Holistic and Analytic Assessments of the TOEFL iBT® Integrated Writing Task

Masumi ONO
Keio University

Hiroyuki YAMANISHI
Chuo University

Yuko HIJIKATA
University of Tsukuba

Abstract
Integrated writing tasks are becoming popular in the field of language testing, but it remains unclear how teachers assess integrated writing tasks holistically and/or analytically and which is more effective. This exploratory study aims to investigate teacher-raters’ holistic and analytic ratings for reliability and validity and to reveal their perceptions of grading the integrated writing task on the Test of English as a Foreign Language Internet-based Test (TOEFL iBT). Thirty-six university students completed a reading-listening-writing task. Seven raters scored the 36 compositions using both a holistic and an analytic scale, and completed a questionnaire about their perceptions of the scales. Results indicated that the holistic and analytic scales exhibited high inter-rater reliability and there were high correlations between the two rating methods. In analytic scoring, which contained four dimensions, namely, content, organization, language use, and verbatim source use, the dimensions of content and organization were highly correlated to the overall analytic score (i.e., the mean score of the four dimensions). However, the dimension of verbatim source use was found to be peculiar in terms of construct validity for the analytic scale. The analyses also indicated various challenges the raters faced while scoring. Their perceptions varied particularly regarding verbatim source use: Some raters tended to emphasize the intricate process of textual borrowing while others stressed the difficulty in judging multiple types and degrees of textual borrowing. Pedagogical implications for the selection and use of rubrics as well as the teaching and assessment of source text use are suggested.

Keywords: integrated writing, reading-listening-writing task, holistic assessment, analytic assessment, rater perception

Integrated writing tasks have increasingly been receiving attention from researchers and educators in the fields of second language (L2) writing and language testing. They are
regarded as a valid and authentic means of writing assessment (Plakans, 2015; Weigle, 2004) and are widely used in English proficiency tests (e.g., Plakans, Gebril, & Bilki, 2016; Shin & Ewert, 2015). In educational settings, the integration of different language skills, such as integration of reading and listening in writing tasks, has become a key feature of the assessment of academic writing abilities (Sawaki, Quinlan, & Lee, 2013; Yang & Plakans, 2012). The Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, where this study is situated, has encouraged English language teachers at the secondary level education to teach integrated language skills (MEXT, 2018). This has also been emphasized at the tertiary level because many academic writing assignments require such skills and related source-use strategies and integrated writing skills serve a vital role in achieving academic success.

Among various types of task, an integrated reading-listening-writing task for the Test of English as a Foreign Language Internet-based Test (TOEFL iBT) has been developed by the Educational Testing Service (ETS). This task requires test-takers to read a passage, listen to a lecture on the same topic, and summarize the main points of both by clarifying the relationship between them. Integrated writing tasks have been researched, yet only a few studies have investigated the assessment of integrated writing tasks (e.g., Ohta, Plakans, & Gebril, 2018; Plakans, 2015; Shin & Ewert, 2015) or raters’ perceptions of the tasks (e.g., Gebril & Plakans, 2014; Wang, Engelhard, Raczyński, Song, & Wolfe, 2017). On writing assessment in general, Schoonen (2005) states that “[t]he assessment of writing ability is notoriously difficult,” and that “even the way in which … traits are scored (e.g., holistically or analytically) may contribute to the writer’s score” (p. 1). Assessment of an integrated writing task is even more challenging because raters must pay attention to multiple dimensions of the task, including text comprehension, connection of sub-skills, and production. Despite this complexity and difficulty, English language teachers need to teach and assess integrated writing tasks. The question, then, is whether and how they can effectively assess students’ integrated writing tasks holistically and analytically. Thus, this exploratory study set out to investigate these teacher-raters’ holistic and analytic scoring results and their perceptions of scoring using the TOEFL iBT integrated writing task in a classroom.

**Literature Review**

**Characteristics of Integrated Writing Tasks**

The integrated writing task can be viewed as a task that combines various language skills (e.g., listening, reading) with writing, where information from source materials is incorporated into one’s written response. Integrated writing tasks are divided into three types (see Weigle & Parker, 2012, for details): (a) text-based writing, including a summary writing task comparing/contrasting the information in one or more source texts (e.g., Asención Delaney, 2008; Yu, 2013); (b) situation-based writing, requiring test-takers to write responses to emails or letters (e.g., Test of English for International Communication by ETS); and (c)
thematically related writing, requiring test-takers to write an opinion or argumentative essay based on source texts and their own knowledge or experience (e.g., Asención Delaney, 2008; Plakans, 2009a; Shin & Ewert, 2015). One popular task of the third type is the reading-to-write task, where test-takers read one or more source texts and write a response essay. Depending on the purpose of the task, the genre of essays can vary between, for instance, argumentative and persuasive. The TOEFL iBT integrated writing task, which is focused in the present study, is a text-based writing task, specifically a summary task which contains three skills (reading, listening, and writing) in interaction with two source texts—a reading passage and a spoken lecture. Test-takers must employ multiple operations, such as reading, listening, comprehending, comparing, synthesizing, and summarizing, to perform the integrated writing task (Plakans, 2009b).

The construct of integrated writing tasks can vary depending on the task type and skills involved. Sawaki et al. (2013) examined the factor structures of responses to the TOEFL iBT integrated writing task as well as scores on independent listening and reading tests. They found that the three skills involved in the integrated writing task (reading, listening, and writing) were correlated constructs. Moreover, the comprehension factor was correlated with two writing subskills, namely, productive vocabulary and sentence conventions. Asención Delaney (2008) also investigated the construct of reading-to-write ability, comparing a summary task and a response essay task. The results indicated that the two tasks had different constructs; the response essay was found to require more critical thinking than the summary task, and furthermore, the reading-to-write ability was weakly correlated with reading ability and was not associated with writing ability. In contrast, Shin and Ewert (2015) found that reading-to-write scores were moderately correlated with both reading score and writing score. Thus, there is inconsistency in this aspect of integrated writing in the previous findings.

Another central construct of the integrated writing task is source text use (e.g., Hyland, 2009), which differs from the construct of the independent writing task, that is, prompt-based writing using one’s knowledge and experience. Hyland (2009) argues that students need to understand when and how citation, quotation, and paraphrasing should be used in their writing and that assessment needs to clearly specify the appropriate textual borrowing. Weigle and Parker (2012) examined students’ source text borrowing, using an integrated reading/writing test. They found no significant differences in the amount and type of source use strategies, regardless of language proficiency. However, Plakans and Gebril (2013) had different findings about textual borrowing when they scrutinized responses on the TOEFL iBT integrated writing tests. They revealed that high-scoring responses included information from the listening material and the main ideas from the sources much more than did the low-scoring responses, which heavily relied on the reading passage and used verbatim phrases. Plakans and Gebril (2012) also investigated source text use in a reading-to-write test using a think-aloud method. The results indicated “that source use serves several functions including generating ideas about the topic and serving as a language repository” (p. 18). Thus, source use is key to completion of the integrated task, but may cause difficulty for some
test-takers.

In essence, any integrated writing task demands intricate processes focusing on multiple dimensions including the use of source texts, depending on task requirements. To date, studies have contributed to a better understanding of the complex characteristics of integrated writing tasks. However, more research needs to address how teachers should teach and assess the multifaceted characteristics of the integrated writing task in classrooms.

**Holistic and Analytic Assessment for Integrated and Independent Writing**

The assessment of integrated writing tasks is often seen as challenging for raters because it is not easy to tease apart the ways in which certain reading-specific, writing-specific, and/or cognitive abilities may affect performance. Regardless of the type of writing task, “[o]ne central aspect in the performance assessment of writing is the rating scale” or rubric (Knoch, 2009, p. 276). The most common rating scales for writing assessment are holistic and analytic (multi-trait) scales (Hamp-Lyons, 1995), developed for various integrated writing tasks for different purposes: testing with holistic scales (e.g., ETS, 2002, 2014), testing with analytic scales (e.g., Chan, Inoue, & Taylor, 2015), research with holistic scales (e.g., Plakans, 2009b), and research with analytic scales (Shin & Ewert, 2015; Yang & Plakans, 2012). More specifically, for the TOEFL iBT integrated writing task, ETS (2014) developed a holistic rating rubric on a six-point scale that has been widely used for various research purposes (e.g., Cho, Rijmen, & Novák, 2013; Knoch, Macqueen, & O’Hagan, 2014).

Focusing on the same task type, Yang and Plakans (2012) developed an analytic rubric based on the holistic rubrics of the TOEFL iBT integrated writing task and the LanguEdge Courseware scoring rubrics employed by Cumming et al. (2005). This means that the ETS (2014) holistic rubric and Yang and Plakans’ analytic rubric can be compared. Yang and Plakans’ analytic rubric has four dimensions: *content*, *organization*, *language use*, and *verbatim source use*. They are defined below, with reference to Cumming et al. (2005):

The band descriptors for *content* are accurate presentations of key points and connections to principal ideas from two sources. The band descriptors for *organization* are coherence and cohesion at the paragraph and essay levels. In terms of *language use*, grammar, mechanics, and overall understanding for the written language are criteria for evaluation. *Verbatim source use* was operationalized as a string of three words or more directly drawn from the sources into student essays. (Yang & Plakans, 2012, p. 88)

Yang and Plakans investigated L2 writers’ task strategies and revealed that high L2 writing ability did not necessarily reflect writers’ appropriate use of L2 source texts; conversely, it was found that high scores on *verbatim source use* did not always result in high scores on the other aspects of the integrated writing task. This implies that skillful source use is somewhat independent of the other subskills yet still important for successful completion of the task.

Much discussion has considered strengths and weaknesses of holistic and analytic
scoring methods using independent writing tasks (e.g., East, 2009; Hamp-Lyons, 2016; Knoch, 2009; Nakamura, 2004; Zhang, Xiao, & Luo, 2015). For example, Weigle (2002) argues that holistic scoring is better in terms of practicality and authenticity while analytic scoring is better in terms of construct validity and reliability. However, empirical studies comparing holistic and analytic scoring in independent writing tasks have yielded mixed results (Ghalib & Al-Hattami, 2015; Nakamura, 2004; Schoonen, 2005; Shi, 2001; Zhang et al., 2015). In terms of validity, generally high correlations have been reported between the two methods (e.g., Bacha, 2001; Nakamura, 2004; Vacc, 1989), although some research has shown low correlations (e.g., Schoonen, 2005). Thus, there is no consensus on the validity of holistic and analytic scales in independent writing, and little is known about this issue as regards the integrated writing task.

Another important aspect of assessment, reliability, has also shown inconsistent findings in independent writing tasks. While some have reported high reliability for both holistic and analytic scoring methods (e.g., Bacha, 2001; Zhang et al., 2015), others have found differences in inter-rater reliability between methods: Barkaoui (2007) found higher inter-rater reliability for holistic scales, and Ghalib and Al-Hattami (2015) and Nakamura (2004) for analytic scales.

In line with this inconsistency, in the assessment of integrated writing, Ohta et al. (2018) examined holistic and multi-trait ratings of reading-to-write tasks done by students in the United Arab Emirates. Five raters scored persuasive essays using both holistic and multi-trait scales (i.e., source use, organization, development of ideas, language use, and authorial voice). The results showed that the multi-trait scores exhibited higher reliability than the holistic ones, and that the source use dimension was the most reliable and consistent among the five dimensions, whereas the language use dimension exhibited the largest rater variance, indicating difficulty in scoring this dimension. This study is important since it used both holistic and multi-trait rubrics and showed insights into reliability and the unique characteristics of each dimension of the analytic rubric. However, it remains unclear which dimension of the analytic rubric contributes most to the overall analytic score in the integrated writing task.

Thus, discrepancies have been observed in previous studies as regards the validity and reliability associated with holistic and analytic scales. It seems that these conflicting findings on reliability are related to the use and nature of the multiple factors. As argued in the literature, raters’ experience (Knoch, 2009), task type (Chalhoub-Deville, 1995), the nature of the rating process, the context of use, and methodological treatments—for instance, “limited writing sample size, limited number of raters, and lack of direct comparison of the two methods” (Zhang et al., 2015, p. 1)—could all influence scoring results. More research needs to examine the validity and reliability of these two scoring methods using integrated writing tasks.
Raters’ Perceptions of Holistic and Analytic Scoring of Integrated Writing Tasks

Rater perceptions are another important aspect of raters’ scoring behaviors and views of (in)effective performance, as well as the challenges they face while scoring. For instance, Wang et al. (2017) examined rating accuracy in US-based raters who scored a reading-to-write task (an informational essay) using an analytic rubric, and their perceptions of scoring challenges. Analyses of rater perceptions indicated that the raters showed inconsistent perceptions on certain characteristics of compositions, including source text use, idea development, and the focus of the essay. In a Japanese context, Hijikata-Someya, Ono, and Yamanishi (2015) investigated ratings of L2 summary writing tasks using the ETS (2002) holistic scale and the perceptions of native-English-speaking (NES) and nonnative-English-speaking (NNES) teacher-raters. They found that the NES group viewed the assessment of content and language use as difficult, while the NNES group regarded the rating difficulties as being related to paraphrasing and vocabulary use. With a more specific focus on textual borrowing, Gebril and Plakans (2014) investigated two raters’ reactions to source use in a reading-to-write task (an argumentative essay) and found that the raters paid particular attention to “(a) locating source information, (b) citation mechanics, and (c) quality of source use” (p. 56). Moreover, Gebril and Plakans indicated that the raters faced the different challenges between lower- and advanced-level compositions, with the latter requiring more careful judgment of the quality and appropriateness of source use. Thus, raters’ perceptions vary in the dimension of the rubrics and the aspect of source text use in the integrated writing task.

Furthermore, raters encounter various challenges in scoring caused by the rating scale they use. Upshur and Turner (1995) and Turner and Upshur (2002) pointed out the following problems. First, the language of descriptors in the rating scale can be problematic: Vague and impressionistic terminology may confuse raters and lead them to interpret the meaning of the descriptions inappropriately or differently from other raters. Second, the score band levels may not be sufficiently distinct. Third, irrelevant features can be included in the descriptors, or features may overlap across band levels. In investigating raters’ perceptions, it should be considered to what extent these points are related to the difficulties raters face in scoring.

Following these studies, it is necessary to clarify not only what challenges teacher-raters face but also what causes the challenges in relation to rating scales for the integrated writing task. In addition, there remains a paucity of research investigating ratings and raters’ perceptions of integrated writing tasks using holistic and analytic scales simultaneously.

Therefore, this exploratory study aims to investigate teachers’ holistic and analytic rating scores in terms of reliability, validity, and teacher-raters’ perceptions when grading the TOEFL iBT integrated writing task in a Japanese university context. We set the following research questions (RQs) to cover each of these realms:

RQ 1: Are there any differences in reliability and validity between holistic and analytic scoring of the TOEFL iBT integrated writing task?
RQ 2: Under teacher-raters’ analytic scoring, how do the four dimensions of scoring (content, organization, language use, and verbatim source use) overlap, and which dimension contributes most to the analytic evaluation?

RQ 3: What challenges do teacher-raters face when scoring the TOEFL iBT integrated writing task holistically and analytically?

Method

Participants

The participants were 36 Japanese students (13 males and 23 females) from a Japanese university. These students were in their first or second year and were majoring in law or political science while also studying English as a foreign language (EFL). They were enrolled in the Intensive English Program in the Department of Law and Political Science, taking English language classes four times a week. A questionnaire survey revealed that about half of the students in this program were planning to study abroad and that 10 out of the 36 participants had had a former experience taking the integrated writing test in the TOEFL iBT. This means that less than 30% of the participants were familiar with this integrated writing task; however, it was deemed suitable for the group of participants because those who wanted to study abroad needed to prepare for the TOEFL iBT or other proficiency tests. In addition, the participants had been taught how to write a summary in their English classes and instructed to use paraphrasing as much as possible in their summaries, instead of using verbatim phrases from source materials.

To measure the participants’ abilities in reading, grammar, and vocabulary, the Quick Placement Test (University of Cambridge Local Examinations Syndicate, 2001) was used. It is a 30-min test consisting of 60 items, all in multiple-choice format. The participants’ mean score was 39.5 out of 60, and the standard deviation was 4.79. Based on the scoring index, the participants’ English language proficiency was regarded as upper-intermediate, equivalent to Levels B1–B2 in the Common European Framework of Reference for Languages.

Seven English language teachers participated in this study as raters from different universities in Japan. University teachers were selected as raters since we intended to reveal the perceptions of university teachers who have taught English language skills, including academic writing, and are accustomed to the learning styles and educational contexts relevant to EFL. The profiles of the seven raters are shown in Table 1. The raters’ teaching expertise varied, ranging from 4 to 24 years, but they were all experts at English language teaching. As for their grading experience, only one rater (rater 4) had previous experience using the ETS holistic scale; thus, the others were regarded as having little familiarity with the task and little experience scoring it.
Table 1  
Profiles of the Seven Raters

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
<th>Rater 6</th>
<th>Rater 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic position</td>
<td>Lecturer</td>
<td>Lecturer</td>
<td>Lecturer</td>
<td>Associate professor</td>
<td>Associate professor</td>
<td>Assistant professor</td>
<td>Lecturer</td>
</tr>
<tr>
<td>Final degree</td>
<td>MA</td>
<td>MA</td>
<td>MA</td>
<td>PhD</td>
<td>MA</td>
<td>PhD</td>
<td>PhD</td>
</tr>
<tr>
<td>Teaching experience at universities (year)</td>
<td>9</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Teaching experience outside universities (year)</td>
<td>0</td>
<td>15</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Grading experience using the ETS scale</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Materials

The TOEFL iBT integrated reading-listening-writing task was selected as the classroom task to be conducted and examined, because it was deemed appropriate to middle- to low-stakes assessments in instructional settings. In addition, some participants needed to tackle this writing task for the actual TOEFL iBT and study abroad. Task prompts were selected from the exercise book for the TOEFL iBT (Wadden, Hilke, & Hayakawa, 2014; see Appendix A). The reading passage consisted of 223 words, and its readability was 9.7 as measured by Flesch–Kincaid Grade Level. The corresponding lecture was composed of 245 words, spoken at a rate of 145 words per minute, and its readability was 8.6. Given the English proficiency of the participants, the difficulty level of the reading passage and the lecture was deemed reasonable. The topic was “a successful career,” and the passage and the lecture conflicted with one another.

Procedures

All 36 student participants took part in the integrated writing task in a computer-assisted language learning classroom, where each of them could use a computer individually. Before commencing the integrated writing task, its general procedures were explained orally, and detailed instructions were provided in written Japanese. The participants received a piece of scratch paper to take notes during the task; dictionary use and Internet access were prohibited.

Some modifications to the task were considered necessary for the classroom environment, since individual students varied in their task familiarity. The detailed procedures for the task were as follows. First, the participants were asked to read the passage shown on the computer screen within three minutes. Second, they listened to the lecture. Unlike the TOEFL iBT, they were allowed to listen to the lecture twice; this was because the
majority of the participants had never taken the TOEFL iBT integrated writing task, and it was necessary to collect an adequate number of written samples for rating given the purposes of the study. Furthermore, if we found a floor effect, it would not enable us to compare the holistic and analytic scales. Third, the students were instructed to identify and write out the main points and the relationship (e.g., conflicting or supplementary) between the reading passage and the lecture in English. Although the participants were allowed to review the reading passage as many times as they wanted during the task, they had to rely on their own notes for the listening part.

In the phase of computer-medium writing, the keywords of the source texts were shown on the computer screen, with the following instructions: Summarize the points made in the lecture you just heard, explaining how they cast doubt on points made in the reading. You may use the following keywords. We used five keywords based on model writings provided by Wadden et al. (2014; see Appendix A). The TOEFL iBT does not provide such keywords in its integrated writing tasks; however, the keywords were shown to the participants because task difficulty needed to be decreased to help them effectively engage in this intricate task for the first time. The participants were directed to write a composition with 150 to 225 words, which was considered a sufficient length to provide an effective response for this task (Wadden et al., 2014). The time limit of the task was extended from 20 to 30 minutes, because more than two-thirds of the participants had never performed the task and were not able to complete it within 20 minutes. Finally, the participants completed a questionnaire investigating their perceptions of the task’s difficulty; the questionnaire was revised from Cho et al. (2013) and translated into Japanese. Analyses of the student questionnaire are beyond the scope of the present study.

Scoring

In the scoring preparation phase, the 36 compositions were randomly rearranged for each rater, to prevent an order effect. However, the presented order of compositions was not changed between holistic and analytic ratings, nor among the different dimensions of the analytic rubric, because rearranging compositions for each dimension would make the rating process time-consuming and thus might not be feasible in the classroom setting. The raters received the task materials, holistic and analytic rating scales, compositions to be graded, evaluation sheets, and a sample of the model composition that was assigned a score of 5 (the highest mark) so that they could understand the characteristics of the task as a whole, what made an effective response to the task, and the degree of expected stringency of scoring for the task. On-site rater training was not conducted because it was difficult to gather the raters, who worked at different universities, and because rater training might affect their perceptions of the rating scales. The raters were instructed to read all the materials carefully before scoring, ask questions whenever they had them, and report on any concerns or problems they faced in scoring the compositions.

As regards the scoring instructions, the raters were first required to evaluate the 36
compositions using the six-point holistic rating scale (ETS, 2014; see Appendix B). In this holistic scale, the descriptors of each score band specify the expected quality and features of the response for various dimensions such as comprehension, coherence, language use, and connections between the two sources. Second, the raters were asked to score the same compositions using Yang and Plakans’ (2012) six-point analytic rating scale, comprising four dimensions: content, organization, language use, and verbatim source use (see Appendix C); the definitions of the four dimensions were noted above and used in the present study. We adopted this scoring order in line with the procedures used in Ohta et al. (2018) so that the results of holistic rating would not be influenced by the dimensions of the analytic rating scale; if the opposite order had been employed, holistic scores may have been influenced by the analytic scores assigned and the descriptors of the dimensions of the analytic rating scale. A counterbalanced design was not used, as the number of raters was thought insufficient. Unlike Ohta et al., we did not set a specific interval between the two scoring methods but gave the raters a period of one and a half months for the scoring task. We assumed that, as long as the raters followed the specified scoring instructions shown above, they would not compare their own holistic and analytic scores during their scoring process.

After scoring the compositions, the raters were asked to complete a questionnaire—in either English or Japanese—that presented 14 question items regarding the raters’ academic and educational backgrounds and their perceptions of the use of the two types of rating scales. Among the 14 items, nine of them were open-ended questions, which were revised based on the questionnaire used in Hijikata-Someya et al. (2015). The raters were asked to provide their perceptions of the respective degrees of difficulty of using the holistic and analytic rating scales, and to report on any challenges they identified in using the scales.

**Data Analyses**

Quantitative and qualitative analyses were conducted. For the quantitative analysis of rating scores for the written compositions, we adopted the following procedures. First, we examined the descriptive statistics. Second, we estimated inter-rater reliability using Cronbach’s alpha. Third, to confirm validity, we examined the correlations between scores assigned with the holistic scale and those with the analytic scale and then estimated correlations between each of the dimensions within the analytic scale.

Next, the questionnaire data, which contained 72 comments from the raters, were analyzed qualitatively using a data-driven approach. Two of the researchers divided the raters’ comments into 87 semantic units through discussion and independently coded them based on a coding scheme developed for this study. The coding scheme consisted of six main codes under which 48 subcodes were created: (a) positive comments (5 subcodes), (b) negative comments (20 subcodes), (c) comments for improvements (11 subcodes), (d) general impressions (4 subcodes), (e) validity and practicality of the scales (4 subcodes), and (f) clarity of the dimensions in the scales (4 subcodes). An inter-coder reliability check showed a high Cohen’s kappa coefficient ($\kappa = .80$). Disagreements were discussed between the two
researchers until all of them were resolved. Based on the coded data, we examined each of the raters’ perceptions of the challenges of using the scales and the reasons behind them.

**Results and Discussion**

**RQ 1: Are There Any Differences in Reliability and Validity Between Holistic and Analytic Scoring of the TOEFL iBT Integrated Writing Task?**

Table 2 shows the overall results of rater scoring—descriptive statistics for both rating scales. While the average score by all the raters using the holistic scale was 3.06 out of 5, that using the analytic scale (see Overall) was 3.38. However, descriptive and basic statistics comparing mean-score differences cannot provide sufficient information to reveal the reliability characteristics of each scoring method’s evaluation results; this is why we conducted the subsequent analyses.

Taking a closer look at the scoring results, Cronbach’s alphas were estimated in order to examine inter-rater reliability. As a result, both the holistic and analytic scales had high inter-rater reliability: $\alpha = .89$ and $\alpha = .90$, respectively. These results indicate that both scales were reliable measures for evaluating students’ compositions in the integrated writing task. Holistic scales in L2 writing tend to display lower inter-rater reliability than analytic scales (Ghalib & Al-Hattami, 2015; Nakamura, 2004); however, the current study confirmed that both the ETS (2014) holistic scale and the Yang and Plakan's (2012) analytic scale exhibit high inter-rater reliabilities. This result supports Bacha (2001) and Zhang et al. (2015), who contend that not all holistic scales exhibit low inter-rater reliabilities, even with little familiarity with the task type and a lack of sufficient rater training. It is also worth noting that, as far as we are aware, the current study is the first one to show high inter-rater reliability for both holistic and analytic scales using an integrated reading-listening-writing task.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive Statistics for Both Rating Scale</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

*Note.* Mean (SD) of 36 compositions. Maximum score = 5. Overall represents the average scores of the four dimensions of the analytic scale.

We then estimated the correlations between the holistic and analytic scales, to closely look at the consistency of the scoring. The results showed high, statistically significant ($p < .05$) correlations between the holistic score and the overall score for the analytic scale, $r = .89$ (see also Table 3). This means that the scoring results for the two scales are generally overlapping and consistent, particularly in terms of the overall scores, showing high concurrent validity between the scoring methods. The results regarding high correlations...
between the two scoring methods are consistent with previous studies addressing independent writing tasks (e.g. Bacha, 2001; Nakamura, 2004; Vacc, 1989). As the ETS (2014) holistic scale was developed for actual TOEFL iBT tests and should have high construct validity for the test, the results indicate that the Yang and Plakans’ (2012) analytic scale also has high construct validity for the TOEFL iBT integrated reading-listening-writing task.

RQ 2: Under Teacher-Raters’ Analytic Scoring, How Do the Four Dimensions of Scoring Overlap, and Which Dimension Contributes Most to the Analytic Evaluation?

Based on the results of RQ 1, we examined how the four dimensions of the analytic scale overlapped and which dimension contributed most and least to the analytic evaluation by examining item–total correlations. This scrutiny within the analytic scale was important to grasp that scale’s features in terms of construct validity.

Table 3 shows the correlation matrix for the total evaluation results by all the raters. Since the number of compositions evaluated was 36, correlation coefficients exceeding $r = \pm .28$ are statistically significant at $\alpha = .05$. As can be seen, every pair of dimensions of the analytic scale had positive correlations between them, ranging from low to high, $r = .21$ to $r = .81$; this indicates that all four dimensions overlapped but covered different aspects of the composition evaluation. Among these four dimensions, organization and content contributed notably to the analytic evaluation; the item-total correlations between these two dimensions and the overall scores of the analytic evaluation were significant and high: $r = .93$ for the organization dimension and $r = .88$ for the content dimension. However, it was found that verbatim source use was the least contributor to the analytic scoring ($r = .52$), and it had weak correlations to the others: $r = .21$ to $r = .37$. This indicates that the teacher-raters may have evaluated verbatim source use differently from the other three dimensions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Content</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Organization</td>
<td>.81*</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Language use</td>
<td>.62*</td>
<td>.67*</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Verbatim source use</td>
<td>.21</td>
<td>.37*</td>
<td>.28*</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Overall</td>
<td>.88*</td>
<td>.93*</td>
<td>.79*</td>
<td>.52*</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>6. ETS holistic</td>
<td>.98*</td>
<td>.84*</td>
<td>.62*</td>
<td>.23</td>
<td>.89*</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. * $p < .05$. Correlation coefficients between holistic and analytic scales are in boldface.

These results imply that the raters perceived and treated this dimension independently from the other evaluation processes when they used the analytic scale. Alternatively, individual variation in rating behaviors and/or perceptions among the raters may have affected the ways in which verbatim source use was scored. Figure 1 is a sample composition.
by a participant showing such an inconsistency of evaluation; that is, only the score of *verbatim source use* is less, and the score of the other dimensions is greater, than the mean scores (see Table 2).

The ideas of the reading passage and the lecture were completely opposite. While the passage states that successful career is a vocation that provides financial stability and material affluence like jobs in medicine or law, the lecture says that money is not everything when choosing a job and rather we should choose it according to our interest. Also, in order to accomplish their aim, each of the two passages gives us ideas about what to do from three perspectives of view. The first point is how to choose university we enter. The reading puts an importance on its reputation, but the lecture says that we should select our school which is fit to our personality. Secondly, regarding major, while the reading writes that we should choose the best major related to such jobs and focus hard on the classes, the lecture says that we should not specialize too early but rather we should explore different fields in order to identify our interest. Thirdly, regarding grades, the reading says that we must focus on study for good grades, but the lecture states that we should also try variety of activities other than studies. [S15: 192 words]

*Figure 1*. Sample composition by a participant. Underscored emphasis added by the authors to show verbatim phrases (three consecutive words or more) from the sources.

The results discussed thus far indicate that the scoring results are generally reliable regardless of scoring method. However, there seems to exist rating variation in a particular aspect of the assessment: different evaluation behavior for *verbatim source use*. Thus, the construct validity of the analytic rating scale needs more scrutiny. This finding reminds us to consider Yang and Plakans’ (2012) study, in which high scores on effective *verbatim source use* did not necessarily result in high overall scores, and vice versa. In other words, the discrepancy between a score on *verbatim source use* and the overall score of the task could be related to not only learners’ task strategy issues but also raters’ scoring issues. To clarify these interpretations, the perceptions of the raters were investigated in the next sections.

**RQ 3: What Challenges Do Teacher-Raters Face When Scoring the TOEFL iBT Integrated Writing Task Holistically and Analytically?**

The teacher-raters assessed the level of difficulty of using the holistic and analytic rubrics on a four-point scale (i.e., *easy*, *somewhat easy*, *somewhat difficult*, *difficult*). The overall tendency is for the use of the analytic scale to be perceived as more challenging than the holistic scale, regardless of their L1 background (see Table 4). However, rater 4 felt that the analytic rubric was easier to use because “it had clear dimensions, and therefore I understood the criteria for each dimension. This helped assess the responses more easily and smoothly, compared to holistic scoring.”
Table 4  

**Difficulty Levels of the Holistic and Analytic Rating Scales as Perceived by the Raters**

<table>
<thead>
<tr>
<th>Scale type</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
<th>Rater 6</th>
<th>Rater 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Analytic</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

*Note. Easy = 1, somewhat easy = 2, somewhat difficult = 3, and difficult = 4.*

As for holistic assessment, only rater 4, who did have previous experience using the ETS (2014) rubric, viewed holistic scoring as more difficult than analytic scoring. Rater 4 explains the reason for his/her difficulty using the holistic rubric as follows:

My own preference for or impression of the response appeared to influence the overall evaluation of the response. Besides, I was not confident enough to assign extreme scores such as score 0 or score 5, and thus I tended to choose score 3. … It was particularly difficult to distinguish score 4 from score 5 without clear keywords for each score band in the rubric. … Overall, I felt that my assessment was relatively subjective in assigning scores. [Rater 4]

Rater 4 is concerned about the influence of subjective evaluation on the final score, which is often impossible to eliminate due to the nature of holistic rubrics. In addition, the difficulty of judging the ambiguous boundaries between score bands is indicated.

Even those who regarded holistic scoring as easy (rater 2) or somewhat easy (rater 1) recognized this issue. Rater 2 pointed out that “Sometimes it was difficult to distinguish between a [score] ‘2’ and a [score] ‘3.’ Sometimes language [of the descriptors] was vague and unclear. What constitutes ‘minor’ as opposed to more serious omission, inaccuracy or vagueness?” Rater 1 showed confusion about “the inconsistent terms such as ‘point,’ ‘key point,’ ‘important point,’ and ‘major key point’” in the descriptors.

As regards analytic assessment, all the raters other than rater 4 encountered various levels of difficulty. Below, the challenges the raters faced using this scale are scrutinized.

(a) **Time-consuming character of the scoring.** Raters 2, 3, and 7 viewed analytic scoring as time-consuming compared to holistic scoring because, for instance, rater 2 “had to keep re-reading the rubric and the response” and rater 7 “had to assign scores for each dimension.” This characteristic of the analytic scale is also identified in the literature (e.g., Weigle, 2002) and seems to help produce the higher degree of difficulty the raters felt in scoring analytically. Although a recursive scoring process is inevitable, the practicality of analytic scoring for classroom use needs to be considered.

(b) **The impracticality of a six-point rating scale.** Raters 3, 4, 5, 6, and 7 pointed out the impracticality of the analytic scale with six score bands. The more the number of points the scale has (the more precise or detailed its measurements), the longer grading takes, since raters need to carefully evaluate and decide which score band is appropriate for each
dimension. As five teachers emphasized the importance of the practicality of the scale, a detailed point scale may not be suitable for middle- to low-stakes assessments in EFL classes and that scales with fewer score bands suit classroom use better (Knoch, 2011).

(c) Ambiguous boundaries between score bands. Raters 2, 3, 5, 6, and 7 identified unclear boundaries between score bands in all four dimensions. This challenge was also found in the holistic rating. Interdependent score bands are often considered problematic, resulting in unfair, unreliable scoring (Turner & Upshur, 2002; Upshur & Turner, 1995). For instance, rater 2 pointed out challenges in the dimensions of organization and language use due to unclear boundaries between scores. These challenges seem to be related to unclear descriptors or to raters’ inadequate or different understanding of the performance expected at each score level. The unclear score boundaries may lead to low inter-rater reliability and inconsistent scoring by individual raters. Since the raters indicated that the score bands were ambiguous, this issue needs to be resolved by modifying the descriptors and providing clear scoring guidelines.

(d) Complicated scoring of the dimension of verbatim source use. Evaluation of verbatim source use was deemed the most challenging by raters 2, 3, 5, 6, and 7. This finding recalls Shin and Ewert (2015), who suggested that reading-specific dimensions such as text engagement (source use) are more challenging than writing-specific dimensions. Our result is also congruent with Wang et al. (2017), who identified textual borrowing as a characteristic of compositions that causes difficulty in scoring. Rater 3 mentioned that “it’s difficult to see how the marker could evaluate whether the student was able to paraphrase but chose not to or did not paraphrase because they could not.” These challenges faced by the raters were also observed in Gebril and Plakans (2014), in which raters of reading-to-write tasks had “difficulty in distinguishing between language cited from source materials and language produced by writers” (p. 66), when writers incorporated source texts inappropriately. Concerning citation use, rater 5 was unsure how to treat quotes employed in the compositions using the current rating scale, another issue also encountered in Gebril and Plakans’ study.

From a different perspective, rater 1 was concerned about the treatment of verbatim source use, as follows:

I think this last dimension, verbatim source use, should have a more weighted score or should not be included as a dimension, as if someone is simply copying the sentences and gets a 0 for this dimension but gets a decent grade in the other dimensions; that is not worthy of a passing grade. In the holistic rubric, if someone is outright copying the text, they would get a 0 but in the integrated [analytic] rubric, it is only 25% of the entire grade. So, more clarification needs to be included here. If it were me, I don’t know if I would include this as a dimension at all. [Rater 1]

Yang and Plakans (2012) implied that individual raters’ perceptions of source use vary because of “various factors, including task characteristics, cultural differences, personal
beliefs, and epistemology” (p. 94). This seems to account for the different tendencies found in the scoring of verbatim source use among the raters, and possibly the individual raters’ perceptions varied as well. In the present study, the raters pointed out the challenge in assessing the dimension of verbatim source use; this result is different from that of Hijikata-Someya et al. (2015), in which only NNES raters encountered difficulty in assessing students’ textual borrowing skills, specifically paraphrasing in a summarization task. Because the current study employed a more advanced integrated writing task using two source types (namely, reading and listening materials), a number of raters may have faced challenges regardless of their L1 background. Gebril and Plakans (2014) reported that “quality of source use” is key to scoring compositions at an advanced level because raters are likely to pay attention “to a more sophisticated area which has to do with the effectiveness of source use” (p. 69), instead of simply focusing on the location of the source information and the mechanics of citation. For effective scoring, rater 5 mentioned the need to have scoring guidelines for the verbatim source use dimension to ensure clear understanding of the difference between effective and ineffective source use; furthermore, raters 5 and 6 indicated that examples of textual borrowing strategies in each score band should be provided along with the rating scale to support teachers and raters.

(e) Unclear descriptors. Raters 2, 3, 5, and 6 identified unclear descriptors in the rubrics. Problems involving ambiguous and impressionistic wording in descriptors were also indicated by Upshur and Turner (1995). Examples in the dimensions of organization and verbatim source use that the current raters believed contained challenging descriptors are as follows: for the organization dimension, rater 3 reported that “I did not understand what it means to say that ‘cohesion is mechanical’” while rater 5 mentioned that “it was difficult to distinguish among ‘minimal,’ ‘few,’ ‘some,’ and ‘many’ instances of verbatim source use.” The descriptors of verbatim source use could be interpreted differently depending on the raters’ understanding, because the interpretation of the terms “minimal,” “few,” “some,” and “many” may vary, as Upshur and Turner argued. These descriptors need to be improved to make grading with the analytic rubrics more efficient.

(f) Indistinct foci of different dimensions. Raters 5 and 6 pointed out that the foci of the dimensions of content and organization were indistinct. This observation could be related to high correlations between content and organization, yet more scrutiny is needed to ascertain the exact nature of this issue. As only two teacher-raters made this observation, raters’ L1 background or individual variation may affect particular dimensions of the rubrics. These raters perceived that the analytic rubric did not specify the expected overall organization of a task response, in terms of cohesion and coherence at paragraph and essay levels, and that the organization dimension did not seem to effectively serve a role in evaluating task performance. However, it is difficult to specify the summary’s overall expected structure in the rubrics because the genre of summary writing does not require any distinctive organizational pattern or structure, unlike in genres such as argumentative or persuasive essays. In addition, the structure of source texts seems to affect how test-takers
compose their writing, because they may intentionally or unintentionally learn features of writing, such as organizational structures and cohesive devices, while exposed to the source texts (Hirvela, 2004). Thus, the organization dimension for this genre may focus more on paragraph structure, coherence and cohesion for easier and more valid scoring.

In short, (a), (b), (c) and (e) relate more to mechanical aspects of band descriptors used for the holistic and analytic assessment whereas (d) and (f) are particularly important for the assessment of integrated writing tasks and more relevant in the difficulty of the analytic assessment. As pointed out by the raters, it seems necessary to prepare scoring guidelines and benchmark responses for different performance levels and to modify the scales in terms of the number of score bands and the descriptors for assessment of integrated writing tasks.

Conclusion

This study contributes to the fields of language assessment and L2 writing research by investigating holistic and analytic ratings using the TOEFL iBT integrated writing task. To our knowledge, this study is the first to examine the reliability and validity of raters’ holistic and analytic scoring results as well as their perceptions of scoring for this task type. New insights obtained from this study are summarized as follows: (a) the holistic and analytic scales exhibited high inter-rater reliability; (b) there were high correlations, that is, high consistency, between the two rating methods; and (c) the dimensions of content and organization were highly correlated to the analytic overall score. However, (d) verbatim source use was found to be peculiar in terms of construct validity for the analytic scale.

The results of the qualitative analyses showed that most of the raters found the use of the analytic scale more challenging than that of the holistic scale. Limited familiarity with scoring this task type and limited previous experience using the target scales could have affected the individual teacher-raters’ perceptions, as discussed in Chalhoub-Deville (1995) and Knoch (2009). The results also indicated various challenges the raters faced in using the holistic and analytic scales. They found verbatim source use to be the most challenging dimension, but with subtly different emphases: Some teachers perceived the difficulty of this dimension as being related to the intricate process and characteristics of textual borrowing strategies, while others emphasized the difficulty of judging multiple types and degrees of textual borrowing.

Despite these meaningful findings, this study also has limitations. First, the number of student participants was small and their L2 proficiency did not vary widely. Including more participants from a wider range of L2 proficiency levels could have yielded more profound results. Second, the integrated writing task used was of only one type and on only one topic. Employing more than one type of integrated writing task would have provided richer data and more profound findings. Third, the number of raters was relatively small, and formal rater training was not conducted; each rater was only provided with one writing sample. Future research should provide rater training and a range of samples.

Pedagogical implications for English language teachers teaching and assessing
integrated writing tasks can be summarized as follows. First, they are encouraged to choose either a holistic or an analytic rubric to match the nature of their assessment goals, since both scales exhibited high reliability and correlations. In choosing, teachers need to understand and decide what aspects they should assess in the integrated writing task, based on the rubric. They are encouraged to prepare several sample responses beforehand and share rubrics with students, so that the challenges the raters in this study faced can be mitigated and a clearer connection between teaching and assessment. Second, teachers need to pay particular attention to verbatim source use in terms of quality and quantity of paraphrasing (see Yamanishi, Ono, & Hijikata, 2019) when scoring integrated compositions. To begin with, teachers and raters should have a clear idea about variation of appropriate source use strategies across integrated writing tasks before implementing a task. To enhance students’ understanding and use of textual borrowing strategies, teachers can show explicit guidelines, a definition of verbatim source use, and rubrics making reference to paraphrasing, patchwriting (e.g., Pecorari, 2003), and plagiarism. For example, Yamanishi et al.’s (2019, p. 19) analytic rubric on summary writing associates different degrees of paraphrasing with each score band, where use of “more than 4 words in a row copied from the original text” is defined as inappropriate paraphrasing. Such explicit instructions are necessary in any integrated writing task in the classroom. We hope these findings help teachers and raters use holistic and analytic rubrics in teaching and assessment of integrated writing tasks.

Future study needs to investigate teacher-raters’ perceptions of analytic and holistic scoring and integrated writing tasks more deeply, with a focus on raters’ experience of and beliefs on teaching and assessment of academic writing. It might be helpful to interview raters about which incidents complicate assessing integrated writing tasks. In addition, test-takers’ perceptions of task difficulty need analysis to better understand of their views of the integrated writing task and challenges they face. The insights obtained thereby may enable us to improve our teaching and assessment methods for the integrated writing task.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP26580121. We gratefully thank the anonymous reviewers for their insightful comments.

References


like and where should the criteria come from? Assessing Writing, 16, 81–96. https://doi.org/10.1016/j.asw.2011.02.003


Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of
the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36, 49–70. https://doi.org/10.2307/3588360
Oxford University Press.
*The Elementary School Journal*, 90, 87–95. https://doi.org/10.1086/461604
Wadden, P., Hilke, R., & Hayakawa, K. (2014). *TOEFL® test: Writing mondai 100 [TOEFL®
test: Writing exercises 100].* Tokyo: Obunsha.
accuracy and perception for integrated writing assessments using a mixed methods
Yamanishi, H., Ono, M., & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary
Yang, H.-C., & Plakans, L. (2012). Second language writers’ strategy use and performance on
an integrated reading-listening-writing task. *TESOL Quarterly*, 46, 80–103. https://doi.org/10.1002/tesq.6
Yu, G. (2013). From integrative to integrated language assessment: Are we there yet?
Appendix A: Materials and Prompts Used for the Integrated Writing Task
(adapted from Wadden et al., 2014, pp. 200–204)

Passage for Reading (reading time: 3 minutes)

When most people think of a “successful career” they mean a vocation that provides financial stability and material affluence. Probably the two most reliable such careers are in medicine or in law. For example, the typical physician earns about $140,000 a year, and the average lawyer around $93,000 a year.

The first question a person who wants to have a rewarding career in a field such as medicine or law must ask is which college to attend. The answer is simple: the more reputable the better. The status of your university will have a big influence on whether you are accepted in medical or law school. Next, after entering the best possible university, you will then need to carefully consider the best major and focus hard on your classes. For medicine, that major will usually be chemistry or biology. Finally, you will need to devote most of your time to studying because you’ll need to get top grades and high scores on tests that qualify you for professional schools, such as the MCAT for medical school or the LSAT for law school. “Strong academic performance in the right major at a top university.” That’s the recipe for achieving your career goals.

Once you have that, the rest will take care of itself. When was the last time you heard of an unemployed doctor? (223 words)

Script for Listening (Lecture)

Now listen to part of a lecture on the topic you just read about.

The first thing you should know about careers is that money isn’t everything. Don’t pursue a career just because it is financially lucrative. Many students enter university as “pre-med” or “pre-law” students only to find after two years or more of taking pre-med and pre-law courses that they are not interested in medicine or law. Some don’t find this out until they are close to graduation. Please keep in mind there are many occupations besides “doctor” or “lawyer.”

For success in college and in life, select a school that fits you personally. If you want close relationships with your professors, consider a small liberal arts college. Don’t worry so much about reputations; you can get a good education at many different institutions—it ultimately depends on what you put into it.

Next, don’t specialize too early. Don’t try to decide your major until you have taken courses in different fields so that you can identify your interests. After that, you then can consider occupations related to those interests.

Finally, study is important, but don’t just focus on grades. Explore. Enjoy a wide variety of activities during your university years. Success in any vocation has a remarkably low correlation with college G.P.A. Most famous people change careers several times during their lives, and the first job you choose right after college probably won’t be your career 20 or 30 years from now. (245 words)

Directions: You have 20 minutes to plan and write your response. Your response will be judged on the basis of the quality of your writing and how well your response presents the points in the lecture and their relationship to the reading passage. Typically, an effective response will be 150 to 225 words. You are NOT allowed to use a dictionary.

Question: Summarize the points made in the lecture you just heard, explaining how they cast doubt on points made in the reading. You may use the following keywords.

Keywords: success career university major grade
### Appendix B: TOEFL iBT® Test Integrated Writing Rubrics

<table>
<thead>
<tr>
<th>SCORE</th>
<th>TASK DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td><strong>A response at this level</strong> successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.</td>
</tr>
<tr>
<td>4</td>
<td><strong>A response at this level</strong> is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. A response is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connection of ideas.</td>
</tr>
</tbody>
</table>
| 3     | **A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following:**  
  - Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading.  
  - The response may omit one major key point made in the lecture.  
  - Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise.  
  - Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections. |
| 2     | **A response at this level contains some relevant information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one or more of the following:**  
  - The response significantly misrepresents or completely omits the overall connection between the lecture and the reading.  
  - The response significantly omits or significantly misrepresents important points made in the lecture.  
  - The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture. |
| 1     | **A response at this level is marked by one or more of the following:**  
  - The response provides little or no meaningful or relevant coherent content from the lecture.  
  - The language level of the response is so low that it is difficult to derive meaning. |
| 0     | **A response at this level** merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank. |

*Note.* Copyright © 2014 Educational Testing Service (ETS). www.ets.org. TOEFL iBT® Test Integrated WRITING Rubrics reprinted by permission of ETS, the copyright owner. All other information contained within this publication is provided by Japan Language Testing Association and no endorsement of any kind by ETS should be inferred.
Appendix C: Analytic Scoring Rubrics for Integrated Writing

Content
5 – A response is effective in selecting principal information from the lecture to support ideas and accurately connecting key points in the lecture and the reading.
4 – A response is effective in selecting principal information from the lecture to support ideas and accurately presenting information from the lecture in relation to the writing. Only minor points from the source content are misrepresented, omitted, or imprecisely connected.
3 – A response contains some key points from the lecture and connects those points with the reading while the points are imprecisely, incorrectly, or incompletely presented or connected. One major key point made in the lecture may be omitted.
2 – A response contains only a few relevant points made in the lecture while they are significantly misrepresented. No, little, or incorrect connections are drawn from the reading and the lecture.
1 – A response fails to use any meaningful or relevant content from the lecture.
0 – A response is not connected to the topic, written in a foreign language, or left blank.

Organization
5 – A response is well organized, cohesive, and coherent at the paragraph and essay level.
4 – A response is generally organized and coherent at the paragraph and essay level although few instances of sentence- or paragraph-level cohesion may be mechanical.
3 – A response is inconsistently and incoherently organized.
2 – A response is poorly organized and lacks paragraph and essay cohesion.
1 – A response has insufficient length to determine organization.
0 – A response is left blank or written in a foreign language.

Language Use
5 – A response displays accurate and appropriate sentence formation and word forms although minor usage and grammatical errors may occur occasionally.
4 – A response generally displays accurate and appropriate sentence formation and word forms. Although some noticeable minor usage and grammatical errors may occur occasionally, the clarity of the response is not affected.
3 – A response contains frequent usage and grammatical errors that may result in noticeable obscured meanings and expressions or unclear connection of ideas.
2 – A response contains usage and grammatical errors that may largely obscure meanings or connection of ideas.
1 – A response is almost incomprehensible due to language errors. Little meaning can be derived from the response.
0 – A response is totally incomprehensible due to language errors or because the response is left blank or is written in a foreign language.

Verbatim Source Use
5 – A response displays no verbatim source use.
4 – A response displays minimal verbatim source use.
3 – A response displays few instances of verbatim source use due to unsuccessful attempts to paraphrase.
2 – A response displays some instances of verbatim source use.
1 – A response displays many instances of verbatim source use.
0 – A response displays wholesale copying from the reading for most parts.