3 topical ideas as defined in Figure 1 were
categorized as LTOP. For example, the question, ‘What four types of cheap communi-
cation helped in the formation of interest groups?’, directly asks about Subtopics 21, 22,
23, and 24 in Figure 1. Likewise, those items that elicit Level 4 and 5 ideas were catego-
rized as LDET. Finally, those items that require test takers to make inferences from the
given information were categorized as LINF. For example, one of the questions classified in
this category is ‘Based on the lecturer’s response to the student, what do you think
the student said?’ Since the student’s words are inaudible, the question requires test tak-
ers to reach the right answer indirectly by figuring out the interconnectedness between
the ideas and making inferences based on this. According to these categories, five items
were designated as LTOP, 10 items as LDET, and three items as LINF.

To obtain observed variables that could be considered continuous for SEM analyses,
item parcels were created by combining two or three items in the same category. In order
minimize potential pitfalls of item parceling, it was desirable to combine items eliciting
information found in the same sentence or paragraph. In addition, items that were
comparatively highly correlated with each other were combined. In this way, eight listen-
ing variables suitable for SEM analyses were obtained (see Table 1).

Quantifying the quality of test takers’ notes. Each test taker’s notes were scored in terms of
the following four criteria;

- **NTOP**: the number of Levels 1 to 3 ideas recorded;
- **NDET**: the number of Level 4 details recorded;
- **NMINOR**: the number of Level 5 or lower level minor ideas recorded;
- **NORG**: well organized (2), partially organized (1), poorly organized (0).

The last index for note quality, NORG, is defined as the extent to which the notes
were recorded in an organized fashion. If all Level 1 to 3 ideas were recorded in a
hierarchical form and Level 4 and 5 details were placed at the right levels in a test
maker’s notes, then the notes were considered well organized (refer to Appendix). If
some Level 1 to 3 ideas were missing, but Level 4 and 5 details recorded were
placed at the right levels, the notes were considered partially organized. Lastly, if
Level 4 and 5 details were recorded following the sequence of the lecture with little
evidence of any hierarchical arrangement of Level 1, 2, and 3 ideas, the notes were
considered poorly organized. Based on these criteria, each test taker’s notes were
scored from 0 to 2 points.

All participants’ notes were scored twice by the researcher in terms of the four note
quality indices. To make these scorings independent from one another, a set of hard copies
were prepared for the original notes and scoring was done on each note sheet in random
order. In this way, four note quality variables were created (see Table 1).

Analyses

To address the research questions of the current study, a multi-group structural equation
model analysis was conducted as follows. First, in order to establish a baseline model for
multi-group analysis, several structural equation models were specified as tentative baseline models and tested for each note-taking group separately. Next, the baseline model selected was tested with both note-taking groups simultaneously to examine cross-group equivalence by imposing between-group equality constraints on the parameters to be tested for invariance.

Establishing tentative baseline models. The models for the current study are premised on the findings of previous research (e.g. Cushing, 1991; Dunkel & Davy, 1989; Tsai, 2004) suggesting that L2 learners’ note-taking may be affected by their target language listening proficiency, since they take notes only of what they understand from the listening input. Thus, it can be hypothesized that the criteria for evaluating note quality, if adequately selected, can be taken as indicators of L2 learners’ target language listening proficiency.

It may be true that regardless of target language listening proficiency, test takers are selective in taking notes depending on their own note-taking styles. That is, some highly proficient listeners might not record much and as a result, produce less complete notes, while other less proficient listeners might write down as many idea units as they can. However, this is not a major concern for this study for the following reasons. First, a pilot study, conducted in an official ESLPE administration session prior to the current study, suggested that in a serious testing situation, test takers are likely to take as many notes on what they consider necessary for the subsequent listening tasks as possible. It seems that most ESLPE test takers take the test quite seriously, since those who fail the test are required to enroll in a series of ESL courses, which might cause them conflicts in schedule and extra financial burden. Moreover, as mentioned earlier, ELSPE test takers are strongly encouraged to take notes while listening to the lecture. Thus, it is expected that their attitudes towards note-taking would be more serious in a testing situation than in a real life lecture. Next, as suggested by previous research (e.g. Meter et al., 1994; Brobst, 1996), selective note-taking is more likely when lecture content is familiar and this, as already noted, is unlikely to be the case with the test version chosen for the current study. Lastly, individual idiosyncrasies in note-taking style were minimized in the current study by excluding from the analysis any incomprehensible notations that could be meaningful only to the test takers themselves (e.g. Meter et al., 1994). For these reasons, it may be assumed that the quality of test takers’ notes can serve as a good indicator of their second language listening proficiency.

Based on the considerations discussed so far, the current study hypothesized a series of structural equation models that included test takers’ scores on the selected note quality indices as well as the open-ended listening items as indicators of the underlying second language listening proficiency. In addition, considering that test takers’ note-taking is also affected by some note-taking-specific factors, including their idiosyncratic note-taking strategies, just as test takers’ performance on open-ended items is affected by open-ended method-specific factors, two types of models, one with and one without method factors were compared in the testing of the model.

This study also compared unitary listening skill models and divisible subskill models in consideration of the debates as to whether separable comprehension subskills exist or not (e.g. Buck & Tatsuoka, 1998; Rost, 1993; Song, 2008b). Unitary skill models
hypothesize that L2 listening is a unitary and integrated skill, while divisible subskill models hypothesize that several divisible subskills exist in L2 listening comprehension and these subskills are affected by a higher order second language listening proficiency. Based on these hypotheses, the current study tested four SEM models, that is, the one-skill model and the three-subskill model which were hypothesized under two different conditions, that is, one without a method factor included and the other including two method factors. Of the four models, only those including two method factors are presented in Figures 2 and 3, owing to space constraints.

Given the divisibility of listening subskills, it may be hypothesized that each note quality index is an indicator of its corresponding subskill factor. That is, the number of main/topical ideas in test takers’ notes is an indicator of their ability to understand main/topical ideas, and the number of supporting and minor details is an indicator of the ability to understand details. Regarding the organization of notes, there is no directly corresponding

Figure 2. One-skill model (with method factors)
subskill factor in the model. In order to answer the questions categorized as inference items in the study, however, it is more crucial for test takers to understand the interconnectedness among the ideas, rather than understand each topical idea or detail. For example, one of the three inference questions used in this study was ‘According to the reasoning in the lecture, why would the number of interest groups increase if the government tried to regulate internet sites?’ This question refers to the first Level 2 idea illustrated in Figure 1, and in order to answer the question, test takers need to produce a statement which integrates both Level 3 ideas supporting this Level 2 idea using an appropriate connecting word for contrast, rather than just mention each idea separately. Therefore, it is to be expected that the organization index may be more associated with higher order comprehension subskills, such as integration or inference, rather than with factual understanding skills, including listening for topical ideas and details (Weir, 1993). Accordingly, it seems reasonable to hypothesize that test takers’ organization of notes is more related to their ability to make inferences than any other subskill concerned in this study.

**Figure 3.** Three-subskill model (with method factors)
Single group model testing. Since each observed variable could be assumed to be ‘a categorized version of an underlying truly continuous, normally distributed variable’ (Bentler, 2004, p. 8), it was decided to analyze the covariance structures of the variables rather than the correlation structures. In addition, the assumption of multivariate normality was not seriously violated for either group, since the normalized estimates of Mardia’s coefficients for the variable set were smaller than ±3 for both groups (Bentler, 2004), which indicates that it is appropriate to use the maximum likelihood (ML) estimation method for the subsequent model tests (Bentler, 2004, p. 5).

The adequacy of the hypothesized models was evaluated based on multiple criteria. For a good model fit, the $\chi^2/df$ ratio, which deals with the effects of large sample size on the $\chi^2$ statistic, should be less than 1.5, but a ratio of 2.5 or less indicates an acceptable fit (Kline, 1998). In addition, the comparative fit index (CFI) should be higher than .90 (Bentler, 1990). Finally, the root mean-square error of approximation (RMSEA), which takes into account model complexity as reflected in the degrees of freedom, should be less than .05 for a good model fit and the values from .05 to .08 are an acceptable fit (Brown & Cudeck, 1992).

Results

The data collected for the current study were analyzed using the Statistical Package for the Social Sciences for Windows Release 11.5 (SPSS, 2002) and the EQS Version 6.1 for windows (Bentler & Wu, 2002) statistical package.

Reliability and descriptive statistics

As preliminary statistical analyses, reliability coefficients and descriptive statistics were calculated. Reliability estimates for the scores obtained from each group were quite satisfactory. Considering that only 18 items were included in the listening test and there were only 28 possible score points (from 0 to 27) in total, the alpha coefficients for the listening test were considered acceptable ($\alpha = .744$ for the blank group, $\alpha = .757$ for the outline group). In addition, the intra-rater reliability coefficients for the note scores, that is, the correlations between the paired scores for each note quality measure, were considerably high, ranging from 0.854 to 0.995. As presented in Table 1, the descriptive statistics for the two note-taking groups show that overall, the outline group received slightly higher mean scores than the blank group in the note quality measures, but the result was not consistent for the open-ended listening measures. One thing to note regarding NTOP, that is, the number of Level 1 to 3 ideas recorded, is that the score differences in NTOP among the participants were mostly a result of the differences in the numbers of Level 3 ideas, since there were very minor differences in the numbers of Level 1 and 2 ideas among the participants.

Correlations among the measures

In order to address Research question 2, that is, to examine the relationships among the open-ended measures and the note quality measures, the correlations among the measures were compared. The correlation coefficients are presented in Tables 2 and 3. Overall, the associations among the measures are firmer in the outline group than in the blank
### Table 1. Descriptive statistics for each note-taking group

<table>
<thead>
<tr>
<th></th>
<th>Blank</th>
<th></th>
<th>Outline</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
</tr>
<tr>
<td>LTOP1</td>
<td>129</td>
<td>2.279</td>
<td>0.927</td>
<td>128</td>
</tr>
<tr>
<td>LTOP2</td>
<td>129</td>
<td>1.333</td>
<td>0.688</td>
<td>128</td>
</tr>
<tr>
<td>LTOP3</td>
<td>129</td>
<td>2.132</td>
<td>1.026</td>
<td>128</td>
</tr>
<tr>
<td>LDET1</td>
<td>129</td>
<td>0.767</td>
<td>0.931</td>
<td>128</td>
</tr>
<tr>
<td>LDET2</td>
<td>129</td>
<td>2.070</td>
<td>1.147</td>
<td>128</td>
</tr>
<tr>
<td>LDET3</td>
<td>129</td>
<td>1.783</td>
<td>0.952</td>
<td>128</td>
</tr>
<tr>
<td>LINF1</td>
<td>129</td>
<td>0.930</td>
<td>0.894</td>
<td>128</td>
</tr>
<tr>
<td>LINF2</td>
<td>129</td>
<td>0.860</td>
<td>0.658</td>
<td>128</td>
</tr>
<tr>
<td>NTOP</td>
<td>129</td>
<td>7.047</td>
<td>1.924</td>
<td>128</td>
</tr>
<tr>
<td>NMIN</td>
<td>129</td>
<td>5.907</td>
<td>4.818</td>
<td>128</td>
</tr>
<tr>
<td>NORG</td>
<td>129</td>
<td>1.434</td>
<td>0.705</td>
<td>128</td>
</tr>
</tbody>
</table>

### Table 2. Correlations among the listening and note quality measures (Blank group)

<table>
<thead>
<tr>
<th></th>
<th>LTOT</th>
<th>LTOP</th>
<th>LDET</th>
<th>LINF</th>
<th>NTOT</th>
<th>NTOP</th>
<th>NDET</th>
<th>NMIN</th>
<th>NORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTOT</td>
<td>1</td>
<td>.85</td>
<td>.83</td>
<td>.72</td>
<td>.36</td>
<td>.61</td>
<td>.22</td>
<td>.27</td>
<td>.57</td>
</tr>
<tr>
<td>LTOP</td>
<td></td>
<td>1</td>
<td>.50</td>
<td>.53</td>
<td>.45</td>
<td>.56</td>
<td>.36</td>
<td>.22</td>
<td>.50</td>
</tr>
<tr>
<td>LDET</td>
<td></td>
<td></td>
<td>1</td>
<td>.39</td>
<td>.07*</td>
<td>.45</td>
<td>.39</td>
<td>.33</td>
<td>.48</td>
</tr>
<tr>
<td>LINF</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.56</td>
<td>.32</td>
<td>.02*</td>
<td>.1*</td>
<td>.35</td>
</tr>
<tr>
<td>NTOT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.58</td>
<td>.94</td>
<td>.49</td>
<td>.64</td>
</tr>
<tr>
<td>NTOP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.72</td>
<td>.38</td>
<td>.29</td>
</tr>
<tr>
<td>NDET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.32</td>
<td>.31</td>
</tr>
<tr>
<td>NMIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.76</td>
</tr>
<tr>
<td>NORG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Correlation is not statistically significant.

### Table 3. Correlations among the listening and note quality measures (Outline group)

<table>
<thead>
<tr>
<th></th>
<th>LTOT</th>
<th>LTOP</th>
<th>LDET</th>
<th>LINF</th>
<th>NTOT</th>
<th>NTOP</th>
<th>NDET</th>
<th>NMIN</th>
<th>NORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTOT</td>
<td>1</td>
<td>.85</td>
<td>.87</td>
<td>.71</td>
<td>.52</td>
<td>.66</td>
<td>.44</td>
<td>.38</td>
<td>.60</td>
</tr>
<tr>
<td>LTOP</td>
<td></td>
<td>1</td>
<td>.56</td>
<td>.55</td>
<td>.56</td>
<td>.58</td>
<td>.50</td>
<td>.21</td>
<td>.50</td>
</tr>
<tr>
<td>LDET</td>
<td></td>
<td></td>
<td>1</td>
<td>.41</td>
<td>.34</td>
<td>.45</td>
<td>.50</td>
<td>.43</td>
<td>.55</td>
</tr>
<tr>
<td>LINF</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.56</td>
<td>.58</td>
<td>.58</td>
<td>.43</td>
<td>.39</td>
</tr>
<tr>
<td>NTOT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.58</td>
<td>.96</td>
<td>.45</td>
<td>.69</td>
</tr>
<tr>
<td>NTOP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.76</td>
<td>.45</td>
<td>.43</td>
</tr>
<tr>
<td>NDET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>.76</td>
<td>.30</td>
</tr>
<tr>
<td>NMIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NORG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
group, in that all the correlation coefficients are statistically significant in the outline group, while there are several insignificant coefficients in the blank group. More detailed discussion on the association patterns among the measures is provided in discussing Research question 2.

**Establishing a baseline model**

First, the one-skill model and three-subskill model were tested with each group. As indicated by the poor model fit indices in Table 4, neither of these models fit the current data well in either group. Next, the same models including two method factors were tested with each group separately. As indicated by the good model fit indices in Table 5, both the one-skill model and the three-subskill model including two method factors worked quite well for both groups.

Based only on the fit indices, either of the models could be selected as a baseline model for further analysis. Given that a baseline model represents the most parsimonious, yet substantively most meaningful and best fitting model to the data (Byrne, Shavelson, & Muthén, 1989), however, the three-subskill model was preferred to the one-skill model. As seen in Tables 2 and 3, in general, the open-ended measures and note quality measures that belong in the same subskill category are more highly correlated with each other than with the measures of the other subskill categories. That is, the association patterns among the open-ended measures and note quality measures suggest that several distinct subskill categories can be empirically identified with the measured variables of this study. Furthermore, as indicated by the fit indices in Table 5, the

---

**Table 4. Testing one skill vs. three subskill factor models with no method factor**

<table>
<thead>
<tr>
<th></th>
<th>Blank</th>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 skill</td>
<td>3 subskills</td>
</tr>
<tr>
<td>$\chi^2$(df)</td>
<td>170.113 (54)</td>
<td>149.092 (51)</td>
</tr>
<tr>
<td>$\chi^2$/df</td>
<td>3.150</td>
<td>2.923</td>
</tr>
<tr>
<td>CFI</td>
<td>0.740</td>
<td>0.781</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.130</td>
<td>0.123</td>
</tr>
</tbody>
</table>

**Table 5. Testing one skill vs. three subskill factor models with two method factors**

<table>
<thead>
<tr>
<th></th>
<th>Blank</th>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 skill</td>
<td>3 subskills</td>
</tr>
<tr>
<td>$\chi^2$(df)</td>
<td>68.011 (42)</td>
<td>62.530 (39)</td>
</tr>
<tr>
<td>$\chi^2$/differ.</td>
<td>5.481 (3), p &gt; 0.10</td>
<td>7.733 (3), p &gt; 0.050</td>
</tr>
<tr>
<td>$\chi^2$/df</td>
<td>1.619</td>
<td>1.603</td>
</tr>
<tr>
<td>CFI</td>
<td>0.942</td>
<td>0.947</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.070</td>
<td>0.069</td>
</tr>
</tbody>
</table>
three-subskill model fits slightly better than the one-skill model for both groups, though the chi-square difference tests between the two models were not significant for either group. Based on these considerations, it seems reasonable to select the three-subskill model with two method factors as the baseline model for further analysis. The standardized path coefficients for this final model are presented in Tables 6 and 7.

As shown in Table 6, some paths from the two method factors, that is, OPENEND and NOTE, were not statistically significant in either group. However, all the paths from the three subskill factors were statistically significant in both groups, which means that each observed variable is a good indicator of its associated subskill factor in both groups. As shown in Table 7, all three first-order subskill factors were also good indicators of the higher-order factor LISTEN, that is, test takers’ L2 listening proficiency, for both groups. Given all the good model fit indices and significant path coefficients from the first- and second-order trait factors, it is appropriate to proceed to test cross-group equivalence of the key parameters.

| Table 6. Standardized estimates for factor loadings in the baseline model (F → V) |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                                  | Blank                           | Outline                        |
|                                 | TOP    | DET    | INF    | OPEN   | NOTE   | TOP    | DET    | INF    | OPEN   | NOTE   |
| LTOP1                           | .46    | -      | -      | .23*   | -      | .31    | -      | -      | .14*   | -      |
| LTOP2                           | .47    | -      | -      | .20*   | -      | .30    | -      | -      | .60    | -      |
| LTOP3                           | .61    | -      | -      | .48    | -      | .59    | -      | -      | .32    | -      |
| LDET1                           | -      | .26    | -      | .15*   | -      | -      | .59    | -      | .09*   | -      |
| LDET2                           | -      | .58    | -      | .19*   | -      | -      | .53    | -      | .28    | -      |
| LDET3                           | -      | .49    | -      | .29    | -      | -      | .49    | -      | .28    | -      |
| LINF1                           | -      | -      | .31    | .31    | -      | -      | -      | .33    | .40    | -      |
| LINF2                           | -      | -      | .47    | .45    | -      | -      | -      | .46    | .47    | -      |
| NTOP                            | .84    | -      | -      | -      | .15*   | .88    | -      | -      | -      | .08*   |
| NDET                            | -      | .33    | -      | .95    | -      | .67    | -      | -      | -      | .61    |
| NMINOR                          | -      | .39    | -      | .63    | -      | .52    | -      | -      | -      | .68    |
| NORG                            | -      | -      | .79    | -      | .08    | -      | -      | .86    | -      | .11    |

Notes: The starred parameters (*) were not statistically significant. The shaded parameters were fixed to identify the model.

| Table 7. Standardized estimates for factor regression coefficients in the baseline model (F → F) |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                                  | Blank                           | Outline                        |
|                                 | LISTEN                          | LISTEN                          |
| TOP                              | .993                            | .990                            |
| DET                              | .959                            | .922                            |
| INF                              | .943                            | .923                            |
Table 8. Standardized estimates for the total effects of the second-order factor (LISTEN)

<table>
<thead>
<tr>
<th>Observed variables</th>
<th>Blank LISTEN</th>
<th>Outline LISTEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTOP1</td>
<td>.454</td>
<td>.308</td>
</tr>
<tr>
<td>LTOP2</td>
<td>.468</td>
<td>.293</td>
</tr>
<tr>
<td>LTOP3</td>
<td>.603</td>
<td>.583</td>
</tr>
<tr>
<td>LDET1</td>
<td>.248</td>
<td>.544</td>
</tr>
<tr>
<td>LDET2</td>
<td>.557</td>
<td>.485</td>
</tr>
<tr>
<td>LDET3</td>
<td>.468</td>
<td>.455</td>
</tr>
<tr>
<td>LINF1</td>
<td>.294</td>
<td>.306</td>
</tr>
<tr>
<td>LINF2</td>
<td>.444</td>
<td>.424</td>
</tr>
<tr>
<td>NTOP</td>
<td>.834</td>
<td>.873</td>
</tr>
<tr>
<td>NDET</td>
<td>.315</td>
<td>.619</td>
</tr>
<tr>
<td>NMINOR</td>
<td>.375</td>
<td>.479</td>
</tr>
<tr>
<td>NORG</td>
<td>.747</td>
<td>.798</td>
</tr>
</tbody>
</table>

Table 8 shows the standardized estimates for the total effect of the second-order factor, that is, LISTEN, on each observed variable. Each of these estimates was obtained by multiplying the first- and second-order factor loading estimates associated with each observed variable, that is, an F→V path coefficient multiplied by the associated F→F path coefficient. Generally, not only are the note quality variables good indicators of LISTEN, that is, test takers’ L2 listening proficiency, especially in the outline group, but some of them are also better indicators of test takers’ L2 listening proficiency than the open-ended listening variables in both groups. More specifically, the note topic measure (NTOP) and note organization (NORG) are the two best indicators of LISTEN in both groups. That is, although the open-ended variables were generally good measures of test takers’ L2 listening proficiency, the note quality variables, especially, the note topic measure (NTOP) and the note organization (NORG) did a better job in representing test takers’ L2 listening proficiency than any of the open-ended listening variables did.

**Multi-group analyses for the baseline model**

Although the three-subskill model with two method factors worked quite well across the groups, the hypothesis of a single set of parameter estimates for that baseline model still remained to be tested simultaneously in a single run for both groups. This study concerns the invariance of two sets of key parameters across groups, that is, equal factor loadings (F→V), which indicates the measurement model invariance, and equal factor regression coefficients (F → F), which indicates the structural model invariance.

The invariance of the factor loadings (F → V) and the factor regression coefficients (F → F) across the groups was tested by constraining all first- and second-order factor loadings to be equal across groups. The results of this multi-group test showed that the hypothesis of cross-group equality of all first- and second-order factor loadings could not
be rejected. The fit indices indicated quite a good overall model fit ($\chi^2$/df = 146.171/102 = 1.433, CFI = .955, RMSEA = .058), and the Lagrange Multiplier (LM) test showed that there was no significant constraint in the univariate tests and none of the constraint increments were significant in the multivariate test, either (p > .05), indicating that all the equality constraints were reasonable. Thus, it was concluded that all the factor loadings were invariant across the two groups. In other words, the relationship between the open-ended variables, note quality variables, and the underlying trait factors were invariant across the two groups.

**Discussion**

*The effect of L2 listening proficiency on note quality and test performance*

The structural equation model selected in the current study suggests that test takers’ second language listening proficiency indirectly affects their performance on the open-ended listening items and note quality measures mediated by the three listening subskills. With regard to the effects of the second order factor, LISTEN, that is, test takers’ second language listening proficiency, on the open-ended and note quality measures, the note quality measures are equally as good indicators of second language listening proficiency as the open-ended measures are, as indicated in Table 8. This is compatible with the indications from previous research that the quality of L2 learners’ notes is significantly affected by their target language listening proficiency (e.g. Dunkel & Davy, 1989; Dunkel et al., 1989). In particular, note topics (NTOP) and note organization (NORG) are the best two indicators of LISTEN regardless of the note-taking format. That is, test takers’ note-taking of topical ideas and the organization of their notes are more strongly affected by their second language listening proficiency than is their performance on any other open-ended or note quality measure.

As seen in Table 6, generally speaking, each note quality variable is the best indicator of its associated subskill factor. That is, NTOP (no. of topical ideas in notes) is the best indicator of the subskill, ‘understanding topical ideas (TOP)’, and NORG (organization of notes) is the best indicator of the subskill, ‘making inferences (INF)’, across the note-taking formats. As for the subskill ‘understanding details (DET)’, NDET (no. of details in notes) is the best indicator in the outline format, while LDET2, an open-ended detail measure, is the best indicator in the blank format. In particular, NTOP and NORG are both optimal indicators of TOP and INF, considering the large factor loading estimates from the subskill factors.

Compared with such strong associations of NTOP and NORG with TOP and INF, the other note quality variables, NDET (no. of Level 4 details in notes) and NMINOR (no. of Level 5 or lower level minor ideas in notes), seem to have slightly weaker associations with the subskill factor, DET, and stronger associations with the method factor, NOTE. This suggests that compared to test takers’ note-taking of topical ideas and the organization of their notes, their note-taking of details might involve more than just an ability to understand details. As discussed earlier, it is likely that test takers’ willingness to record details may be partly a matter of personal preference as well as being inhibited in some cases by other factors such as slow handwriting.
The results reported so far may be discussed as follows. First, comparing open-ended questions and note-taking as different types of listening tasks, an open-ended task limits test takers’ demonstration of their comprehension by eliciting responses to a set of selected items, while a note-taking task allows test takers to demonstrate whatever they understand, although they might be selective in recording this. Therefore, it could be argued that a note-taking task is a better indicator of their listening proficiency than an open-ended task, as long as test takers try to record whatever they consider important to complete subsequent test tasks, as may be the case in a serious testing situation. Test takers’ notes of topical ideas and organization of their notes may be especially good indicators of their academic listening proficiency given the tendency observed in Song’s (2008a) study for test takers to take notes of as many topical ideas as they can. On the other hand, test takers’ note-taking of details may depend on their individual note-taking styles as well as on their listening proficiency, and this is corroborated by the large standard deviation values for the note detail variables, NDET and NMINOR, as seen in Table 1. In short, it appears that test takers’ notes of topical ideas and the organization of their notes may reflect their L2 listening proficiency better than their notes of details do.

Comparing the two note-taking formats, the multi-group structural equation model analysis indicated that the note-taking format did not cause significant differences in the structural relationships among the measured variables and underlying factors. Although the current study did not detect statistically significant differences between the formats, the differences in the factor loading estimates of the note quality measures, especially the two detail measures, NDET and NMINOR, are quite noticeable between the formats. Also, as mentioned earlier, NDET is the best indicator of the subskill factor, DET, in the outline format, while this is not the case in the blank format. In other words, test takers’ note-taking of details seems to have slightly stronger association with their listening proficiency in the outline format than in the blank format.

According to Song’s (2008a) study, the outline format seems to facilitate or induce high level test takers to take more complete notes of every level of information and organize their notes more clearly. On the other hand, low level test takers may have difficulty deciding which level to assign information to in the outline format and, therefore, be reluctant to record details. In contrast, in the blank format, without the restrictions imposed by an outline, they might be more inclined to take notes of whatever they want to, including minor details, even if they do not clearly understand the interconnectedness among the ideas. Perhaps for this reason, the outline format seems to produce a more noticeable difference in the quality of high and low level test takers’ notes, especially with respect to details. That may be one reason why the note quality measures, especially the detail measures in this study, have slightly stronger associations with test takers’ L2 listening proficiency in the outline format than in the blank format.

The relationship between the quality of notes and test performance

The following discussion is based on the correlations among the open-ended and note quality variables as summarized in Tables 2 and 3, since the structural equation models selected in this study do not address the relationships among the observed variables directly.
Overall, each note quality variable is significantly correlated with the open-ended measures across the note-taking formats, except for a few cases in the blank format. More specifically, the note topics (NTOP) and the note organization (NORG) are quite highly correlated with all the open-ended listening measures, not just with their open-ended counterparts, LTOP or LINF. As previously discussed, the quality of test takers’ notes of topical ideas and the organization of their notes are the best two indicators of test takers’ second language listening proficiency. In other words, those who are able to record many topical ideas and organize their notes in a structured way are very likely to be highly proficient second language listeners and, accordingly, are expected to perform well on all types of subskill items. Furthermore, considering that the topical ideas defined in this study include Level 1 to 3 ideas, the result of the current study is not incompatible with the result of Kiewra et al.’s (1987) finding that the records of the intermediate level points have predictive value for immediate test performance.

Unlike NTOP and NORG, the detail measures of notes, that is, NDET and NMINOR, are highly correlated only with the open-ended detail measure, LDET, and not with the other open-ended measures. That is, the quality of L2 test takers’ notes of details is closely associated with their performance on items measuring ability to understand details, while it is not related to their performance on other types of subskill items and, accordingly, does not indicate their listening proficiency reliably. As pointed out by previous researchers (e.g. Slotte & Lonka, 1999; Song, 2008a), test takers are very likely to review their notes during a test to gather the necessary information to respond to given items and note-taking as a memory/storage aid may work better for details than topical ideas. Thus, test takers might benefit from their notes of details even if they do not understand the interconnectedness among the ideas clearly.

The open-ended detail variable (LDET) is highly correlated not only with the detail measures of notes, that is, NDET and NMINOR, but also with the other note quality measures, across the note-taking formats. Considering that the points assigned to the detail items cover approximately 52% of the possible maximum score of the ESLPE listening test (14 out of 27 points), it is expected that those who get high scores on the detail items are very likely to get high total test scores. Therefore, the high correlations between LDET and all the note quality measures indicate that test takers’ overall performance on the listening test and the quality of their notes are highly correlated with each other, which offers support for the practice of using note quality indices as listening measures.

The open-ended inference measure (LINF) is more highly correlated with the note topic variable (NTOP) and note organization variable (NORG) than with the note detail variables (NDET and NMINOR) across the note-taking formats, and this is more noticeable in the blank format. This lends support to the assumption underlying the structural equation models tested in this study, namely that there would be an association between test takers’ performance on inference items and their ability to organize their notes. Moreover, the strong association of LINF with NTOP may be understood considering that the right answers for the three inference items used in this study were, coincidentally, all topical ideas from the given lecture. However, it is not clear whether this is only the case for the ESLPE inference items or may be true of other inference items in general. Therefore, the association between LINF and NTOP should be the subject of further empirical research.
Finally, comparing the two note-taking formats, the correlations among the open-ended variables and note quality variables are slightly higher in the outline group than in the blank group. As seen in Tables 2 and 3, all the coefficients were significantly different from zero in the outline format, while there were several insignificant correlation coefficients in the blank group. In particular, the note detail measures, that is, NDET and NMINOR, were not highly correlated with the open-ended listening measures in the blank group. This suggests that the quality of test takers’ notes of details obtained in the blank format cannot be relied upon as a predictor of their performance on the subsequent listening test. As previously discussed, test takers might record in the blank format whatever they listen to even if they do not clearly understand the organizational connections among the ideas, while they are less likely to do so in the outline format. Thus, they might not always be able to relate their notes in the blank format to the associated test items.

**Conclusions and implications**

Based on the discussions so far, the following conclusions and implications are drawn. First, the study has shown that it is possible to identify what indices of note quality or levels of information in notes are more reliable indicators of L2 learners’ academic listening proficiency. The number of topical ideas recorded, especially Level 3, that is, intermediate level ideas, and the organization of notes may be better indicators of L2 test takers’ listening proficiency than the extent to which test takers supply other levels of information. On the other hand, the numbers of lower levels of information found in test takers’ notes might not be very reliable indicators of their listening proficiency, possibly because recording details are not only dependent on test takers’ listening proficiency but also other factors including their note-taking styles.

Second, the findings of this study suggest that note quality indices could be regarded as equally good indicators of test takers’ second language academic listening proficiency as other kinds of listening measures. Thus, test developers might want to consider adopting note-taking as a listening test task, as is already the case of some listening tests, such as the listening sub-test of the Occupational English Test (McNamara, 1996) and the table completion task of IELTS. As for the format for note-taking, it would seem that notes taken in the outline format in particular, because it constrains the contents of test takers’ notes to a given framework, might have more potential as a listening measure than notes in the blank format. Also, it seems advisable based on this study’s results to adopt a note-taking task focusing on topical ideas rather than details as a measure of L2 academic listening proficiency. For example, completing partially filled outlines that focus on topical ideas could be a practicable note-taking task even for a large-scale listening test. Ideally, L2 test takers’ academic listening proficiency could be more accurately measured by employing both a note-taking task as a measure of their abilities to understand the topical ideas and organizational structure of a lecture and other types of traditional listening tasks as measures of other listening subskills.

As discussed in the previous section, there are some suggestions that the outline format might work differently depending on such factors as textual characteristics of listening inputs or test takers’ L2 listening proficiency. Therefore, it may be worthwhile to investigate the effect of the note-taking format using lectures which have different textual characteristics.
from the one used in this study. Finally, it may also be worth comparing the effect of the note-taking format on test takers who are at different L2 proficiency levels. In particular, the extent to which the outline format works differently for different proficiency groups should be investigated in follow-up studies by employing both quantitative and qualitative approaches.

Note
1. Bachman and Palmer (1996) state that ‘we (test developers) should attempt to design our tests to elicit test takers’ best performance’ (p. 66). Swain (1985) refers to this principle as ‘bias for the best’.

References


Appendix: An example of test takers’ notes in the outline format