The New Possibilities of and Problems with Writing Assessment in the Context of Integrated-skills Teaching

Kahoko Matsumoto

International Education Center, Tokai University

mkahoko@tsc.u-tokai.ac.jp

Abstract

As secondary- and tertiary-level English classes in Japan have become more skills-integrative, the need for integrated-skills assessment has been increasing. This also reflects the requirements of the so-called "The 21st Century Skills" and the increasing trend of teaching meaningful contents in English or any other foreign language, notably by the CLIL method. As a general rule, effective assessment aligns with both the class contents and the teaching method, namely, what is being taught and how it is taught. Therefore, teachers who want to make an effective connection between input and output should learn how to create valid, reliable, and feasible integrated-skills tests. In this paper, the new possibilities and merits of using and assessing writing in various integrated-skills courses will be summarized while the problems and/or difficulties related to the use of writing final products for the assessment for such skills-integrative, communicative courses will be discussed. In doing so, the results of studies and the perspectives of authorities such as Plakans (2013 & 2015) and Cumming (2013) will be referred to. Finally, more systematic test specifications and rubrics for valid and reliable integrated-skills assessment through writing will be proposed based on the analysis of students' data collected by the TOEFL®-iBT.

Keywords

Integrated-skills writing assessment, Assessment literacy, High-level cognitive constructs

1 Introduction

This paper was presented as a part of the symposium at the 21st PAAL Conference, the theme of which was "L2 English Writing Instruction in Asian Context: Theory, Methodology and Evaluation." Many insights were gained from 3 distinguished panelists' presentations (Dr. Ken Hyland, Dr. Kyoko Oi, and Dr. Barley MAK) on the directions and present conditions of writing instruction in different Asian countries and the problems we should solve for advancement in this area. As the last speaker, I, as an assessment specialist, tried to introduce a new perspective, that is, integrated-skills assessment, which has been attracting increasing attention, reflecting the requirements of using language in actual academic or professional situations. In other words, any communication entails the connection between input and output, and also, the more complicated communication becomes, the more high-level cognitive skills are required (Sasaki & Hirose, 1996), which Cummins (1979) explained as CALP and are also manifested as important skills in "The 21st Century Skills" (Ananiadou & Claro, 2000).

In fact, integrated-skills assessment has been common in ESP and EAP courses where presentations and research papers are counted heavily towards the final grade. Also, other courses which involve project-based, communicative tasks often use some portfolio-type holistic assessments, which are naturally skills-integrative. The problem is that there have been few systematic studies on how to create a valid and reliable rubric for these courses, hence the assessment is done mainly based on each teacher's instincts gained from their experience.

As 4-skills integrated instruction in Japanese secondary school English education has become more prevalent in response to the Ministry of Education's initiatives (MEXT, 2014), there is an increasing need for classroom teachers to adopt integrated-skills assessment. It is quite natural that students who receive such instruction need to be assessed on what has actually been taught. As a general rule, effective assessment aligns with both the class contents and the teaching method, namely, what is being taught and how it is taught. Because teaching and assessment are the two sides of a coin, teachers who want to make an effective

connection between input and output should learn how to create valid, reliable, and feasible integrated-skills tests and/or do more long-term project-based, portfolio-type assessment. This in turn will facilitate evaluating the students' progress and the effectiveness of the teaching. In addition, gate-keeping tests such as the TOEFL-iBT® have become skills-integrative, reflecting the requirements of using language in actual academic or professional situations.

An academic group to which this writer belongs (JACET Testing SIG) has been giving annual workshops on integrated-skills teaching and related test creation for prospective and in-service teachers with an eye toward raising Japanese English teachers' assessment literacy (Taylor, 2009), where some typical difficulties with integrated-skills teaching and assessment have been observed. At the same time, it was discovered in a questionnaire-based study that secondary school teachers generally feel much difficulty to create integrated-skills tests (Murray, et. al, 2011, 2013). Especially, the biggest challenge facing us is how we can incorporate this kind of authentic, real-life oriented assessment into the existing rigid entrance examination system in Japan.

2 Merits and Challenges of Integrated-skills Assessment

2.1 Definition

Integrated-skills assessment can be defined as the use of tests that combine two or more skills, such as reading/writing or reading/listening/speaking (Plakans, 2011, 2012, 2013; Cumming, 2013; Read, 2015). Typically, "integrated skills" tasks require test takers to use the information provided by the reading and listening passages to answer with written and/or spoken responses.

2.2 Assessment Literacy

Before discussing specific difficulties and challenges in creating and using integrated-skills tests, the important requirements for a valid, reliable and feasible assessment should be confirmed based on assessment literacy (Inbar-Lourie, 2008). He used 3 lay words (Why, What and How) to explain the elements to be considered in making any good test.

1) 'Why' refers to the rationale behind the testing. For instance, a major purpose of using an integrated-skills assessment is to evaluate real-life communicative skills that future global citizens need.
2) 'What' deals with the current theories regarding assessment involving validity, reliability and practicality. So, in case of integrated-skills assessment, this covers many considerations necessary to construct valid and reliable integrated-skills tests which measure authentic communicative skills, especially in everyday teaching practice
3) 'How' focuses on test construction, development and the role of assessment in a language curriculum. It is inevitably related to other 2 aspects; choice of methods/tasks that fit the purpose (Why) should be made while their validity and reliability (What) should be checked. Also, the chosen methods/tasks should be doable and manageable.

2.3 Merits of integrated-skills assessment

An evidence of the increasing attention to integrated-skills assessment is that *Language Assessment Quarterly* devoted an entire issue (Volume 1 of Number 10 in 2013) to the assessment of integrated-skills writing. The reason why this established journal has not dealt with integrated-skills speaking assessment at that time is probably an insufficient accumulation of research because assessment of speaking itself encompasses many aspects with its rubrics being far from well-established (Taylor, 2011). First, the following are the advantages of integrated-skills writing assessment summarized mainly based on the introductory article of the above-mentioned journal that Cumming (2013) wrote unless otherwise stated.

1. It provides realistic and higher-level literacy education including problem-solving skills that will lead to future academic and professional activities.
2. It engages examinees in writing as a meaningful communicative activity that requires the expression of one's ideas about specific content or for a specific purpose. In fact, different language skills naturally interact with each other, hence integrated-skills assessment has more authenticity (Plakans, 2012).
3. It counters the test method or practice effects associated with conventional types of writing assessment

and helps create a new model of teaching and assessing.
4. It evaluates language abilities at a more complex, interactional level as examinees integrate and reconstruct multi-modal information. In other words, it fits the increasing requirements of multi-literacy education (Cope & Kalantazis, 2000).
5. It provides more constructive feedback in teaching and self-assessment and more positive washback effects for students are expected (Plakans, 2012, 2013).
6. There may be some positive effect of limiting external intervening factors such as existing knowledge or experience so that targeted abilities can be measured more accurately (Gebril & Plakans, 2013a).

2.4 Problems with integrated-skills assessment

Naturally, integrated-skills assessment poses more challenges in that the constructs to be measured are both input and output related, which makes the evaluation method and rubric more complicated. Though it has long been used for ESP and EAP courses as previously mentioned, it has not been viewed, discussed or studied from the viewpoint of integrated-skills assessment. The following are the points Cumming (2013) pointed out as problems in the same article.

1. It is hard to distinguish the constructs related to the comprehension of stimulus material (reading passages and/or listening material) and those related to writing.
2. In the use of evaluation results, the purpose of assessment often becomes unclear or ambiguous. Thus, its use for diagnostic feedback or self-assessment can easily be confused with high-stakes decision making and vice versa.
3. The selection of the genres and tasks with insufficient reliability are often observed.
4. Most integrated-skills assessment involves tasks that require certain level of proficiency (=threshold), in addition to the fact that various factors affect the results. Therefore, fair and accurate comparison based on equivalency, which is needed for any high-stakes decision making is hard to be achieved.
5. In terms of actual evaluation of the output (=writing), it is hard to determine which part is attributable to the examinee's comprehension of source material and which part is realized by his or her writing ability.

2.5 New Developments

Yet, there are some new studies emerging that made attempts to ascertain the manifestation or reflection of source material in the final artifact (Sawaki, et. al, 2013; Weigle, Yang & Mntree, 2013). Plakans herself published a summary article titled "Integrated Second Writing Assessment: Why? What? How?" (2015) in line with the Inbar-Lourie's framework of assessment literacy. The following is her idea of 3 elements (Why, What and How) as related to integrated-skills assessment of writing.

1. Why?
   First, she referred to the representativeness of the construct which can be made possible by applying the newly created concept "discourse synthesis" (=how the content of input is integrated rhetorically at the discourse level). Secondly, authenticity was emphasized as done by Cumming, on the ground that learning through meaningful, realistic problem-solving is ever more required in this age of multi-modal communication. Lastly, she pointed out the fact that such way of learning will be preferred by students and raise their motivation.
2. What?
   In her quest for a better method of integrated-skills assessment, she categorized the accumulation of research in this area into 3 types.
   (1) Comparison between independent writing evaluation and that of integrated-skills writing;
   (2) Studies about the correlations between integrated-skills writing evaluation and general English proficiency;
   (3) Studies about the relationships between integrated-skills writing scores and the use of source material (=input).
3. How?
   She raised 3 considerations as necessary for the advancement of research on integrated-skills writing evaluation.
   (1) Selection of source (stimulus) material that fits the purpose of evaluation and has sufficient validity in light of it.
   (2) Necessity of proper task design and clear instructions so that examinees have little difficulty

grasping the requirements of the final writing output. In accomplishing this, the characteristics and past experiences of particular examines may have to be taken into account.

(3) Increase and accumulation of more studies on systematic inquiries into rubric construction and evaluation training.

3 A Study on the Integrated-skills Parts of TOEFL®-iBT

3.1 The Purpose of the Study

An attempt was made to find the correlations between integrated-skills writing scores and independent writing scores of TOEFL®-iBT using Japanese university students. Also, additional qualitative analysis was done in order to tease out the factors that led to good scores for the integrated-skills writing test items with particular attention paid to the new construct "discourse synthesis."

3.2 Subjects

One hundred and three students who were enrolled in 3 advanced-level TOEFL preparation courses whose TOEFL®-iBT scores are 60 and over participated in this experiment as an in-class activity. The experiment was made sufficiently beneficial to the students because the teachers gave detailed feedback to all the written and spoken responses. For a close statistical analysis, the subjects were divided into 2 groups by English proficiency; the Advanced Group (scores between 80 and 100) and the Intermediate Group (scores between 60 and 79).

3.3 Methodology

Because the official TOEFL®-iBT score report does not provide the number of points for each item, the test items in the Official Practice Book published by ETS were used for the experiment. Though the focus was on integrated-skills writing assessment, the data was collected for both speaking and writing items to get a general tendency as well as to find the effect of integration. Each subject answered one independent test item and one integrated-skills test item from both the writing and speaking sections. Then, 2 raters who are well versed with TOEFL®-iBT and its preparation evaluated the students' writing and speaking responses. When their points did not agree, the third rater decided the final score considering the 2 raters comments. After the statistical analysis, a qualitative analysis of writings representative of the 3 groups identified by the cluster analysis was conducted.

3.4 Statistical Results

Actually, the division of the subjects into the Advanced and Intermediate Groups was not done arbitrarily; when checking the correlations between the integrated-skills test items and independent ones, it was found that the correlations change around the score of 80 points. Thus, the balance of the number of students belonging to the 2 levels is skewed. The correlations between the integrated-skills test items and independent ones were significantly higher for the Advanced Group (34 subjects), which were 0.74 for writing and 0.78 for speaking. On the other hand, for the Intermediate Group (69 subjects), not only the correlations were lower (0.64 for writing and 0.61 for speaking), but also the standard deviation and variance were a lot larger. It means that there are many students in the Intermediate Group whose subskill combinations are varied.

In order to find the general tendency on the relationship between the integrated-skills tasks and independent ones, a cluster analysis was conducted, which found the following 3 groups (=clusters) plus outliners.

1. Those whose scores of integrated-skills test items and independent ones showed significantly high correlations (45 subjects)
2. Those whose scores of independent test items were considerably higher than the scores of integrated-skills ones (28 subjects)
3. Those whose scores of integrated-skills test items were considerably higher than the scores of independent ones (16 subjects)
4. Outliners (14 subjects) = some irregularities or inconsistencies were found in their data.

The breakdown of 2 levels in each cluster is shown in Table 1.

Table 1: Breakdown of Each Cluster

Cluster	Advanced-level	Intermediate-level
1	22 (65%)	23 (33%)
2	4 (12%)	24 (35%)
3	6 (17%)	10 (15%)
Outliners	2 (6%)	12 (17%)

As previously noted, another reason why both speaking and writing data was analyzed first, was to see the effect of meta-cognitive skills in responding to the integrated-skills test items, and incidentally, to eliminate outliners. Predictably, most subjects in the Advanced Group were included in the clusters #1 and 3. This may prove that the nature of high-level meta-cognitive skills required to respond to these test items either in writing or speaking seem to be quite similar; namely, first, understanding the source material, secondly, making judgment of what content to be reflected after analysis, and lastly, integration of the necessary information into proper discoursal or rhetorical development. Also, the above result is the testimony of a certain level of English proficiency being needed to achieve sufficient level of "discourse synthesis." At the same time, the variance found in the Intermediate-level subjects may have been caused by unbalanced subskill combinations. Some of them may possess high proficiency in receptive skills (reading and listening) while lacking in the ability to reflect the inputted information in terms of productive skills (speaking and writing) or vice versa. For others, lack of generic, meta-cognitive skills may have hindered them from obtaining high scores for integrated-skills test items. Also, there are a small number of subjects in the Advanced Group who were in cluster #3, meaning that they may not have sufficient meta-cognitive skills, but obtained high overall scores by doing well mostly in independent tasks.

3.5 Qualitative Inquiries

The general statistical tendency was quite similar between 2 kinds of integrated-skills test items; one that solicits speaking as an output and the other with writing as an output. Since the focus of the symposium was writing, a qualitative analysis was done to a few writing samples that seemed representative of the 3 clusters. The following are the descriptions of Level 5 (best score) and Level 3 (middle score) of TOEFL®-iBT integrated-skills writing items. The descriptions are holistic and do not explicitly allude to the construct for measurement, but it is quite clear what kinds of constructs are embedded, which are shown by underlines and annotations in parentheses.

<Level 5>
A response at this level successfully selects the important information from the lecture (listening skill + semantic/organizational judgment) and coherently and accurately presents this information in relation to the relevant information presented in the reading (reading skill + organizational/rhetorical skill for synthesis). The response is well-organized and occasional language errors that are present do not result in inaccurate or imprecise presentation of the content or connections (grammatical and lexical ability).

<Level 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 (insufficient listening and reading skills + insufficient skill for synthesis).
- The response may omit one major key point made in the lecture (insufficient listening skill + insufficient semantic/organizational judgment)
- Some key points made in the lecture or the reading, or the connections between the two, may be incomplete, inaccurate or imprecise (insufficient organizational/rhetorical skill for synthesis).
- 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 (insufficient grammatical and lexical ability).

Actually, the holistic rating based on these descriptions was not so difficult if the raters have enough training and experience. However, finding the reasons for inadequate output, in this case, the resultant writing is quite difficult. For example, if a major point in reading material is not connected with that of the listening material, it is hard to determine whether the inadequacy is caused by lack of either reading or listening skill, or an inability to integrate or synthesize the two important points gleaned from different

source materials.

In the first cluster (group) with the subjects whose scores of integrated-skills test items and independent ones showed high correlations, 22 out of the 45 Advanced-level students were included. Probably because of their all-round proficiency, their writings fulfilled most of the Level 5 requirements (=embedded constructs annotated in parentheses) though with varying degrees. In the same vein, their scores of the independent writing task were also high, which put them in the Advanced Group. On the other hand, the outputs of the integrated-skills writing task of Intermediate-level subjects in this cluster showed either a lack of receptive skills or an inability for synthesis, or both. Again, it is hard to tell what caused poorer scores unless some deeper analysis such as think-aloud protocol is done. As found in Gebril and Plakans' study (2013b), there was a tendency for lower-level subjects to use more direct quotes from the reading material while more attempts to paraphrase were detected in the writing outputs of higher-level subjects.

The second cluster with the subjects whose scores of independent test items were considerably higher than those of integrated-skills test items mostly consisted of Intermediate-level students (24 out of 28 subjects). It is natural that the students' subskill combinations are more varied at the lower level; thus, these Intermediate-level students have good writing skills, but couldn't utilize them fully in the integrated-skills task due to a lack of receptive skills. The most typical problem with their outputs was the inability to connect key points of reading and listening materials though some showed quite good writing skills, especially in the part opinions were expressed.

Lastly, the third cluster included 6 Advanced-level subjects and 10 Intermediate-level ones. In general, both of these Advanced- and Intermediate-level students appeared to offset a relative lack of writing skills with either good listening and reading skills and/or high meta-cognitive abilities to synthesize whatever they comprehended from source materials, of course at different levels. Here also, Intermediate-level students tended to use direct quotes while Advanced-level students resorted to more paraphrasing.

4 Implications and Propositions

With the changes of the secondary school curriculum and the prospect of increase of problem-solving and project-based instruction to nurture the 21st Century Skills, we, teachers have to be prepared to become able to create valid, reliable and feasible integrated-skills tests. There still are problems and challenges concerning such measurement as previously explained, but we should try to augment the research in this area as well as create training courses based on recent studies and models such as Assessment Production Cycle proposed by Green (2014). One worksheet developed for such training is attached as Appendix A, which should help teachers raise their assessment literacy and pay more attention to important concepts such as validity, reliability, and practicality.

5 References and Appendix

5.1 References

Ananiadou, K., & Claro, M. (2009). *21st Century skills and competencies for new millennium learners in OECD countries*. OECD Education Working Papers, No.41, OECD Publishing.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Cope, B., & Kalantazis, M. (2000). (Eds), *Multiliteracies: Literacy learning and the design of social futures*. London, UK: Routledge.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10(1),* 1-8.

Commins, J. (1979) Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19,* 121-129.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, UK: Cambridge University Press.

Demauro, G. (1992). Examination of the relationships among TSE, TWE, and TOEFL scores. *Language Testing, 9*(2), 149-161.

Gebril, A., & Plakans, L. (2013a). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(2), 217-230.

Gebril, A., & Plakans, L. (2013b). Toward a transparent construct of reading-to-writing tasks: The interface between discourse features and proficiency, *Language Assessment Quarterly, 10*(1), 9-27.

Green, A. (2014). *Exploring language assessment and testing*. London and New York: Routledge.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385-402.
MEXT (2014). Policies for the future improvement of English Education in Japan. Retrieved from http://www.mext.go.jp/b_menu/shingi/chousa/shotou/102/houkoku/attach/1352464.htm
Murray, A., Ito, Y., Kimura, K., Matsumoto, K., Nakamura, Y., & Okada, A. (2011). Current trends in language testing education in Japan. *JALT 2010 Conference Proceedings*, 106-118.
Murray, A., Akiyama, T., Matsumoto, K., Miyazaki, K., Nakamura, Y., and Tsuchihira, T. (2013). Fundamental issues surrounding integrated tests in terms of assessment literacy - The case of integrated speaking tests. *Selected Papers of the 17th Conference of Pan-Pacific Association of Applied Linguistics*, 31-41.
Plakans, L. (2012). Writing integrated items. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp.249-260). Abingdon, UK: Routledge.
Plakans, L. (2013). Assessment of integrated skills. In C.A.Chapell (Ed.), *The Encyclopedia of applied linguistics, Vol. I* (pp.204-212). Hoboken, NJ: Wiley-Blackwell.
Plakans, L. (2015). Integrated second language assessment: Why? What? How? *Language and Linguistic Compass*, *9*(4), 159-167.
Read, J. (2015). *Assessing English Proficiency for University Study.* UK: Palgrave Macmillan.
Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46(*1), 137-174.
Sawaki, Y., Quinlan, T., & Lee, Y. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*(1), 73–95.
Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics. 29*, 21-36.
Taylor, L.(Ed.) (2011). *Examining speaking (Studies in language testing #30)*. Cambridge, UK: Cambridge University Press.
Weigle, S.C., Yang, W., & Montee, W. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly, 10*(1), 28-48.

5.2 Appendix A. Worksheet to form ideas of an integrated-skills test

(Based on Language Assessment Literacy Concepts)

1. "WHY" do you want to create this test item (related to teaching objectives)?
   Ex.: in order to develop the ability for students to write a research paper (or a response paper) based on what they have read or listened to, by effectively incorporating important facts in their papers

2. "WHAT (construct)" do you want to measure?
   Ex.: the ability to write a good summary extracting and connecting important points from a reading (or listening) material

3. "HOW" do you want to measure it?
(1) Content and Nature of the Task
   Ex.: a news report on global warming to be summarized as an individual, formative task

(2) Test Item Format
   Ex.: a scaffolded (guided) task with a half of important points and some transition devices given

(3) Difficulty Level
   Ex.: Lower-intermediate (TOEIC® 400-450)

(4) Assessment Criteria with Weighting

Ex.: reflection of important points (50%), sentence accuracy (20%), use of proper expressions (10%) and genre-based cohesive discourse progression (20%)

*If the task is timed, writing fluency can be added.