Collecting Research Data on Students Writing Skills

by
Mohammad Adnan Latief
adnanlatiefs@yahoo.com

University of Pittsburgh
State University of Malang
2009
COLLECTING RESEARCH DATA ON STUDENTS WRITING SKILL
by Mohammad Adnan Latief

Abstract

Data on students writing skills can be obtained from professional writing teachers who keep good records of students writing progress through authentic writing assessment as well as through formal summative evaluation, or directly from students through the process of writing assessment. Students writing skills can be assessed indirectly by giving them an objective test to measure the students skill and knowledge of the components of writing. From the scores of the mastery of the writing components, then the students writing skill is predicted. Direct measurement of students writing skills requires the prompt and the scoring scale to be prepared. The prompt which may provide students with one topic or more than one alternative topics for students to write about, must be stated very clearly so that the students are sure of what they have to write. A scoring scale can be developed using holistic, analytic, or primary-trait scoring technique. In holistic scoring, the quality of each text is graded with one score based the general quality of the text, while in analytic scoring, the quality of each text is graded with a profile of several scores each representing the quality of different components. Primary-trait scoring is in between the holistic and analytic scoring techniques. A long writing scale of many points can differentiate quality levels of students texts but it is difficult to develop the scale and difficult to score the texts. A short scale of few points, on the other hand, is easy to develop and it is easy to score the texts, but it cannot separate different levels among good texts or among poor texts. Good data on students writing skills have to be supported with evidence of construct validity, content validity, empirical validity evidence, as well as good estimate of reliability

Key Words: writing assessment, validity, reliability, error of assessment, correctness, preciseness, holistic, analytic, primary-trait scoring.

Research procedures involve several activities from determining research problems, and objectives, defining the constructs and coverage of the variables involved, determining the design of research, selecting the sources of data, developing instruments to collect data, collecting data, verifying the validity and reliability of the collected data, analyzing the data, concluding the finding, discussing the finding, to drawing the conclusion or product of the research. Each of these activities has to be done correctly or otherwise the product or the conclusion will not be convincingly trustworthy. If the source of data is wrongly selected, for example, the obtained data will be wrong, and the
conclusion of the research will be wrong. Even if the source of data is trustworthy, if the instrument to collect data is not appropriate, the data collected will again be wrong, and in turns the conclusion of the research will be questionable. Similarly, research product based on correct source of data, with correct choice of data collection instrument, but with the wrong process of data collection will also be questionable. It is very important, therefore, that each of these activities be understood by any researcher before conducting a research project.

This article, which is taken from one chapter of the writer’s book, Research Methodology on Language Learning, discusses one activity in the process of research project on Language Learning. It is obligatory for any researcher on language learning to know well enough how to collect research data on language learning. This article presents some issues to be considered in collecting research data when students writing skill is involved as the variable of the research data. The discussion covers how to get research data on students writing skill from writing teachers and directly from students, how to assess students writing skill directly and indirectly, how develop the writing prompts, using one topic or alternative topics to write about, how to score the students writing texts, using long scale or short scale, applying holistic, analytic, or primary-trait scoring technique, how to validate scores on students writing texts, and how to provide estimate of reliability for scores on students writing texts.

**COLLECTING DATA ON STUDENTS WRITING SKILL**

Data on students writing skill can be collected from writing teachers’ documents on students writing skills or by conducting writing assessment to the students. Professional writing teachers conduct formative authentic writing assessment as well as summative formal writing evaluation. Therefore, writing teachers can provide research data on students writing skills.

**Getting Research Data from Writing Teachers Documents**

Formative authentic and alternative classroom writing assessment refers to any activity of the writing teacher that involves systematic collection of information about the writing skill the students are learning/acquiring, administered throughout the whole
process of learning, to monitor the progress of learning, for the purpose of facilitating the teachers in giving the students maximum help in developing their writing skills (Hoy & Greg, 1994:4). The alternative classroom assessment for the students’ progress in learning writing may take the form of (the teachers’) examining and recording the students written texts from their journals, notebooks, wall magazines, personal letters to their classmates or to the instructor. Condon (2009) said if students are asked to write their own experiences as writers or as learners, the resulting texts would be useful research data to reflect the students writing skill much better than if students are given a formal writing test.

Summative writing assessment refers to formal writing assessment administered at a certain time (in the middle and at the end of a quarter or a semester), to evaluate the achievement in learning writing or the proficiency in writing, for the purpose of judging the quality of students writing skills, or for the purpose of making decision for the students (giving scores, regrouping, passing/failing, keeping them at the same grade or promoting them to the next grade).

In Indonesian schools, formative assessment seems to be mixed up with summative assessment. Some teachers use midterm test scores and the scores from daily assessment and the scores from summative assessment to determine the students final score to represent the students learning achievement.

Assessing Students Writing Skill

Students writing skill can be assessed directly by asking the students to produce a written text and scoring the quality of the text or indirectly by giving them an objective test, like multiple choice type test, and computing the scores from their answers.

The indirect assessment is aimed at measuring certain components which are believed to be the prerequisite for the students to be able to write well (Stiggins, 1994:106). The people who use the indirect writing assessment believe that if the students have developed the pre-requisite skills/knowledge of writing, like the skill in choosing correct words, in applying the right use of grammar, in ordering the ideas in the sentences and the paragraphs, in using the right cohesive devices, they can be predicted to have a good writing skill. They argue that it is not possible for anyone to be able to produce a
good piece of writing until the knowledge and skills of the sub-components of writing be acquired. It is possible, therefore, to assess the students writing skill by giving them an objective type test, which assesses their knowledge of skill in the writing components without necessarily asking the students to actually write a piece of writing (Breland et al., 1987:1)

Direct writing assessment is done by asking the students to actually produce a piece of writing or by selecting a sample of each student’s writing and scoring the students writing using certain standards of writing judgment. The scoring is done by writing scorers who have been trained or skillful in making judgment about the quality of students writing to reflect students writing skill. People who prefer direct writing assessment argue that the students writing skill can only be assessed from the students actual piece of writing.

Each of the direct and indirect assessment methods has its own drawbacks. The direct assessment method has been criticized as suffering from low reliability resulting from many possible sources of errors, such as (a) each scorer’s personal errors, (b) errors introduced by the limited sampling of topics on which students can write, (c) errors the scorers may make because of the influence of, for example, essay length and handwriting neatness, and (d) errors in judgment due to the influence of the judgment given to the immediately preceding essays (Breland et al., 1987). The indirect writing assessment, on the other hand, has been criticized as suffering from low validity. It does not measure an actual sample of the students production of writing (Greenberg, 1986).

**Direct Method of Assessing Students Writing Skill**

The preparation before assessing students writing skill is developing the research instrument to collect the students texts, as the correct data collection instrument becomes the evidence of the validity of the data to be collected. With well developed research instrument, the researcher collects the research data. The research instrument can be selected from any available instrument with the criteria that the instrument is developed to measure the same purpose as the data to be collected for the research. Or if the available instrument is only partly appropriate with the research data to be collected, the
A researcher can revise the instrument. If there is no available instrument appropriate to the purpose of the research, the researcher has to develop the instrument from the scratch.

**Developing Data Collection Instrument**

In developing data collection instrument, a researcher has to start by defining the construct of what is meant with the students writing skill. The wrong definition of the construct of what the students writing skill is will lead to the wrong choice of the instrument and in turns will lead to the result of data that do not represent the students writing skill. The correct definition of the construct which will lead to the correct choice of the prompt, and will in turns lead to the result of data that reflects the students writing becomes the construct validity evidence of the data.

When the construct of the variable, the students writing skill, is correctly defined, then the researcher has to define the coverage domain of the task to be involved. A certain level of students writing skill might involve and be reflected by the correct choice of the words, the right use of grammar, the right order of the ideas in the sentences and the paragraphs, the right use of cohesive devices, the clarity of the main idea to be proposed, the evidence to support the main idea, etc. All of these aspects have be included in the development of the task of the instrument to provide evidence of content validity of the data to be collected, or otherwise, the data collected may suffer from content validity problem.

Based on the defined construct of the students writing skills and the content coverage in the tasks to be included in the instrument, the research instrument is selected or developed. Then, the research instrument has to be tried out to see if the data from the try out correctly reflect the students writing skill. This test can be done by comparing the data with the data resulted from another more trusted measure. The match between the try out data and the data from another more trusted measure can be used as the empirical validity evidence for the data to be collected, or otherwise, the data to be collected will suffer from empirical validity problem.

Two instruments have to be prepared to collect data on students writing skills. The first instrument is the prompt to ask the students to produce a written text. The second instrument is the scale to score the students texts.
Developing the Prompt to Ask students to Produce Written Texts

To get students written texts, a researcher has to develop a prompt that requires the students to actually write a text which can be scored to reflect the text quality, which in turns reflects their writing skills. A decision has to be made whether to ask all the students to write about the same topic or to provide alternative topics one of which has to be selected by the students to write about.

A good prompt is the one that can help students show their maximum writing competence in producing good written texts. White (1985) suggests some requirements for a good prompt. First, the task required must be clear to the students, so that the students can see clearly what they have to write. Some students with good writing competence may not be able to produce a high quality written text if they are not sure of what to write. Second, the topic must be equally challenging to all students to write about. A student with good competence in writing may not be able to produce a good written text optimally because the topic is not challenging enough (Ruth & Murphy, 1988). Third, the prompt must require all students to do the same task (White, 1985). If the students are allowed to show different skills, some writing narration, others description, exposition, or arguments, then the students written texts are not comparable to each other.

Asking all the students to write about the same topic or providing alternative topics one of which to be selected by the students to write about has some trade off. Asking all the students to write the same topic makes it easy for the scorers to apply the same standard for scoring all the student texts, but it is not easy to guarantee that the topic is equally challenging to all the students. On the other hand, providing alternative topics, one of which to be selected to write about allows each student to select equally challenging topic to write about, but if the topics vary so greatly that require students to show different writing quality based on different writing skills, the writing texts are not comparable in scoring. For any of the techniques selected, the researcher has to be aware of the weaknesses and has to minimize the weakness.
Developing Scoring Scales for Students Written Texts

One of the three techniques of scoring the quality of written texts; holistic, analytic, or primary trait scoring scales can be selected. Holistic scoring views a text as a whole, analytic scoring views a text as composed of several components making up the quality of the text, primary trait scoring views a text from its rhetorical quality. Holistic scoring is based on the belief that the quality of a text can only be seen when we see it as a whole, not as combination of its parts. In this belief, the whole is believed to be more than just the combination of each component. In other words, the combination of all good components do not necessarily make up a good whole. On the other hand, in analytic scoring, the belief is that the whole is made up of the combination of the components. When each component is good, then the whole is good (White, 1985)

In holistic scoring, each text is given one score to represent the quality of the text as a whole. This one score is given by a scorer based on the general impression of the quality of the text. Hieronymus et al. (1987) gives an example of a scoring scale for primary school students narrative text holistically (see Table 1). In analytic scoring, each written text gets several scores from one scorer, each score representing the quality of one components in the text. These scores are averaged after being weighted, a more important component being weighted more than the other component. Cooper & Odell, (1977) gives an example of an analytic scoring scale for students personal narrative writing texts (see Table 2). In primary-trait scoring, just like in holistic scoring, each text only gets one score representing the quality of the rhetoric of the text.

Each of the scoring techniques has its own strength and drawbacks. Holistic scoring gives general information on the quality of a text in a relatively short time. Therefore, this technique can be used to assess writing skills of a big number of students or for the purpose of selection (Carlson et al., 1985). This technique, though, is not appropriate to diagnose students writing weakness and strength. The problem with holistic scoring technique is that it is not easy to come up with a single score to represent the general picture of the quality of each text (Spandel & Stiggins,1990). This technique, therefore, can only be used by a highly skilled and trained scorer.
Table 1
Holistic Scoring Scale for Elementary Pupils Narrative Texts
By Hieronymus et al (1987)

<table>
<thead>
<tr>
<th>Narrative Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The narrative task- to tell a story- is probably the most familiar to the early elementary pupils. Even the writer of a “1” paper in this mode appears to have no trouble with the nature of the assignment. But the narrator is unable to tell a good story because of his or her mechanical limitations. The reader has only the vaguest idea of what is happening, to whom, and in what situation. Specifics of plots, character, and setting are obscured by indecipherable spelling, unconventional vocabulary, and/or incohesive sentence structure.</td>
</tr>
<tr>
<td>2. The writer of “2” narrative tells a clear but insufficiently elaborated story. Mechanical problems have not disappeared, but they do not intrude on the story. The reader understands what is happening. Unfortunately, very little is happening in a “2” story. The writer’s problem is generativity. He or she seems unable to get beyond the prompt to any additional detail of his or her own. In many ways, a “2” paper bears more resemblance to a description of the sequence of drawings which constitutes the prompt than to a well-developed story.</td>
</tr>
<tr>
<td>3. A “3” paper tells a good story. Mechanics are adequate. The reader begins to be able to differentiate the story of one writer from that of another. Some writers at the “3” level are beginning to use dialogues; others describe the emotions of their central characters. In both cases, characters are beginning to display individuality. The setting of the story generally is unelaborated in the “3” paper, but the plot and characterization are strong enough to compensate for this omission.</td>
</tr>
<tr>
<td>4. The dominant characteristic of a “4” paper is uniqueness. No “4” paper can’t be mistaken for another. The writer excels on some dimension, and as a result, his or her story is memorable. The writer’s conceptualization more than compensates for mechanical errors which may still appear in a paper at this level. Something-vivid language, realistic detail, a highly elaborated or unusual context for the story, a dramatic twist to the plot or extraordinary complexity of characterization makes the “4” narrative outstanding.</td>
</tr>
</tbody>
</table>
Table 2
Analytic Scoring Scale for Personal Narrative Writing
(Cooper & Odell, 1977)

Rater: _______________

<table>
<thead>
<tr>
<th>I</th>
<th>General Qualities</th>
<th>LOW</th>
<th>MIDDLE</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Author’s Role</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>B: Style</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>C: Central Figure</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>D: Background</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>E: Sequence</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>F: Theme</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II</th>
<th>Diction, Syntax, Mechanics</th>
<th>LOW</th>
<th>MIDDLE</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Wording</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>B: Syntax</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>C: Usage</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>D: Punctuation</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>E: Spelling</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

Analytic scoring can be used to represent the profile of the quality of each student writing. Therefore, it is very helpful for the purpose of diagnosing students writing weaknesses and strengths (Carlson et al, 1985). The profile is very important for a classroom teacher to make instructional decision in helping students develop their writing skills. The problem with this technique is that it is difficult to assign several scores representing different components of writing from a single reading of the text. Raters who are not well trained may find it difficult to assign an independent score for each component, free from the influence by the score of another component (Spandel & Stiggins, 1990). This scoring technique can, therefore, be used only by highly skillful and trained scorers.

Primary-trait scoring technique may be the easiest to score students texts, as it requires a scorer to come up with only one score to represent only one prominent component of a text, like the clarity of the main idea, the strength of the argument, the coherence, etc.
The Length of Scoring Scales for Students’ Writing

Another important issue in developing the scoring scale is the length or the number of points for the scoring scale. The shorter the scale (the fewer the number of points on the scale) is, the easier it is to develop the descriptor for each point, but the weaker it is in differentiating text qualities. A short scale of two points, point “1” and point 2 is easy to describe, as “1” representing poor quality, and “2” representing good quality, but different qualities among the good quality texts is not identified.

Similarly, the longer the scale (the more the number of points on the scale) is, the better it can differentiate levels of text quality, but the more difficult it is to develop, since many points have to be described. A long scale of 100 points, with “1” as the lowest quality and “100” as the highest quality is very good in separating among poor, medium, and high quality texts. But of course, it is very hard to provide 100 descriptors. It is even more difficult using a long scoring scale in analytic scoring where the weighting for each component is different. It is even more complicated in analytic scoring when one scale is used to score all components of text quality, for example, the choice of vocabulary is scaled from points 1 to 10, correctness of using grammar is scaled from 11 to 30, clarity in organization is scaled from 31-45, the strength of the argument is scaled from 46-70, and the clarity of the proposed idea is scaled from 71 to 100.

The moderate way is using a -4, -5, or -10 point scale. In a 4-point scale, the very bad text is scored 1, a bad text with some good points is scored 2, the very good text is scored 4, and the very good text with some weak points is scored 3.

To test the clarity of descriptors of a scoring scale, two or more raters can be asked to use the scale to score one student text. If the two raters give the same score for that text, then the two raters have the same understanding of the descriptors, which means that the descriptors are very clear. But if the raters do not agree on the score for the same text, then it must be either because the descriptors are not very clear, or the raters do not have enough understanding of the descriptors.
VALIDITY OF THE DATA ON STUDENTS WRITING SKILLS

Only with correct data on students writing skill, a conclusion of research on students writing skills can be convincingly valid. The correctness of research data in representing the research variable, students writing skill, has to be supported with evidence, otherwise, the research data and the conclusion drawn from the data suffers from validity problem. The stronger the evidence that supports the validity of the research data, the more valid the research data and the conclusion are.

Defining Validity of the Data on Students’ Writing Skills

The data on students writing skills are valid if they correctly reflect the writing skills of the students, correctly predict the way students would perform on another writing assessment conducted for the same purpose, and correctly predict similar writing performance in a different situation. If the data on students writing skills do not correctly reflect the students writing skills, do not correctly predict the way students would perform on another writing assessment conducted for the same purpose, or do not correctly predict similar writing performance in a different situation, then the data cannot be used as valid representation of the students writing skill. The level of validity of data on students writing skills, then, can defined as the correctness of the data in representing the students skill in writing. Validity, therefore, is the characteristic of the data and the conclusion drawn from the data, not the characteristic of the data collection instrument. Frisbie (1986:90) said

Validity traditionally has been regarded as a test characteristic, generally the most important quality of a test. But the current thinking of measurement experts is that validity should be associated with the use to be made of the scores from a test. In particular, validity refers to the appropriateness of making specific inferences or of making certain decisions on the basis of scores from a test. The question, then, is not “Is the test valid?” But “Is it valid to use the scores from this test to make these kinds of interpretations about these examinees?”

The evidence to support the validity of the data on students writing skills can be provided from the instruments that have been used to collect the data which represent
theoretical validity evidence and the evidence that shows what writing skills the students can perform which represent empirical validity evidence.

**Theoretical Validity Evidence**

If the instrument to collect data on students writing skills contains a writing task, a prompt which requires students to produce a piece of writing, and a scoring scale which is designed to measure the quality of the students texts, the data obtained will have construct validity evidence. If the prompt does not require students to produce a written text from which the writing skills of the students are drawn, then the data cannot be claimed as the valid representation of the students writing skills. If the prompt asks the students to perform an oral presentation, for example, and based on the oral presentation the students writing skills are assessed, then the data obtained will suffer from validity problem, and therefore, cannot be used as valid representation of students writing skills.

Ebel and Frisbie (1986: 91) said “The meaning of the test scores is derives from the nature of the tasks examinees are asked to perform.”

Another source of validity evidence from the instrument is the coverage of the tasks. If the tasks cover the samples of all contents of the domain of the variable to be assessed, the data obtained will have content validity evidence. Gronlund (1985:59) said

> The essence of content validation is determining the adequacy of sampling. More formally, content validation is the process of determining the extent to which a set of test tasks provides a relevant and representative sample of the domain of tasks under consideration.

Construct validity evidence and content validity evidence are both important for the data to establish the trustworthiness of the conclusion of the research.

**Empirical Validity Evidence**

Unlike the theoretical validity evidence that can be provided from the test instrument, empirical validity evidence can be obtained from another set of data which are collected to represent the same variable (Heaton, 1988:161). Ebel and Frisbie (1986:95) calls the evidence as criterion-related evidence for validity.

The empirical validity evidence is provided by correlating the scores on the students writing skills that have been collected with another set of scores (criterion-
related), which are also collected to measure the same students writing skills, like from the teachers’ records of the students classroom writing achievement. The criterion-related scores must, of course, be the scores that can be believed to be highly valid representation of the students writing skills. The high correlation between the two sets of scores represents strong empirical evidence of the validity of the data collected. Likewise, the low correlation between the two sets of scores shows that the data have empirical validity problem.

RELIABILITY OF THE DATA ON STUDENTS WRITING SKILLS

While validity refers to the degree of correctness of the scores of writing (to what extent the scores of writing actually assess the skills valued in effective writing, predict the way students would perform writing skills on another assessment conducted for the same purpose and predict similar writing skill performance on a different situation), reliability of the scores of writing refers to the preciseness of the writing scores in representing the actual level of the students writing skills. The writing scores have high reliability if the scores precisely represent (very close to, or not too far away from, or give good estimate of, or do not overestimate or underestimate) the true level of the students writing skill. In other words, if the writing scores are too far away different from the true level of the students writing skills, then the writing scores have low reliability. The distance between the true level of the students writing skill and the writing scores, then, determines the degree of reliability; the bigger the distance is between the writing scores and the actual level of the students writing, the lower the reliability of the writing scores. The distance between the writing scores and the actual level of the students writing represents errors of the writing scores. In other words, the bigger the errors in the writing scores are, the bigger the distance is between the students writing skills and the actual level of the students writing, and the lower the reliability of the writing scores is.

Mathematically, the relationship between the writing scores (X), the true level of the writing skill (T), and the errors (E) can be formulated as $X = T + E$ (Allen & Yen, 1979: 57, Ebel & Frisbie, 1985: 72) The formula explains that each of the students writing score (X) contains the mixture of the true level of the writing skill (T) and the
error (E). The amount of error (E) determines the degree of the reliability of the writing scores (X). The bigger the error (E) is, the lower the reliability of the writing score (X) is, and similarly, the smaller the error (E) is, the higher the reliability of the writing scores (X) is. See Allen & Yen as quoted below.

“As reliability of a test increases, the error score variance becomes relatively smaller. When error variance is relatively slight, an examinee’s observed score is very close to his or her true score. However, when error variance is relatively large, observed scores give poor estimates of true scores” (Allen & Yen, 1979:73).

Some language testing experts define reliability as referring to consistency of the scores resulted from the assessment (See Djiwandono, 1996: 98, Gronlund, 1985: 86 for examples). Consistency is an important indicator for reliability, meaning that if an assessment result is (or the test scores are) consistent from one assessment to another, then the assessment result has (or the test scores have) high reliability. However, consistency is not the meaning of reliability, it is only an indicator of reliability. The meaning of reliability (of a language skill assessment result) is preciseness (of the assessment result or the closeness of the X to T).

Factors Affecting the Degree of Reliability of the Scores of Students Texts

The main factor affecting the validity of students writing scores is the appropriateness of the procedure of the assessment (the appropriateness of the choice of instrument). An assessment of students writing skill using a paper-and-pencil test that requires the examinees to select one of the alternative expressions provided writing, for example, will result in the writing scores with low validity (having construct-validity problem), which means that the writing scores better represent the knowledge of writing rather than the skill of writing. If the assessment of writing skill is administered using an essay test, which requires the examinees to show the skill of writing by actually writing and based on the writing product the writing skill is estimated, then the result of the writing assessment (or the students writing scores) will have higher validity (with higher construct validity evidence).

“When we interpret test scores as a measure of a particular construct, we are implying that there is such a construct, that it differs from other
constructs, and that the test scores provide a measure of the construct that is little influenced by extraneous factors. Verifying such implications is the task of construct validation.” (Gronlund, 1985: 72).

However, if the writing test only requires the examinees to write loose short answers to the questions in the test, then the students writing scores will still have low validity (content validity problem) because the students are not required to performs all parts of writing skills, like organizing the ideas, using connecting devices, showing unity of the essays, etc.

“Content validation is the process of determining the extent to which a set of test tasks provides a relevant and representative sample of the domain of tasks under considerations.” (Gronlund, 1985:59).

While low validity of the students writing scores means that the scores wrongly represent another skill other than writing skill, low reliability of writing scores means that the scores resulted contain big errors and so give poor estimates for (overestimate or underestimate) the true level of the students writing skills. The poor estimates of the students writing scores may be caused by (1) the inability of the students to show their best writing performance, (2) the inability of the instrument to solicit the students best writing performance, (3) the inability of the raters to give objective judgment about the level of the students writing skills. Ebel & Frisbie (1985: 73) said:"Reliability depends on the nature of the group tested, the test content, and the conditions of testing."

Following are several conditions which cause the reliability of students writing scores to be low.

Not the Best Writing Performance of the Students

Errors in writing assessment that cause the scores to underestimate the true level of the students writing skill may happen because the students are not in their best condition when the writing assessment is being administered due to the physical as well as emotional constraints. They may be sick, tired, hungry, emotionally disturbed, not concentrating, sleepy while the writing assessment is conducted. If for some reason the students do not submit their own writing, instead they submit others’ writing which has better quality, on the other hand, the writing scores will overestimate the actual level of their writing skill. To avoid the errors of either underestimating or overestimating the true
level of the students writing skill, therefore, the writing assessment should be conducted in such situation that the constraints can be minimized. The data collector should select the best conducive atmosphere to make sure that the students being tested are not having those constraints while the writing assessment is being administered.

**Not the Scorers’ Most Objective Judgment.**

Like the errors coming from the examinees’ physical as well as emotional constraints, errors in writing assessment that cause the scores to underestimate the true level of the students writing skill being assessed may happen because the raters who give the judgment to the quality of the students writing are not in their most natural and objective physical as well as emotional mode. They may be sick, tired, hungry, emotionally disturbed, not concentrating, sleepy while giving judgment. Doping or any situation that makes the raters over-active or excessively happy, on the other hand, can also cause the judgment to result in writing scores which overestimate the actual level of the writing skill being assessed. To avoid the errors of either underestimating or overestimating the true level of the writing skill being assessed, therefore, the judgment process should be conducted in such situation that the constraints can be minimized. The raters should select the best conducive atmosphere to make sure that they are not having those constraints while giving the judgment to the quality of students writing

**The writing Assessment of too short paragraphs.**

A writing assessment which requires the students to write an essay of several paragraphs will result in the writing scores with higher reliability than the same writing assessment which asks the students to write only a paragraph. In other words, the writing raters could see the level of the students writing skill better from the students essay of several paragraphs than just from the students short paragraph, or the scoring from the students essay of several paragraphs will result in scores with higher reliability than the scoring from the students short paragraph. (Ebel & Frisbie, 1985: 84).

**The Writing Assessment Tasks being too Difficult or too Easy.**

A writing assessment which gives students a too difficult topic to write, or a topic which the students do not know much about will result in writing scores which do not
accurately reflect the actual level of the students writing skill as the students will have to work too hard on thinking about the topic rather than on expressing their ideas. In the same way, if the topic is too popular, the students might have read or heard a lot about that topic so that they do not have to think of how to express the ideas about the topic, they will only need to recall and then write what they have read and heard a lot. (Ebel & Frisbie, 1985: 85). So, the level of difficulty of the task in the writing assessment influences the degree of reliability of the students writing scores.

**Uncomfortable Place and Time for Writing Assessment**

A writing assessment conducted in an uncomfortable room; too hot, too cold, too windy, too crowded, too small, or too noisy, and in uncomfortable time; at 2.00 p.m. after the examinees have worked all morning, for example, will result in the writing assessment scores with low reliability. This uncomfortable place and time will not make the students able to produce their best writing, so the writing scores based on the students writing produced in these uncomfortable situations will have low reliability.

**ESTIMATING THE DEGREE OF RELIABILITY OF DATA ON STUDENTS’ WRITING SKILLS**

Estimating the degree of reliability of students writing scores refers to verifying or confirming whether the writing scores have high degree of preciseness in reflecting the actual level of these students writing skill. The evidence assumed to indicate high degree of preciseness of writing scores is the consistency of the scores. It is assumed that if the writing scores precisely represent the actual level of the students writing skill being assessed or if the scores do not contain too big mistakes (if the distance between the scores and the actual level of the skill being assessed is not too big), then the writing will be consistent. Conversely, if the writing scores do not precisely represent the actual level of the students writing skill being assessed or if the students writing scores contain too big mistakes (if the distance between the students writing scores and the actual level of the students writing skill being assessed is too big), then the students writing scores will not be consistent.

Estimating the degree of reliability, therefore, refers to an effort to collect evidence of consistency to verify or to confirm the reliability. In other words, if the scores or the
results of assessment are provided with evidence of high consistency, then the writing scores convincingly have high degree of reliability/ preciseness. Different sets of scores can be obtained from retesting (the students are asked to write again) or inter rater scoring (different raters are asked to score the students texts using the same scoring scale).

The evidence of students writing scores consistency may be obtained by computing the co-relational index of the two sets of students writing scores. (See Djiwandono, 1996:99 and Ebel & Frisbie, 1986: 75 for more details).

**CONCLUDING REMARKS**

Each step in a research project must be done correctly, otherwise the research conclusion is not convincingly trustworthy. Data collection, one important step in research, on students writing skills has to be planned and conducted in the best possible way to get the best judgment that correctly represents the writing skills being assessed or does not wrongly represent another skill (have high construct and content validity) and precisely represents the level of the students writing skills, or doesn’t too far overestimate or underestimate the true students writing skill (have high reliability).

Low validity of data on students writing skills indicates that the data do not correctly represent the students writing skills and this may be caused by wrong tasks the students are required to perform and wrong coverage of the tasks components of writing. The low validity can also be measured by correlating the data with another data obtained from criterion-related assessment.

Low reliability of students writing assessment indicates that the scores too far overestimate or underestimate the students writing skill and this happens because there have been big errors in the process of writing assessment. The errors may happen because the students cannot show their best performance during the assessment process, the scorers cannot give their most objective judgment, the assessment process is not conducive enough for the students in doing the tasks, the writing assessment tasks are too difficult or too easy. The evidence of high consistency of students writing skills can be collected by computing the co-relational index from two sets of students writing scores obtained from the method of retesting with the same or parallel writing instruments.
REFERENCES


