L2 writing assessment in context: The case of a Japanese EFL secondary school

Gordon MYSKOW *

This article situates the discussion of L2 writing assessment in a particular context by describing the assessment practices used in one Japanese EFL secondary school course. The article begins by discussing a key contextual factor in designing L2 writing assessment, the purposes of testing. It then describes the development of two course objectives and discusses some considerations involved in designing rubrics and prompts to assess the extent to which students achieved these objectives. Though this article focuses on assessment in a secondary school context, many of the considerations may be of relevance to teachers in related contexts.

Key words: Japanese EFL Context, foreign language writing, classroom assessment, course objectives

Bachman (1990) observes that language assessment does not take place in "a values-free psychometric test tube" (p. 279). The ways an instrument is designed and administered in specific contexts are contingent upon a multitude of local factors, including logistical constraints, the values and interests of stakeholders, and the many possible purposes that the test could be used for. Such varied considerations lead Hamp-Lyons (1991b) to conclude that, "The depth and breadth of our knowledge are not sufficient for us to be able to judge the appropriacy of any given writing assessment across the board" (p. 326).

This article discusses key areas of L2 writing assessment literature in the context of a specific educational setting, a Japanese EFL secondary school. While much of the assessment literature focuses on large-scale, high-stakes test development, this article views many of these same concepts in the context of day-to-day classroom assessment. It is my hope that situating the literature in a particular setting may help to generate some specific insights that cannot always be gleaned from discussions of assessment in other more general or

＊ A lecturer in the Faculty of Economics, and a member of the Institute of Human Sciences at Toyo University

distant contexts.

The article begins by discussing a key contextual factor in designing second language writing assessment, the purposes of testing. It then looks at the development of two course objectives for an academic summary and an argumentative essay. Finally, the article discusses some of the many considerations involved in designing rubrics and prompts to assess the extent to which students achieved these course objectives.

Purposes of writing assessment

One of the most important things to consider before designing any assessment tool is its purpose. According to Bachman and Palmer (1996), the overriding purposes of any language tests are to first make inferences about language ability, and secondly to make decisions based on those inferences. Three common types of inferences that a test writer might make are (a) the extent to which learning has taken place at the end of a unit or course of study (*an achievement test*) (b) the particular strengths and weaknesses of a group of students at the beginning of a course so that the teacher will be able to adapt instruction to best meet learner needs (*a diagnostic test*) and (c) the general language abilities of students (*a proficiency test*) which is often used for the purpose of deciding admission to a program or institution.

In order to help ensure that the best type of test is used to make the best possible decision, Brown (2002) recommends distinguishing between tests that are criterion-referenced and those that are norm-referenced. Norm-referenced tests are used to compare the relative abilities of examinees in order to “disperse the performances of students in a normal distribution” (Brown, 2002, p. 2). In order to bring about this normal distribution of scores (also known as the bell-curve) a norm referenced-test aims to include items that are capable of distinguishing among the relative abilities of a broad spectrum of examinees.

The primary purpose of criterion-referenced tests on the other hand, is not to learn how an examinee’s performance compares with that of other test-takers, but to “describe the amount that they know of a specific domain of knowledge or set of objectives” (Brown, 2002, p. 5). As the name suggests therefore, in a criterion-referenced test, an examinee’s performance is referenced against a set of pre-existing external criteria rather than the performances of other test-takers.

Since the purpose of this article is to discuss writing assessment in a classroom setting, the remainder of this article will focus only on the use of criterion-referenced achievement for the purpose of making inferences about students’ ability to perform specific, clearly defined writing tasks.

Educational Setting and Sample Course Objectives

It is not possible to develop an effective test without a very clear understanding of what it is exactly that one wants to assess and how it is to be assessed. It is important therefore that the writing tasks be clearly defined beforehand, not just in terms of the types of tasks students are to perform, but the conditions under

which they will complete them and the standards to which they will be expected to perform them. According to Brown (2002), well-developed criterion-referenced tests can “in a very real sense serve to operationalize a set of course objectives” (p. 30)

Using the familiar SWBAT acronym (students will be able to) Table 1 details objectives for an argumentative essay and an academic summary for a high school EFL writing course that the author developed in collaboration with colleagues at a private secondary school in Tokyo. A brief background discussion of the setting in which these objectives were developed follows.

Educational Setting

Kanto International High School is a co-educational private school specializing in foreign language instruction. Unlike many other high schools in Japan, the majority of English classes are taught by native English speaking teachers who are mostly highly-qualified, many of them having completed Masters Degrees in TESOL or in the process of acquiring one. The objectives here are from a second year honors writing course called the Super English Program. According to the program handbook this course “caters to the needs of students who endeavor to pursue serious English studies at university and beyond” (p. 3). While there are generally very few students enrolled in this program who have spent significant time in English speaking countries (returnees), most of the students are quite highly motivated to study English. A major consideration for this writing course is that students develop the skills to succeed on university entrance exams that contain writing sections. Another notable feature of this teaching context is that the program is highly coordinated both across and within years. Teachers regularly meet to ensure that they are teaching to the same course objectives.

Sample Course Objectives

In the first objective detailed in Table 1 the far left column indicates the writing task that students are to perform. This includes such information as the format of the task (an essay), and the rhetorical mode or text type (argumentation and summary).

Table 1: *Course Objectives for an Argumentative Essay and an Academic Summary*

Task	Conditions	Standards
SWBAT compose an argumentative essay	● Within 50 minutes. ● On a context-enriched topic. ● In approx. 350 words. ● With prior exposure to the topic and some class time spent preparing for the topic.	● appropriately written for an audience of peers and the teacher. ● Displaying only minor lexico-grammatical errors that do not interfere with communicative intent. ● Employing a wide range of cohesive devices. ● Effectively organizing ideas in a clear and logical way. ● Providing sufficient and relevant support for the argument.
SWBAT summarize an academic text	● In an out-of-class course assignment without time constraints.	● appropriately written for an audience of younger peers unfamiliar with the academic text. ● with only minor lexico-grammatical errors ● Using a wide range of cohesive devices ● Effective use of paraphrasing techniques and reported speech

Adapted from the *Super English Program Handbook*, p. 36

Describing objectives in terms of task, conditions and standards is by no means new (See Majer, 1975; Brown, 1995), and it is not without its critics (Valdman, 1975; Tumposky, 1984). Brown (2002) argues that many of the criticisms of detailed instructional objectives are “knee-jerk reactions” stemming from an overly-simplified association of instructional objectives with the operant conditioning of behaviorist traditions in psychology (p.37). It is beyond the scope of this article to consider all the arguments for and against the use of clearly defined instructional objectives. But if the observation that “you get what you assess” (Resnick & Resnick, 1992, as cited in Johnson, et al., 2009, p. 89) is even remotely accurate, then we would do well to be as explicit as possible about what it is exactly we want to get, and as we will see, instructional objectives can play a vital role in helping to make assessment expectations clear. It is worth emphasizing though that course objectives need not be considered rigid, inflexible directives, but “they should serve as the basis for developing classroom tests… and other curriculum elements” (Brown, 2002, p. 37).

One condition for the argumentative essay in Table 1 stipulates that students will have “prior exposure to the topic” and some time in class to prepare for it. It is important to keep in mind when specifying the task in this way that the conditions under which the task is performed could have a knock-on effect to the actual cognitive skill that is being assessed. Though this task requires students to write an argumentative essay, the associated cognitive skill of evaluation is actually not under focus since students will have had the opportunity to prepare their arguments by consulting others and possibly even writing drafts beforehand. If this condition instead specified that students were to write under timed conditions on an unfamiliar topic, the evaluative skill would move back into focus and teachers might be better able to make inferences about students’ ability to

employ this skill. Such considerations reveal the importance of carefully considering the conditions and standards of a task prior to assessment and instruction. (See Anderson & Krathwol, 2001 and Marzano & Kendall, 2008 for a detailed discussion of cognition and knowledge in educational objectives).

One aspect of the objectives included in Table 1 that readers may find unusual is the way in which they are formatted into three separate columns. This way of formatting course objectives has been found to have several practical benefits in this particular teaching context. First, it helps to address one of the major shortcomings of this type of instructional objectives; they are very difficult to write (Richards, 2001). Course planners may find themselves engaged in a sort of syntactic gymnastics when trying to force a string of prepositional phrases describing the various conditions and standards of a task into the parameters of acceptable English grammar. Writing them in columns avoids this problem and provides a clear and easy resource for teachers to consult when preparing classes and designing assessment. This is particularly useful where, as in the aforementioned teaching context, there are a number of teachers working in a coordinated curriculum. This format of objectives is also very easy to revise by just adding, deleting or changing elements in each column. Finally, and most importantly for this discussion is that these types of objectives can provide a blue print of what is expected in terms of assessment. Carefully writing objectives at the beginning of a course can help save a lot of uncertainty at the end of the semester as teachers assess students' assignments or tests and realize that they had not focused enough on a particular aspect of writing throughout the course or not been explicit enough to students about what is expected.

These types of objectives might be referred to as indexical course objectives for the way in which they specify characteristics of a task in a one-by-one, list-like manner that can be easily referenced. The next section will detail some considerations involved in developing a rubric for an argumentative essay and an academic summary based on the two indexical course objectives described in Table 1.

Considerations in Designing Achievement Rubrics

The following section discusses six key considerations in rubric design and development: 1. deciding the type of rubric to use; 2. aligning course objectives with rubrics; 3. trialing rubrics; 4. determining the level of explicitness in a rubric; 5. defining audience and purpose; and 6.developing criterion-referenced scales.

1. Deciding the Type of Rubric to Use

One consideration that confronts a writing teacher when developing an assessment rubric is deciding the type of rubric to use. *The holistic rubric* as advocated by White (1984; 1985) provides a single impressionistic score based on broad bands of performance ability. Probably the most well-known holistic rubrics are the ones used for the independent and integrated writing sections on the Test of English as a Foreign Language (TOEFL®). The obvious benefit of this type of rubric is that it requires very little time to assess essays since

only a single score is reported. However, several studies on the reliability of holistic rubrics used in the North American high school classroom context (Markham, 1976; Sloan & McGinnis, 1982) have shown that such ancillary task requirements as the quality of student handwriting can affect the scores that teachers assign to student scripts. Another disadvantage of holistic scoring is that it is not designed to provide any feedback to writers on their specific strengths and weaknesses. It is therefore better suited to large-scale proficiency tests than to classroom contexts in which feedback is an important part of a writer's development.

Analytical rubrics on the other hand, are capable of providing feedback to writers on specific areas such as content, organization, and grammar. The most well-known of these is the ESL Composition Profile (Jacobs et. al., 1981) which has differential weightings built into a scale with content receiving more weight than other indicators such as language and mechanics. Analytical rubrics like the one used by Jacobs et al. make use of a broad range of criteria (i.e., content, organization) and the same rubric can often be used to assess any number of writing assignments from a narrative to an essay. But in terms of classroom assessment, this versatility of analytical rubrics is also a weakness; many of them are not designed to assess the features specific to the genres students have been asked to emulate and therefore only the most general feedback can be provided.

One type of assessment rubric that is designed to give highly specific feedback to test-takers is the primary-trait rubric developed by Lloyd-Jones (1977) and a variation of this scoring rubric called a multiple-trait rubric advocated by Hamp-Lyons (1991a). A primary-trait rubric clearly identifies a characteristic or "trait" of a narrow range of discourse (e.g. an argumentative essay or an academic summary) and assesses students' ability to complete an assignment in accordance with the characteristics of this trait. The obvious weakness of this type of scoring is the time it takes to develop a new rubric for each writing task. Lloyd-Jones (1977) even estimated that creating a scoring guide takes an average of 60-80 hours per task. Obviously, for most teachers in the EFL contexts who are already burdened with many responsibilities, such an investment of time on a single rubric is not realistic.

The following discussion aims to illustrate how rubrics that are carefully aligned with course objectives and able to provide detailed feedback might be developed in a more reasonable amount of time. Though the rubrics discussed here do not meet all of the requirements of the development process for primary trait or multiple trait rubrics as advocated by Lloyd Jones (1977) or Hamp-Lyons (1991), they do aim to provide fine-grained, genre specific feedback for learners.

2. Aligning Assessment Rubrics and Course Objectives

Table 2 shows a sample rubric for an argumentative essay that was developed to assess the argumentative essay objective in Table 1 (formatting of this rubric was inspired in part by Christianson & Palmer, 2005, as cited in Bachman & Palmer, 2010, p. 347). The rubric was designed to provide extensive genre-specific

feedback and includes a number of indicators, such as topic sentences and support for the argument that are tailored specifically to the argumentative rhetorical mode. Each of these indicators also corresponds to standards in the far right column of the objective in Table 1. One of the standards in Table 1 that states "appropriately written for an audience of peers and the teacher" is addressed in the rubric with references to reader engagement in the introduction section of the rubric, including "an inviting hook" and "suitable background information to orientate the reader".

There is also a section of this rubric devoted to "overall cohesion and accuracy", reflecting the objective (Table 1) that the essays contain only minor lexico-grammatical errors that do not interfere with communicative intent. It would of course be possible to remove this section on cohesion and accuracy from the rubric, since language-use could be measured implicitly in the students' ability to effectively perform the task. In this particular case however, one of the conditions of the objective in Table 1 states that students will write on a context-enriched, or familiar topic in which they can draw on their own experiences to complete the task. Thus, a primary purpose of this assignment was for students to gain control of the linguistic and rhetorical features of the argumentative essay genre within the comfort of a familiar topic before moving onto a context reduced topic such as the environment or a social issue. Since language-use was intended to be a central focus during the instructional process, it was included in the rubric as a discrete category.

Table 2: *A Sample Argumentative Essay Scoring Rubric*

Argumentative Essay Scoring Rubric						
	C	E	M	L	A	Teacher's Comments
Introduction Has an inviting hook that engages the reader	8	6	4	2	0	
Contains suitable background information to orientate the reader	8	6	4	2	0	
Has an effective thesis statement that focuses the reader	4	3	2	1	0	
Body Includes appropriate topic sentences that signal main ideas	8	6	4	2	0	
Provides appropriate examples and details to support the argument	20	15	10	5	0	
-Logical organization of examples and details	16	12	8	4	0	
Conclusion -Restates thesis in an interesting way	4	3	2	1	0	
-Closes the essay in a logical manner	8	6	4	2	0	
Overall Cohesion and Accuracy Effectively uses of a variety of transitional expressions	12	9	6	3	0	
Uses a variety of grammatical and lexical structures	12	9	6	3	0	
Scoring Guide **C**omplete= All Features Present **E**xtensive= Most Features Present **M**oderate= Some Features Present **L**imited = Few Features Present **A**bsent= No Features Present						**Total:** **/100**

But whether language-use is addressed implicitly or explicitly in the rubric, it is important to have clear descriptors that are able to discern how effectively the task was completed. It is usually not very informative to learn simply whether or not the task has been completed. In a rather amusing example, Brown (2002) observes that: “After all, there are many ways of carrying out the task of acquiring bread at a supermarket--some involve language and some involve weapons” (Brown, 2002, p. 23).

3. Trialing the rubrics before administering the writing task

Since the rubric shown in Table 1 was designed for what Crusan (2010) calls “the little **a**” of classroom assessment rather than “the big **A**” of high-stakes standardized assessment, the method in which it was developed was decidedly less rigorous than that which would be necessary for more high-stakes tests such as placement or proficiency tests. There were no attempts to use statistical analyses to ascertain the reliability among raters using the rubric. Employing such statistical analyses for each classroom writing assignment would be too time-consuming in most teaching contexts. This is not to downplay the powerful role that statistics can play even in everyday classroom assessment. Brown, (2002) describes in a highly accessible way a number of easy-to-use statistical techniques that can be used to improve the validity and reliability of criterion-referenced tests. But in a course that aims to provide finely-grained, genre-specific feedback for each writing assignment, such analyses can be extremely time-consuming.

One practical way to help ensure that rubrics are making the type of inferences they were designed to make is to trial them on student scripts as much as possible before administering them. When trialing a rubric, sample essays might be used from diagnostic pre-tests or previous achievement tests. Whenever possible it is also a good idea to involve colleagues in the trialing process. Many teachers in the secondary school context discussed here share the burden of marking large numbers of essay exams with their colleagues, and take part in “norming sessions” to increase inter-rater reliability. This process has promoted rich discussion not only about issues regarding grading, but problems with using the rubrics. Unfortunately however, by this time it is often too late to modify the rubric to accommodate teachers’ suggestions. Of course the techniques for determining the validity and reliability of any rubric will depend on a variety of institutional factors including the availability of time and resources. Whenever possible however, teachers could benefit from even very brief sessions when developing rubrics for classroom tests or assignments.

4. Determining the level of explicitness in assessment rubrics

When developing a classroom achievement test, (and arguably any test) it is important to consider not only reliability and validity, but the *instructional value* of the test. As Hamp-Lyons (1991a) points out when discussing large-scale L2 writing assessments: “any method of testing which fails to utilize the educative potential of the test itself permits a disjunction between teaching and assessment” (p.244).

One way to increase the educational potential of a rubric is to be as explicit as possible about what is required to successfully complete the writing task. Crusan (2010) strongly advocates “transparent assessment” practices (p. 33) and argues that “if you want to count something [when assessing a writing assignment], then you need to tell your students what counts” (Crusan, 2010, p. 59). A cavalier, gut-referenced approach to assessment provides little direction for students trying to improve their writing—or accountability for teachers assessing it. Developing a detailed rubric that is clear about how students are to be assessed and distributing it well in advance of the exam or assignment due date, are simple ways to increase transparency. (See Crusan, 2010 for additional suggestions for communicating assessment expectations to students).

It is important however, to distinguish between being *highly explicit* in an assessment and being *overly-prescriptive*. The former implies complete openness about how students will be graded, while the latter means imposing too many constraints on what an acceptable response should entail. With the dual aims of many writing teachers in the Japanese EFL context to simultaneously prepare students for tests that contain discreet point grammar questions while developing their academic writing skills, there may be a temptation for teachers to prescribe the use of certain target grammatical structures such as present perfect in an introduction to an essay or the third or fourth conditional in the body of the essay. While well-intentioned, the prescription of specific linguistic items in the test or assignment could very well give students a wrong impression of composition as a kind of large-scale cloze exercise. Likewise, some well-meaning teachers may prescribe on writing rubrics the use of specific signal words that have been taught in the course. This risks rendering a rubric to the role of a linguistic checklist in which students are forced to focus on the inclusion of the required words rather than communicative potential of a writing task.

It is also possible to be overly prescriptive on the rubric in terms of larger stretches of discourse such as the number of sentences in a paragraph or the number of paragraphs in an essay. Carson (2000) recommends avoiding specifying length in terms of sentences or paragraphs because the structural units will emerge from how students approach the task. If there is any doubt that such over-specification of structural units can have a highly negative impact on students’ views of writing, a colleague recently mentioned that one of his advanced university writing students in Japan was baffled by the sample essay that he distributed because her composition teacher had told her paragraphs have five sentences and this one didn’t!

A general rule that can be used when deciding what language or rhetorical constraints to impose on a task is whether they can be considered optional or obligatory features of the writing task. In other words, can the writing task be effectively completed without this specific linguistic or rhetorical feature? It is not possible to complete an essay without making use of verb forms, prepositions or vocabulary, so depending on the purpose of the assessment, these may very well be included in the rubric as they are in Table 2. But it is perfectly conceivable to write an essay without the use of present perfect and it is therefore an optional linguistic resource that should not be included in the rubric. Likewise, while it is not easy to imagine an essay without

an introduction, body and conclusion (See Table 2), the way in which these stages of discourse are realized should not be prescribed in terms of the number of paragraphs. Just as it is possible for a student to write two, three or four paragraphs in the body of an essay, it is also conceivable that a student might choose to break the introduction into several paragraphs. In summary, a rubric should be explicit about the features of a successful response while allowing the writer enough rhetorical space to approach the task in a novel way.

5. Addressing Audience and Purpose in Scoring Rubrics

Written genres like the argumentative essay or the academic summary are not just static sets of rhetorical patterns but are dynamic, socially-embedded resources that fulfill specific social purposes. As Paltridge (2006) puts it, they are "the ways in which people get things done through their use of language in particular contexts" (Johns et al., p. 235). One potential danger of scoring rubrics is that writers, in their attempts to meet all rubric requirements, may lose sight of the fundamentally social and dialogical nature of academic discourse. (Bazerman, 1994; Bhatia, 2004; Hood, 2010; Martin, 2005; Swales, 1990; Tardy, 2009).

Some students may come to treat composition as a process of assembling various parts of a writing task detailed in a rubric (i.e., an introduction and topic sentences) with little consideration of how to effectively engage readers. Simply telling students to include an interesting or surprising fact in their introduction or to anticipate possible objections to their arguments may be of little help to conscientious writer who point out that "*the way to write an essay depends on who's reading it*". What may be considered interesting or surprising to a teacher reading a students' essay about the negative aspects of school uniforms may be common knowledge and thus of little interest to a readership of peers. Likewise, views or opinions about school uniforms that might be taken for granted and beyond reproach for a group of students may require more elaboration and justification when written for a teacher (See Martin 2005 for a detailed analysis of the types resources used to engage a readership). Being explicit about the purpose and audience of a writing task not only helps to provide focus for the writer, but reinforces the notion that academic discourse, like other forms of communication, is socially situated.

Table 3 shows a sample academic summary scoring rubric that includes highly specific references to audience and purpose in the descriptors. Though many L2 writing assessment experts (e.g. Crusan, 2010; Weigle 2002) highlight the importance of clearly defining the audience of a writing task, it is somewhat unusual for such detailed considerations to be included in the actual rubric task descriptors.

Table 3: *A Sample Academic Summary Scoring Rubric*

Academic Summary Scoring Rubric						
	C	**E**	**M**	**L**	**A**	**Teacher's Comments**
-The summary is effectively written for an audience unfamiliar with the topic of the academic text.	8	6	4	2	0	
-The academic text is appropriately paraphrased for an audience with limited English ability	8	6	4	2	0	
-All main ideas of the academic text are clearly referenced	12	9	6	3	0	
-Appropriate reported speech devices are used to refer to the author's ideas	8	6	4	2	0	
- Effective use of a variety of transitional expressions	4	3	2	1	0	
-There are no unnecessary details	4	3	2	1	0	
-Language is used accurately	4	3	2	1	0	
Scoring Guide **C**omplete= All Features Present **E**xtensive= Most Features Present **M**oderate= Some Features Present **L**imited = Few Features Present **A**bsent= No Features Present						**Total:** **/48**

The decision to foreground audience in this way came about after observing that the summaries students were submitting were not meeting teachers' expectations, which led the author to conclude that students did not fully understand the purpose of summary writing. An earlier rubric that was used to assess the academic summary required only that students paraphrase all the main ideas from the text, that they use reported speech to refer to the author's ideas, and that it be cohesive and accurate. The results unfortunately were often lists of sentences containing main ideas copied from the text separated by signal words and such phrase as "the author states that". In other words, students merely spliced together the required elements or structural units described in the rubric with little attention to the fundamental social purpose of a summary, to succinctly communicate the main ideas of a text to save the reader the time of actually reading it. In subsequent years, course objectives, instructional practices and rubrics were revised to ensure that purpose and audience occupied a more central role in instruction and assessment.

In this rubric the audience is defined not as the teacher or peers but as people who are "unfamiliar with the topic of the academic text" and have "limited English ability". When the assignment was distributed, the audience was further delineated as a group of students in the previous year who would be reading the texts that the summaries are based on when they become 2nd year students. Thus, the purpose of the writing task was to help younger peers understand what types of texts they would be reading in the following year. Providing a purpose for writing and identifying a real group of people to write for, helped to show that summary writing can serve a real and meaningful social purpose. The fact that students had to write for their younger peers also helped to make paraphrasing the text a more meaningful exercise; it was not just a

question of substituting different words to avoid plagiarism, but selecting words that their audience is more likely to understand.

Though it is often very difficult in the EFL context in Japan to find another audience for students to write to other than the teacher, the junior-senior (*kohai,-sempai*) relationships in Japanese secondary schools and universities represent one opportunity for teachers to vary audiences (See Myskow & Gordon, 2010 for other suggestions of audience variation at the senior secondary school context in Japan).

6. Developing Criterion-referenced Scales

Another aspect of these rubrics that also deserves mention is the scoring categories and weightings. In keeping with Politt's (1990) observation that it is overly-optimistic to expect a test of writing to reliably distinguish between 5 point scales or more, the criterion-referenced scale used in both the summary and essay rubrics differentiate between no more than five categories of performance. At the top of the rubrics in Tables 2 and 3 is the acronym CEMLA which is defined in the scoring guides at the bottom of each rubric as "complete", "extensive", "moderate", "limited" and "absent" to describe the extent to which target features of the genres are evident in the students' scripts. This encourages teachers to engage in a process of actually "reading for evidence" of a specified criterion. (Gordon, 1999, as cited in Johnson, et al., 2009, p. 205) Vague assessment categories such as "excellent", "good" and "poor" fail to clarify what is representative of each of these categories. While the process of reading for evidence of specific criteria by no means eliminates these uncertainties, it does prime assessors to focus on what is present or absent in a paper rather than employing their personal value judgments to determine the extent to which they feel certain features are appropriate or not.

The question of course still remains for teachers to reliably differentiate between scale categories. In order to avoid what Crusan (2010) calls the personal "rubric in the head" of teachers, she suggests that teachers develop a clear understanding of what type of responses represent what level of criterion prior to assigning any scores (p. 50). For instructional purposes, it is recommended that the differential criterion levels be determined before instruction starts, so that teachers are keenly aware throughout the course of the levels of responses and the types of problem areas they expect. This can be done by identifying anchor papers that represent different levels of responses and referring to these throughout the course and marking sessions. Possible sources of these anchor papers could be the pre-course diagnostic tests or achievement tests and assignments from students in previous years.

These anchor papers can also play a valuable role in instruction. Crusan (2010) strongly encourages teachers to share these papers with students prior to the assessment so that they too have a clearer idea of what is expected of them. This will of course depend on the availability of papers that illuminate the levels of criteria, and the extent to which they are sufficiently different from the actual writing prompt so that students

do not borrow extensively from the highly-rated papers.

One final point about the use of criterion scales is the means by which weightings and composite scores are determined. In the argumentative essay rubric in Table 2 the total composite score is 100, while in the academic summary rubric in Table 3 the total score is 48. The total scores are not important, but many students may like to receive a score that is out of 100. This can be easily done by converting the number to a percentage and writing it at the top of the paper. It is more important that test writers focus on the rationale for the weighting of sub-skills rather than whether or not the total number of points adds up to a desired number.

Determining the comparative weightings of each sub-skill however is not an easy task. For high-stakes tests, many experts recommend avoiding a composite score and reporting the various scores separately—only when absolutely necessary arriving at a composite score through consultation with a statistical expert (Bachman, 2010; Hamp-Lyons, 1991a). Needless to say most teachers do not have access to in-house statisticians to help them calculate composite scores. It is recommended therefore that considerations in weighting sub-skills plays a central role in the rubric trialing process discussed above.

Considerations in writing task instructions and prompts

Writing Task instructions

Bachman and Palmer (1996) recommend three basic guidelines for writing task instructions: they should be simple enough to understand; they should be short enough to not take up too much time; and they should be detailed enough for test takers to know what to do. Weigle (2002) also suggests that the instructions include an indication of audience and purpose. Note that these considerations of audience and purpose may also be specified in the actual rubrics as they are in Table 3 & 4.

In order to make the instructions comprehensible to EFL students, it may be necessary to write them in their native language. Of course this also applies to information in the rubrics and in some, albeit rare situations, possibly even the prompt itself. When administering a writing test or assignment, it is most important that a teacher is able to use the piece of writing to make inferences about the extent to which students achieved course objectives. Introducing such “ancillary task demands” (Haertel and Lynn, 1996, as cited in Johnson, et al., 2009, p. 65) as the ability to understand instructions that may be written at a level higher than the writing task itself, may hinder a teacher’s ability to make such inferences. Of course it may be an instructional objective for students to develop such test literacy skills as interpreting task instructions in the target language. In which case it is recommended that ample time be spent on this skill beforehand.

Considerations in designing writing prompts

Crusan (2010) states that “a good prompt consists of all the information students need to develop an

appropriate response to the assignment" (p. 68). One important consideration is developing a context for the prompt. Depending on how much context is needed for students to clearly understand the prompt, descriptions could vary from 1-2 sentences to a full paragraph. The following prompt developed for the argumentative course objective (Table 1) contains only three simple sentences detailing the context. The prompt below also specifies the format (an essay) and the mode (argumentation).

Example 1. Sample Argumentative Essay Prompt

Some people think that the school week in Japan should only be five days. Others think that five days is not long enough. They think that the school week should be six days. Which view do you agree with? Write an argumentative essay explaining your opinion. Be sure to use specific reasons and details to support your opinion.

Prompts such as the following that are too open-ended and do not specify the genre that students are to write may result in responses that do not meet the task requirements:

Example 2. A Poorly Constructed Argumentative Essay Prompt

English education in Japan should start in Kindergarten. Discuss

Prompts that contain idiomatic language that the students may not be familiar with are also potentially problematic. In Example 3 for example, the word "pointless" is likely to be interpreted by students in any number of ways.

Example 3. Another Poorly Constructed Argumentative Essay Prompt

Some students feel that PE classes are pointless. Others feel that they are not.

Though this article focuses mainly on writing assessment in the context of in-class examinations, it is important to consider whether or not a writing task even needs to be done under timed conditions at all. In the case of the argumentative essay, it was thought best that the task be completed under timed conditions since one of the major goals of the course was to help students prepare for university entrance examinations that contain writing prompts. But as Hyland reminds us: "there is no need for students to write under timed writing conditions unless those are the conditions under which the genre is used in real life" (p. 165).

Conclusion

By way of extended example of an argumentative essay and an academic summary writing task, this

article has aimed to share with readers some of the many issues and considerations that have arisen when assessing writing in a specific Japanese EFL context. Unfortunately, the broad scope of this article has prevented any particular consideration to be treated in detail. It is my hope that the article generates some discussion that will build on (and improve) the suggestions offered here. But however assessment issues are addressed in particular contexts, we would do well to consider Hamp-Lyons (1991b) appeal that "we keep always at the center of our attention the well-being of those writers who take our tests and whose lives are affected, sometimes greatly, by the results" (p. 328).

Bibliography

Anderson, L., & Krathwol, D. (2001) *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of Educational Objectives*. New York: Longman.

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

Bachman, L.F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L.F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In A. Freedman, & P. Medway (Eds.), *Genre and the new rhetoric* (pp. 79-101). London: Taylor & Francis, Ltd.

Bhatia, V.K. (2004). *Worlds of written discourse: A genre-based view*. London: Continuum.

Brown, J.D. (1995). *The elements of language curriculum*. Boston: Heinle & Heinle.

Brown, J.D., & Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

Carson, J. G. (2000). Reading and writing for academic purposes. In M. Pally (Ed.), *Sustained content teaching in academic ESL/EFL* (pp. 19-34). Boston: Houghton-Mifflin.

Carson, J. G., Chase, N.D., & Gibson, S.U. (1992). Literacy demands of the undergraduate curriculum. *Reading Research and Instruction*, *31*(4), 25-50.

Crusan, D. (2010). *Assessment in the Second Language Writing Classroom*. Michigan: The University of Michigan Press.

Hamp-Lyons, L. (Ed.). (1991a) Issues and directions in assessing second language writing in academic contexts. In L. Hamp-Lyons (ed.), *Assessing second language writing in academic contexts* (pp. 323-329). Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L., (Ed.). (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (ed.), *Assessing second language writing in academic contexts* (pp. 241-266). Norwood, NJ: Ablex Publishing Corporation.

Hood, S. (2010). *Appraising research: Evaluation in academic writing*. London: Palgrave Macmillan.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.),

Assessing second language writing in academic contexts (pp. 71-85). Norwood, NJ: Ablex Publishing Corporation.

Hyland, K. (2004). *Genre and second language writing*. Michigan: University of Michigan Press.

Jacobs, H., Zinkgraft, S., Wormuth, D., Hartfiel, V. & Hughey, J. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Johns, A. M., Bawarashi, A., Coe, R.M., Hyland, K., Paltridge, B., Reiff, M.J., & Tardy, C. (2006). Crossing the boundaries of genre studies: Commentaries by experts. *Journal of Second Language Writing*, *15*(3), 234-249.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance*. New York, NY: The Guilford Press.

Kanto International High School *Super English Program Handbook* (2009)

Leki, I., Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in different university courses. *TESOL Quarterly, 31*(1), 39-70.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper and L. Odell (eds.), *Evaluating writing* (pp. 33-69). NY: National Council of Teachers of English.

Majer, R.F. (1975). *Preparing instructional objectives*. Belmont, CA: Fairon-Pitman.

Marzano, R., & Kendall, J. (2008). *Designing and assessing educational objectives: Applying the new taxonomy*. Thousand Oaks, CA: Corwin Press.

Markham, L.R. (1976). Influence of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13*(4), 277-283.

Martin, J. R. & White, P.R.R. (2005). *The language of evaluation: Appraisal in English*. London: Palgrave Macmillan:

Myskow, G. & Gordon, K. (2010). A focus on purpose: Using a genre approach in an EFL writing class. *ELT Journal*, *64*(3), 283-292.

Politt, A. (1990). Response to Charles Alderson's paper: 'Bands and scores.' Alderson, J.C. (1991). In J.C. Alderson and B. North (Eds.) *Language Testing in the 1990s: The communicative legacy* (pp. 87-91). London: Modern English Publications/ British Council/ Macmillan.

Richards, J.C. (2001). *Curriculum development in language teaching*. Cambridge: Cambridge University Press.

Rothery, J. (1986). Teaching genre in the primary school: a genre-based approach to the

development of writing abilities. In *Writing project-report 1986* (pp. 3-62). Sydney: University of Sydney, Department of Linguistics.

Sloan, C., McGinnis, I. (1982). The effect of handwriting on teachers' grading of high school essays. *Journal of the Association for the Study of Perception, 17*(2), 15-21.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge

University Press.

Tardy, C.M. (2009). *Building genre knowledge*. West Lafayette, Indiana: Parlor Press.

Tumposky, N.R. (1984). Behavioral objectives, the cult of efficiency and foreign language learning: Are they compatible? TESOL Quarterly, *18*(2), 295-310.

Valdman, A. (1975). On the specification of performance objectives in individualized foreign language instruction. *Modern Language Journal*, *59*(7), 353-360.

Weigle, S.C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S.C. (2006) Investment in assessment: Designing tests to promote positive washback. In P. Matsuda (Ed.), *The politics of second language writing*. (pp. 222-244). West Lafayette, Indiana: Parlor Press.

White, E.M. (1984). Holisticism. *College Composition and Communication*. *35*(4), 400-409.

White, E.M. (1985). *Teaching and assessing Writing*. San Francisco, CA: Jossey-Bass.

L2 ライティング評価：日本 EFL の高等学校の場合

ゴードン・ミスコウ*

この論文は、L2 ライティング評価の論考を日本の EFL 高等学校の評価の実践に位置付ける事を目的としている。論文は L2 ライティング評価を作る際に鍵になる文脈上の事実、テストの目的を論じる事から始まる。そしてその後２つのコース目標の発展を説明し、また学生がこのコース目標を達成したかどうかを評価するためのルーブリックとプロンプトを作成する際の考察を論じる。この論文は初等教育の評価に集中しているが、多数の考察は相関のある状況の教師にとっても関連性があると考えられる。

キーワード：日本 EFL、ライティング、クラス内評価、コース目標

＊　人間科学総合研究所研究員・東洋大学経済学部