Title	Basic considerations in designing an effective L2 writing test
Sub Title	ライティングテスト作成の基礎
Author	中村, 優治(Nakamura, Yuji)
Publisher	慶應義塾大学日吉紀要刊行委員会
Publication year	2007
Jtitle	慶應義塾大学日吉紀要. 言語・文化・コミュニケーション (Language, culture and communication). No.39 (2007.) ,p.1- 10
JaLC DOI	
Abstract	
Notes	
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN10032394-20071220-0001

Basic considerations in designing an effective L2 writing test

Yuji Nakamura

1. Introduction

The ability to write effectively has become an essential tool for us in today's global community. It is one of the one major means that allows for the ease of communication among individuals of differing cultures and backgrounds. Consequently, instruction in writing is assuming an increasing role in both second- and foreign-language education (Weigle, 2002).

When the acquisition of a specific language skill is seen as important, it becomes equally important to test that skill, and writing is no exception. Accordingly, as the role of writing in second-language education increases, there is an ever greater demand for valid and reliable ways to test writing ability. It is generally accepted that a test of writing involves at least two basic components: one or more writing tasks, or instructions that tell test takers what to write, and a means of evaluating the writing samples that test takers produce (Weigle, 2002). However, there seems to be more to know about the construction of a writing test.

To establish the scope of writing assessment, Weigle (2002) gives a comprehensive overview by asking seven questions (p.2): 1) What are language testers trying to test? 2) Why do language testers want to test writing ability? 3) Who are the test takers? 4) Who are the raters and what criteria or standards are used? 5) Who will use the information obtained from the tests? 6) What are the constraints? 7) How can language testers make the test valid and reliable?

Bachman and Palmer (1996, p.17) claim that the most important consideration in designing and developing a language test is the use for which it is intended, thereby designating its usefulness as the most important quality of a test. Their definition of test usefulness includes the following six qualities: reliability, construct validity, authenticity, interactiveness, impact and practicality. Since the relative importance of these six qualities varies from situation to situation, they suggest that

test developers should strive to maximize overall usefulness in individual situations (Bachman and Palmer, 1996).

Cambridge ESOL's traditional approach—the VRIP approach—is also useful. It stands for Validity (the conventional sources of validity evidence: construct, content, criterion), Reliability, Impact and Practicality (Shaw and Weir, 2007).

According to Messick (1989), the construct validation process should include both collection of empirical evidence and a theoretical rationale. Chapelle (1998, p.51) discusses five types of evidence for construct validity: 1) content analysis, 2) empirical item investigation, 3) task analysis, 4) relationships between test scores and other measures, and 5) experimental research identifying performance differences over time, across groups and settings, and in response to experimental interventions. Weigle (2002, p. 51) claims that construct validity must be demonstrated in at least three ways: 1) the task must elicit the type of writing that we want to test; 2) the scoring criteria must take into account those components of writing that are included in the definition of the construct; and 3) the readers must actually adhere to those criteria when scoring writing samples.

2. Purpose of the paper

This paper focuses on construct validity, which is important and the core among these qualities in the test construction, as well as discussing other qualities in relation to construct validity. By taking into account Weigle's (2002) ideas, the paper discusses what we should consider when we make decisions about designing writing assessment tasks or scoring procedures for a better judgment of students' writing ability. Among the issues discussed will be: 1) the definition of the construct or the definition of writing ability, 2) the tasks (the test tasks and the response tasks), 3) the scoring (the rating) and the raters.

2.1. The definition of the construct of "writing ability"

First of all, we need to take into consideration the definition of writing ability. What do we mean by writing ability, i.e. what is the theoretical construct of writing? We will look at the construct from four viewpoints: a) the nature of writing, b) linguistic theory, i.e., applied linguistics, second language acquisition theory, and psycholinguistic theory, c) the test format of existing tests, and d) the connections between writing and other language skills, especially speaking and reading.

The nature of writing ability has changed with the advancement of technology—from traditional writing ability to technology-mediated writing ability

(cf. Chapelle and Douglas, 2006). The nature of writing can be characterized as the ability to get the message across, the ability to use linguistic knowledge to express our ideas, and the ability to organize the ideas logically. However, we must also consider the new options opened by testing through technology. The development and use of computerized tests challenges and expands the imagination for writing. In other words, it can offer more authentic testing settings in which test takers can feel as if they are writing in a real world. Therefore, we need to reconsider the nature of writing that is called upon in technology-mediated interactions and communication. Thus we need to rethink test construct (cf. Chapelle and Douglas, 2006).

Approaching the definition of writing theoretically is another way to outline the component of language proficiency. Buck (1994) claims that during test construction, a number of theoretical decisions must be made. Weir (1993) also maintains that the test should be theory driven. Thus, the theoretical part of test construction is an obvious starting point. Second language acquisition theory has added new aspects to the component of writing ability. For example, Carson (2001) has approached L2 writing in terms of second language acquisition.

The common commercially standardized tests have provided new evaluation items as well as the new facet of writing ability. One innovation of the new iBT TOEFL is that it includes tasks that resemble those performed by students in North American colleges and universities. These "integrated skills" tasks require test takers to use the information provided in the reading and listening passages in essay and/or spoken responses, in addition to the independent task in which test takers are asked to write an essay under a given topic (cf. Jamieson, 2005). The integrated task is scored on the quality of writing (organization, appropriate and precise use of grammar and vocabulary) and the completeness and accuracy of the content, while the independent task is scored on the overall quality of writing (development, organization, and appropriate and precise use of grammar and vocabulary).

The idea of connections between writing and speaking is a current issue in performance assessment (cf. Nakamura, 2003; Weigle 2002). Nakamura (2003) suggests the necessity of two-dimensional performance assessment involving both speaking and writing, although two types of performance tests can share some rating items. Weigle (2002) claims that though speech and written discourse draw on many of the same linguistic resources and can be used in many cases to meet the same communicative goals, writing differs from speech in a number of important ways, both in terms of textual qualities and in terms of the factors that

govern the uses of each modality. Weigle (2002) maintains that written language is a distinct mode of communication, involving among other things very different sociocultural norms and cognitive processes.

The construct of writing has been changing along with such factors as the nature (e.g the advancement and the use of technology), the theory(e.g the idea of second language acquisition findings), and the existing tests (the employment of the change of the revised standardized tests).

2.2. Test tasks and response tasks

We should take into account the test tasks or the response tasks. In other words, how is the test conducted? As a task component we can think of either one or more writing tasks, or instructions that tell test takers what to write (Weigle, 2002). There have been independent tasks where test takers are required to respond to a single topic. As Hughes (2003) claims, tests should check one ability in one test and tests should give test takers as many fresh starts as possible. It has recently been suggested that test takers be required to listen and write, or read and write in order to accurately assess their combining of these skills (cf. Cumming, 2005).

Today language tests are different from traditional tests mainly because they are administered on computer rather than with paper-and pencil or with an interviewer, and because many of them make use of computer technology to branch test takers to different subsets of items or tasks. These tests can be adaptive or linear, and they can be administered via the web, CD, or a network. Each method of delivery offers potential benefits and problems. However, many of these tests do not provide us with an alternate construct of language ability. So, although the tests are innovative in terms of technology, they are not particularly innovative in their operationalization of communicative language ability (cf. Jamieson, 2005).

One innovation of the new iBT TOEFL is that it includes "integrated skills" tasks that require test takers to use the information provided in the reading and listening passages in essay and/or spoken responses. The same passages also serve as the input for reading and listening comprehension questions. This use of more authentic tasks provides evidence for the representation inference in the validity argument for the new TOEFL. It has also forced language testers to grapple with the traditional, yet conflicting, inferences of performance on language tests between underlying ability and task completion (Jamieson, 2005). Authentic tasks in general require test takers to complete tasks. For example, they read a passage and write an essay. In this case, we wonder where the results come from— their

reading ability or their writing ability.

Cumming et al. (2005) show that the discourse produced for the integrated writing tasks (involving writing in response to print or audio source texts) differed significantly from the discourse produced in the independent essay for the variables of lexical complexity (text length, word length, ratio of different words to total words written), syntactic complexity (number of words per T-unit, number of clauses per T-unit), rhetoric (quality of propositions, claims, data, warrants, and oppositions in argument structure), and pragmatics (orientations to source evidence in respect to self or others and to phrasing the message as wither declarations, paraphrases, or summaries).

Moreover, this decision to include integrated tasks resulted in the violation of the assumption of the Item Response Theory (IRT) that tasks on a single scale are conditionally independent; and without IRT, the necessary psychometric base for calibrating items for computer-adaptive delivery was lost. The current need for human raters of speaking and writing tasks also precluded the use of computer-adaptive format. The decision to develop computerized tasks that better represent authentic language use rather than to be constrained in task development by relying on known technology and psychometrics marks a new direction in large scale, high-stakes testing (cf. Jamieson, 2005).

Here is an outline of the different tasks (Independent and integrated).

Independent task
—given topic and writing
Integrated task
—1) integrated reading-writing task
—2) integrated listening-writing task
—3) integrated reading, listening-writing task

2.3. The scoring (the rating) and the raters.

We should consider how the scoring procedures, which include the rating criteria, the rating, and the raters, are conducted. In other words, how is the rating administered? Is it in a holistic way or in an analytic way? (cf. Weigle, 2002; Nakamura, 2004). Is the test rated by computers (machine rating) or by human raters (human rating)? How are they separated or combined? (cf. Chapelle and Douglas, 2006). And what are the rater characteristics?

2.3.1. Rating criteria

Rating is a very important element in writing. We need to consider whether it should be holistic or analytic depending on the test purpose or the testing context. For example, for the diagnostic purpose we need detailed results which require analytic rating, whereas for the placement purpose when we have time constraints we usually tend to use holistic rating. Holistic ratings use overall band scales such as 5,4,3,2,1. and each level should have a descriptor for high inter-rater reliability. The problem is that there is an assumption that a leaner has a unified ability at each level. Analytic ratings use rating items decided in advance, such as grammar, vocabulary, coherence. The rating system is more reliable and diagnostic, although it takes time

Holistic, analytic, or impressionistic methods of rating written compositions have featured in tests and examinations (Cumming, 2007). These methods, through descriptive criteria, rating scales, and the specification of the content of such assessments have been implemented through the training, monitoring, and moderating of raters to ensure their reliability in scoring (Cumming, 2007).

Nakamura (2004) examines the strengths and weaknesses of holistic and analytic scoring methods, using the Weigle adaptation of Bachman and Palmer's (1996) framework, which has six original categories of test usefulness, and explores how we can use holistic or analytic scales to better assess student compositions.

It is also important to decide the weighting of rating items. The reason for this is that each individual item makes a different contribution to the whole writing ability.

2.3.2. Rater issue

An automatic scoring machine like the E-rater produced by Educational Testing Service has particular features which are based on four kinds of analysis: syntactic, discourse, topical, and lexical. Since essay length was found to be "the single most important objectively calculated variable in predicting human holistic scores" essay length as measured by number of words was included in e-rater's updated feature set (Jamieson, 2005). Although the trend of using machine rating is controversial among writing teachers, these automated systems have none the less reported high correlations with human raters (Weigle, 2002; Jamieson, 2005).

Traditionally, inter-rater reliability has been the main focus in this area. Recently, the idea of raters' rating characteristics (rater bias, rater severity etc.) have been considered for rater training using the FACETS model (Engelhard, 1994; Nakamura, 2007) or adopting the idea of think aloud protocols (Lumley, 2005).

Nakamura (2002), for example, suggests a paired-rating procedure with high inter-rater reliability to minimize the labor constraint.

Cumming (2007) maintains that descriptive criteria, rating scales, and benchmarks etc. are just the tools that guide writing assessments. He further claims that the actual practices of writing assessments occur through raters' individual interpretations and judgments while they score written texts. Cumming (2007) refers to Connor-Linton (1995) by saying that we must look into rater's minds to examine what composition assessments really involve.

Lumley (2005) deals with this rater problem in detail. He claims that an important aspect of investigating validity is concerned with how the process is managed. His major goal of the study was to examine and attempt to describe the relationship between the simplicity of the scale and the complex judgements of the raters. Although, as Cumming (1997) notes, the investigation of the rating process, and the basis of raters' decisions is still at a preliminary stage, Lumley is able to provide insights into the complexities, nuances, difficulties, and contingencies of composition assessments (Cumming, 2007).

Among the many findings of Lumley's study (2005), three are useful for this present paper. In the first place Lumley (2005) concludes there is a single fundamental process that raters follow during analytic rating: a first reading (reading and thinking about the quality of the text), a scoring stage (giving scores and justification in relation to categories), and a third stage (score revising and confirming).

Another finding came through Lumley's attempt to deal with the nature of the rating process. Lumley (2005) investigates the validity of a test by examining the validity of the rating process. He says rating (including reading) is at one level a rule-bound, socially governed procedure that relies upon a rating scale and the rater training which supports it, but it retains an indeterminate component as a result of the complexity of raters' reactions to individual text. In spite of the tension and indeterminacy, rating can succeed in yielding consistent scores provided raters are supported by adequate training. Furthermore, Lumley (2005) also claims that the role of the rating scale, and especially its analytic categories, is to help the raters justify rating decisions, rather than to describe test performances. Lumley (2005) argues that the validity of writing assessments resides more in the complex cognitive processes of human scoring than the necessarily underspecified, descriptive criteria and scales that conventionally define a writing test.

Using think aloud protocols, introspective research techniques to reveal the rating process, Lumley illuminates details of the thinking processes of four

experienced composition raters. Although verbal reports have limitations and they are descriptive, idiosyncratic and contingent, Lumley's (2005) study still can provide insights about how raters rate writing internally.

Nakamura (1996) used the technique of verbal report by the raters and questionnaire analysis in a speaking test: 1) to gather information on how raters actually made their choices on their rating sheet of students' speaking ability, 2) to determine what criteria teachers actually use, as well as those they think they use in rating students' speaking ability, and 3) to make the best use of the results for the training of raters, for both classroom teaching and classroom learning. The result of the study shows that raters should realize that they need to share with the students either the topic or the context. This would allow the students to expand during the testing situations and in general, help them to learn how to speak more comfortably. Also, students (test takers) should be reminded that they can be listener-friendly as well by trying to make themselves understood in English with a clear and audible tone in addition to being more open to subject area change. Co-operative attitudes between the two sides will make the communication easier and the judgement more accurate. Although this result is based on a speaking test, it can give us useful suggestions for writing assessment, because both writing and speaking are categorized in a performance test where test-takers perform and raters are involved in scoring.

3. Conclusion

The issues discussed were: 1) the definition of the construct or the definition of writing ability, 2) the tasks (the test tasks and the response tasks), 3) the scoring (the rating) and the raters.

Unlike the usual or traditional way of looking at the test issue, the present paper has examined exclusively the validity with special focus on raters instead of reliability and practicality. In doing so, it has emphasized the importance of the raters in the performance test even for the consideration of the test validity.

While Cumming (2007) claims that the actual practices of writing assessments occur through raters' individual (or sometimes collaborative) interpretations and judgments, the validity of composition tests is determined by describing and evaluating the cognitive processes of scoring written compositions (Connor-Linton, 1995). Lumley (2005) argues that the validity of writing assessments resides more in the complex cognitive processes of human scoring than in the necessarily underspecified, descriptive criteria and scales that conventionally define a writing

test.

The present paper has argued the necessity of considering the basic elements of designing writing tests. They are 1) traditional ideas (validity, reliability and practicality), and 2) the fact that assessing writing involves a constant interplay between making interpretations and judgments as well as attention to diverse features of the language, discourse organization, and ideational content of the text being assessed (Cumming et al.,2002, also cf. Shaw and Weir, 2007).

Note

Note 1: Cumming et al. (2005) show that there were significant differences between the discourse that examinees wrote for the independent essays and the integrated reading-writing or listening-writing tasks in respect to:

1) Lexical sophistication (in terms of word length and different words produced),
2) Syntactic complexity (in terms of words per T-unit and clauses per T-unit),
3) Argument structure (in terms of propostions, claims, data, warrants, and oppositions),
4) Voice in source evidence (in terms of specifying the self or other sources as evidence), and
5) Message in source evidence (in terms of proportions of declarations, paraphrases, and summaries).

Note 2 : Cambridge ESOL's traditional approach to validating tests—the VRIP approach (V=Validity, R=Reliability, I=Impact, P=Practicality), and the work of Bachman (1990) and the early work of Bachman and Palmer (1996) underpinned the adoption of the VRIP

Note 3: Shaw and Weir (2007) show that three critical components of any language test constitute an innovative conceptualization of construct validity. They are:

1) cognitive validity (the test-taker's cognitive abilities)
2) context validity (the context in which the task is performed)
3) scoring validity (the scoring process)

Bibliography

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

Buck, G. (1994). The appropriacy of psychometric measurement model for testing second language listening comprehension. *Language Testing* 11, 145–170.

Carson, J. G. (2001). Second language writing and second language acquisition. In T. Silva & P. K. Matsuada (Eds.), *On second language writing* (pp. 191–199). Mahwah, NJ: Laurence Erlbaum.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In > Bachman and A. Cohen (eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Chapelle, C. & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge University Press.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly* 29, 762–65.

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing*, 18, 2, 207–224.

Cumming, A. (2007). Book review: Lumley, T. 2005: Assessing second language writing: the rater's perspective .*Language Testing*, 24, 2, 287–291.

Cumming, A. & Berwick, R. (eds.) (1996). *Validation in language testing*. Clevedon: Multilingual Matters Ltd.

Cumming, A., Kantor, R. and Powers, D. (2002). Decision making while assessing ESL/EFL writing: A descriptive framework. *Modern Language Journal* 86, 67–96.

Cumming, A., R. Kantor, K. Baba, U. Erdosy, K. Eouanzoui, and M. James. (2005). Differencesin written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 1, 5–43.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, Summer 1994,* 31, 2, 93–112.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228–242.

Lumley, T. (2005). Assessing second language writing: the rater's perspective. In Grotjahn, R. and Sigot, G. (eds.) *Language testing and evaluation*. Vol. 3. Berlin: Peter Lang.

McNamara, T. & Roever, C. (2006). Language testing: the social dimension. In Young, R. (ed.) *Language learning monograph series.* Oxford: Blackwell.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* 18, 2, 5–11.

Nakamura, Y. (1996). A Study of Raters' Scoring Tendency of Speaking Ability through Verbal Report Methods and Questionnaire Analysis. *TKU Communication Studies* No. 5, 3–17.

Nakamura, Y. (2002). Effectiveness of paired rating in the assessment of English compositions. *JLTA Journal*, 5, 61–71.

Nakamura, Y. (2003). Two-dimensional performance assessment: speaking and writing tests. *The TKU Journal of Communication Studies*, 18, 7–15. Tokyo Keizai University.

Nakamura, Y. (2004). A Comparison of holistic and analytic scoring methods in the assessment of writing. *Proceedings of the 3rd Annual JALT Pan-SIG Conference,* 1–9, JALT.

Nakamura, Y. (2007) . Effective use of differential item functioning information for rater training and language teaching. *ICU Educational Studies.* 49, 10–19.

Shaw, S. D. & Weir, C. J.(2007). *Examining Writing: Research and practice in assessing second language writing,* Studies in Language Testing 26, Cambridge: Cambridge University Press.

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

Weir, C. J. (1993). *Understanding and developing language test*s. Hemel Hempstead: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.