Feasibility Criteria of Language Testing
Ye Jun\textsuperscript{1,a} and Huang Tao\textsuperscript{2,b}

\textsuperscript{1}School of Foreign Languages, Wuhan University of Technology, Wuhan, P.R. China, 430070
\textsuperscript{2}Office of Laboratory and Equipment Management, Huazhong Normal University, Wuhan, P.R. China, 430070
\textsuperscript{a}email: yejun725@163.com, \textsuperscript{b}email: huangtao@mail.ccnu.edu.cn

Keywords: feasibility; criteria; language testing

Abstract: There has not been a system of evaluation criteria specifically for feasibility of language testing so far. This paper aims at building a system of feasibility criteria which can be implemented in various language tests. The study is based on the current domestic study and abroad of language testing. It promotes testing expenses, testing administration, testing scoring and scores reporting as the first grade indexes of feasibility criteria and each one has further grade indexes. The promotion of such criteria is necessary and crucial to evaluate and serve for Chinese language testing.

1. Introduction
Language tests are instruments used to measure language ability or aptitude. The quality of language tests strongly effects the realization of testing purposes. This paper aims at building a system of feasibility criteria which can be implemented in various language tests.

1.1 Significance of the study
There has not been a system of evaluation criteria specifically for feasibility of language testing so far. Therefore, this study has both theoretical and practical significance.

1.1.1 Theoretical significance.
First of all, the exploration of such criteria can enrich the language testing theory. Experts in China and foreign countries have proposed some evaluation criteria for language testing, however, those criteria are difficult to be implemented in practice. Secondly, the exploration of such criteria can enrich the theory of psychological measurement. Language testing has been an independent subject since 1950s. In the development it has absorbed and learnt from the theory of psychological measurement.

1.1.2 Practical significance.
Firstly, the feasibility criteria can systematically evaluate the practice of language testing. So far, a number of discussions in China and other countries are focused on the quality of language tests themselves. However, high quality language tests require high quality administration. Therefore, it is of necessity to discuss such criteria in practice. Secondly, the feasibility criteria can better instruct the tests practice, which is the main purpose of such criteria. Language testing is a long process, including tests design, administration, scoring and use of the scores. Thirdly, the feasibility criteria can promote language teaching in China. It is acknowledged that backwash of language testing does exist. That is to say, successful language can in turn promote language teaching and learning. This is another purpose of the study.

1.2 Overseas and Domestic Research Status
1.2.1 Overseas Research Status
Bachman and Palmer (1996) considered practicality as an important factor of language testing. They define practicality as the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities. This relationship can be represented as the following:
Practicality = \frac{Available \ resources}{Required \ resources}

In their study of test usefulness, they pictured a model based on design, operationalism and test administration. They used one chapter to elaborate the definition, allocation and management of various resources. Nick Saville held that the principles of quality management and their relation with validity can be widely implemented not only in various design of tests, but also every step of tests, from the production of testing material and the explanation of the scores. The differences between school tests and large-scale tests lie in the complex levels of organization of resources and management structure, including human, material, time-related factors, intended use of the test results and importance attached to decisions made about the test takers.

1.2.2 Domestic Research Status

The theoretical and practical study of language testing has been developing rapidly since 1980s. In the meantime, a number of influential large-scale language testing in China, such as CET, PETS, TEM, have come into existence because of the growing popularity of English learning. Those high-stake tests exert a great influence to both examinees and the English language teaching. Therefore, a good command of the quality in each step of the testing, including the design, operationalization and using the testing results, is extremely important. However, there is not a systematic criteria which is suited to national conditions.

Wang Zhenya (2009) considered the practicality as the third consideration following validity and reliability. He held that practicality consists of human resource, material resource and time resource. Specifically, human resources include “adequate organizers, designers, administrators, and raters”; material resources include “adequate material and appropriate machine for the tests, such as paper, recorder, computer and test-scoring machine”; time resource consist of “designing time, testing time and scoring time”.

Li Xiaoju (2001) didn’t mention the definition of practicality directly, but she elaborated the high-quality modern testing process. She stated that three groups of professional people would be needed in each step of tests. Bu Aihua (2011) held that the evaluation criteria of tests consist of the following aspects: reliability, validity, difficulty index, discrimination, practicality and backwash. Among them, practicality means “whether the tests can be administrated easily”.

Domestic large-scale tests are also high-stake tests in which numerous testees are involved. In the meantime, those tests have backwash to language teaching. Therefore, it is urgent to build such criteria for the design, evaluation and use of those tests in China. The feasibility criteria of language testing should not only meet the requirements of general characters in language testing, but also take the features in China into consideration; it should not only consider the traditional hot issues in this field, but also include the latest development in language testing these years.

2. Feasibility criteria of language testing

2.1 What is feasibility?

Feasibility is an important norm to measure a test. It refers to whether the current conditions can meet the requirements of design, implementation and scoring in the aspects of human resources, capital and material resources (Wang Zhenya, 2009). Firstly, it requires competent and professional workers to design, administrate, score and manage the whole process of tests. Secondly, material resources will be necessary condition to guarantee the tests, such as the testing paper, testing machine, recorder, computer, scoring machine, etc. Thirdly, financial support will be the guarantee to the items above. Besides, the difference of testing scales makes different requirements to the resources.

2.2 What affect feasibility?

To set a test is like to paint a blueprint, from the first step to the last. A test starts from the decision of the testing range and categories to the use of the scores, from the design of the whole structure to
every single item, from the testing paper printing to the confidentiality, from the scoring to the statistics of the scores. Generally, feasibility involves in three aspects: human resources, capital and material resources.

Human resources consist of three groups of professional people:
1. Administrative personnel (People who are professional in administrating test. If not, trainings are necessary.)
2. Statistical analysts (People who are professional in educational statistics or computer operation)
3. Professional personnel (People from language teachers)

Material resources include paper printing, testing machine, recorder, computer, scoring machine, etc.

Financial resources are all the costs in the test, including the cost to hire relevant personnel, to afford relevant machines and the confidential expenses. Lack of fund may affect the workers activity or deter the testing process from moving on smoothly.

2.3 The index

2.3.1 The first grade index
Feasibility is an important norm to measure a test. Language testing mentioned in this assay refers to both Chinese language testing and other language testing. It refers to the large-scale language testing but not small classroom testing.

2.3.2 Categories of the first grade index
According to the testing steps in the literature above, the first grade index “feasibility” consists of the following indexes as the second grade.
1) Testing expenses
2) Testing administration
3) Testing scoring
4) Scores reporting

Testing expenses examines the financial support from the test design to the implementation. Testing administration examines the development of test implementation. Testing scoring evaluate the scoring process. Scores reporting considers whether the test results can be used properly.

2.3.3 Categories of the second grade index
Testing expenses refer to all the costs needed in the test, which is important financial support. Costs in tests include the fund on relevant people and material.

Human resources consist of three groups of professional people:
1. Administrative personnel (People who are professional in administrating test. If not, training is necessary.)
2. Statistical analysts (People who are professional in educational statistics or computer operation)
3. Professional personnel (People from language teachers)

Testing administration examines the development of test implementation. A standard test consists of the following steps: 1) to decide the testing purpose 2) to decide the testing range; 3) to choose the testing method; 4) to design the sample test; 5) to pretest and adapt the sample test; 6) to decide the final test; 7) to administrate the test and analyze the testing result; 8) to evaluate the testing quality (Zou Shen, 2008).

Bachman and Palmer (1996) summarize the above steps into three stages. The first one is the design stage, including defining the testing target, range and describing the testees’ features and requirements. The second is operationalization stage, which aims at designing the frame of the test and choosing the testing methods and designing testing items based on the target and range. The third one is the administration stage, including administrating, analyzing testing result and evaluating the testing quality. The above three stages are interdependent.

A real-time test is limited by the feasibility. Specifically speaking, a real-time test should take
the following three aspects into consideration: administration segments, administration stages and
administration conditions.

Scoring is a critical segment for the testing reliability. Failing in this part can make the efforts in
the test in vain. There are machine scoring and human scoring according to different subjects. There
are subjective scoring and objective scoring according to different items. There are on-the-spot
scoring and afterwards scoring according to different scoring time. As the development of
technology, computer scoring has been widely used.

Reporting and explaining the scores is to express the testing result in the way that is convenient
for the designers and users of the tests(Zou Shen, 2008). So far the measurement theories of
psychology and education fall ahead of those in physics, therefore, it is of great importance in the
result reporting. In this sense, reporting of testing results serves for testing validity.

2.3.4 Categories of the third grade index
According to the analysis above, the second grade indexes consists of the following indexes as the
third grade.

1) Testing expenses
   (1) Expenses on human;    (2) Paper printing;
   (3) Testing facilities;      (4) Confidential measures
2) Testing administration
   (1 ) Administration steps;   (2) Administration conditions;
3) Scoring
   (1 ) Scoring subject;        (2) Scoring object:  (3) Scoring time;
4) Reporting and explaining the scores
   (1) Original scores;         (2) Standard scores

2.3.5 Categories of the fourth grade index
Expenses on human refer to all the costs for hiring relevant people, including the administrative
personnel, statistical analysts and professional personnel.

Testing facilities are necessary for testing administration. The essential part is the construction
of the modern examination rooms which can fully make use of the Internet. In order to guarantee
the fairness and validity, examination rooms should be strictly monitored to avoid cheating. It is
also essential to cut the information connection from inside to the outside of the classrooms.

Take confidential measures for testing include asking the examiners to sign the confidential
agreement and the security of the testing contents. The more important the test is to the examinees,
the more important the test security is.

Administration steps refer to the order to finish a whole testing. Language testing usually
consists of listening, speaking, reading, writing and translating. Which part comes first is worth
considering. Besides, when to hand in the paper sheet, when to play the record and the testing time
range are important elements to consider. Once there is an emergency, contingency plan is needed.

Administration condition refers to all the hardware and software needed in the testing. Hardware
refers to the testing paper and relevant machines. Is there any printing mistake? Is the paper printed
clearly? Are the machines easy to use? Software is main about human and the rules and regulations.
Only all the factors work well together can the examination goes on smoothly.

Scoring methods consist of human scoring and machine scoring according to different objects.
Human scoring consists of paper scoring and non-paper scoring. There are also subjective scoring
and objective scoring according to different subjects.

Original scores are the paper scores. The standard score is the signed number of standard
deviations by which the value of an observation or data point is above the mean value of what is
being observed or measured. Each has its advantages in expressing the testing results.
3. Conclusions

Large-scale language testing in China often involves millions of examinees and exert critical backwash to language teaching. Therefore, it is necessary and crucial to build a series of criteria to evaluate and serve for Chinese language testing. There are a number of research results in this field both in practice and theory for us to learn in foreign countries. In general, the feasibility criteria do not only consider the common points of language testing but also Chinese individual features; the criteria do not only consider the hot topics in language testing, but also show the latest development.

Acknowledgement

This paper is supported by the teaching reform project of Wuhan University of Technology. Project’s name is *Positive washback of achievement test on College Students' English Autonomous Learning*. Project’s number is 2013102.

References


