The Multi-analysis of the Reliability of the Language Testing

Li ZHAO

Xi'an International University, Xi'an Shaanxi, China

601933109@qq.com

Keywords: Testing Reliability; Test Design; Test Takers; Test Administration

Abstract. Reliability is one of the most important qualities in language testing. For a language teacher, it is necessary to realize that there are so many factors affecting the reliability of the language test, from the test design to the test administration, from the test takers to the scorers, etc. The language teachers should not regard the testing as the only valid means to evaluate the students' language abilities, but make rational use of the language tests to serve their teaching.

Introduction

Reliability refers to the consistency and stability of the test scores. It is not only one of the most important ways to evaluate different levels of the language tests, but also an absolutely essential quality of tests, for unless the test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure. So only if the test result is reliable, the test can be valid and successful. Only this kind of test can be used to cultivate and select the persons with talent.

Specialists have long recognized that the examination of reliability depends upon our ability to distinguish the effects (on test scores) of the abilities we cant to measure from the effects of other factors. That is, if we wish to estimate how reliable our test scores are, we must begin with a set of definitions of the abilities we want to measure, and of the other factors that we expect to affect test scores. Now we will discuss these factors in the testing procedure, such as the test design, the test administration, and test scoring.

Validity

The reliability of the test is closely related to the test validity, which is thought to be the main qualities characterized by all the good tests, combined with the reliability. A test can be thought to be reliable only when it is valid, vice versa. Validity is the first factors needed to be considered. Generally, we think that if a test measures accurately what it is intended to measure, we can say it is a valid test. It is important to reach high validity to achieve high reliability. Then how to achieve high validity?

Content validity. In order to achieve it, what we need to do is to make a careful analysis of the skill or an outline of the course at the early stage of test construction, then write items based on the analysis and make sure that important areas are adequately represented. Besides, we also need to decide the proportion of the representative sampling of the content to be covered, such as the importance of the different areas, the classroom hours devoted to them and the difficulty level of each of them.

Construct validity. The construct validity is used to refer to the extent to which we can interpret a given test score as an indicator of the abilities or the construct, we want to measure. Construct validity also had to do with the domain of generalization to which our score interpretations generalize. In test writing, we have to take the TLU (target language use) task into consideration to achieve construct validity.

  DOI: 10.25236/icssae.2019.075

Backwash validity. Hughes thinks that the effect of testing and learning is known as backwash. We have to consider achieving beneficial backwash effects in test writing. What we can do may be: to sample widely and unpredictably; to use direct testing; to make testing criterion-referenced; to ensure that the test is known and understood by students and teachers and to provide assistance to teachers where necessary.

However, there is always some tension between reliability and validity. We have to balance gains in one against losses in the other. Even in this condition, it is necessary to take all these three factors into accounts in testing writing. Only this way, we can achieve the high validity. That is to say, we can identity that this test measures the abilities we want to measure, and to some extent, we can conclude it is a reliable test to some extent, though we have a long way to go to achieve high reliability in language testing.

Features of the test itself

The second factor we should pay attention to is certainly the features of the test itself.

Category of the test. The category of the test is the main factor to decide whether the test is reliable. If the form of the test itself is reliable, it is easier to achieve high reliability. For example, if the test is in the form of multiple choices, the result is almost perfect, while in form of writing, high reliability is hard to achieve. But the reality is that we cannot simply write the whole test paper in the form of multiple-choice in order to achieve high reliability, for this kind of test is invalid: there is not any situation to ask the test takers to choose the correct answer from the choice marked A, B, C, D in the real communicative setting. As a result, though the test in the form of it can be achieved perfect reliability in scoring, it is not reliable from the angle of the test itself and the aim of the testing. We have to employ such kind of tests in the subjective form to test the real communicative abilities of the test takers.

The test needs to allow the test takers a great of freedom in the way that they answer the ones that have chosen. But there is such a situation: the more freedom that is given, the greater difference the candidates' answers, thus the high reliability is harder to achieve. So the proportion is what we should to control in an ideal balance, which is why so many tests employ the combined test form of objective and subjective testing.

Adequacy of the sampling. The adequacy of the sampling is also an important factor. The more items that you have on a test, the more reliable the test will be. This seems intuitively right. If we want to know how good an archer someone is, we will not rely on the evidence of a single shot at the target. The one shot could be quite unrepresentative of their reliability. To be satisfied that we have a really reliable measure of the ability we could want to see a large number of shots at the target. This is the same case with the testing.

But we can't run to another extreme to give too many items in one test. If the testing paper is too long, the test takers may be tired of it and then guess the answers because of the limited time. Thus the reliability of this test is reduced. We can decide the quantity of the test according to such aspects, such as the age and the language level of the test takers, the types of the testing, and so on.

Test instructions. The third feature the test has in affecting the reliability is the test instructions. Because in testing, the test designers cannot be in the testing room and the invigilators are not allowed to explain the test in any way, the test takers should take a lot of time to read the instructions before finishing the testing. If the instructions are ambiguous and unclear, the examinees will be faced with the additional tasks which are not ones meant to measure their language abilities. Their unwanted time-wasting on the instructions will lower the reliability of a test. So the test instructions must be clear enough, and the test should be legible. It is better that the test takers can be familiar with the test instructions in advance.

Test discrimination level. The discrimination level refers to how well a test can tell the difference between high and low candidates. A test with high discrimination level can tell the different levels of the test takers' language abilities; while a test with low discrimination level cannot tell the different level so it is not a reliable test. It is closely related with the difficulty level of items included in the test

paper. Generally, the test paper should be carefully designed and organized. The levels of the tests should be reasonable and stay in an appropriate proportion.

Layout of the test. In test design, except for the sampling and discrimination, the layout of the test should be taken into account. You can imagine if a test is organized in double-side printed, at the same time, the pages are not significantly marked, the test takers may be too careless to miss some pages or items because of the psychology factor. As a result, the test is unreliable to evaluate the language ability of the test takers. The best way to avoid it is to make clear marking under the page.

In all, the test design is so important that nowadays a number of formal testing employ quite a lot testing specialists to construct the test in full time. If a test itself is relatively reliable, high reliability in the whole testing procedure is easy to achieve.

Test administration

The test administration is such a factor that some people neglect it. But there are so many random factors that will affect the high reliability in language testing to achieve.

Testing conditions. The uniform and non-distracting conditions of administration should be provided in order to achieve high reliability. In order to make sure the test administration is uniform, detailed instructions should be prepared for the invigilators and the subject of a meeting with them is also necessary. This point is especially important in oral testing when the whole testing process should be recorded. Once there is something wrong with, the test takers may suffer a lot, so the careful preparations show great importance.

Besides, the test room must be quite so as not to influence the examinees: the room should be large enough and the room between the students also needs to be large enough to avoid cheating of passing some information between the friends. The layout of the testing room, including the placing of the desks, should be arranged well in advance. The testing room should be clean; there is not any paper or other material which may be related to the testing on the ground or in the seat.

It is essential that the invigilators should time the test precisely, making sure that everyone starts on time and does not continue after the time is up. The examinees should be allowed to leave the room only one at a time and accompanied by an invigilator in order to avoid cheating by the excuse of leaving seats or room.

Invigilators. The invigilators should be responsible for the whole test procedure. Every test taker must be treated equally. The invigilators must give clear instructions about what the examinees should do, and instruct the examinees to provide the required personal information, such as the examination number, on the answer sheet or test booklet. It is quite important because we have found that many students haven’t given such information in the testing that they are scored zero in their testing results, especially in CET and PRETCO.

If some kind of equipment is needed in the testing, the invigilators have the duty to be familiar with the operation of the facility in advance. For example, in a listening test if the tape recorders are needed, all the invigilators in different testing rooms must be required to play the recorders at the same time and end it at the same time. Only this, it is fair to all the examinees.

In the process of the testing, the invigilators shouldn’t make any big noise or give too many warnings about the time, which will have negative effects on some examinees. Sometimes, to certain examinees, the attitude and the nature of the invigilators also affect the examinees; the invigilators can try best to avoid it.

Testing equipment. Some kinds of testing equipment are needed in many kinds of language tests, such as the tape recorder or listening facility in the listening test. The test administrator should check such equipment in advance to make sure they are in good working condition. If there is something wrong with the equipments, repair or replace them at once.

In addition, the testing room must have satisfactory acoustic qualities in order not to influence the examinees’ listening effects. These conditions are also needed to be checked in advance by the inspectors or the test administrators.

Attributes of the test takers. Attributes of individuals are the factors unrelated to the language ability include characteristics such as cognitive style and knowledge of particular content areas, and group characteristics such as sex, face, and ethnic background. For example, if an individual's field is likely to affect his performance on one test. And knowledge of economics is likely to affect an individual's performance on any test in which affect an individual's performance on any test in which economics is included as propositional content. This kind of knowledge may affect the performance of the test takers while the test is supposed to evaluate the language ability of the test takers.

Different factors will affect different individuals differently. Just as individuals vary in their level of language ability, so they will vary in the extent to which they are affected by different methods of testing. Some individuals, for example, may do very well on a multiple-choice test but perform poorly on a composition test. Or it may be the case that members of one racial or ethnic group may do better or worse on a given test than do members of other groups. Or some test takers may be well rested and mentally alert on the day of the test while others may have stayed up too late preparing for the test and consequently not be able to perform at their highest level of ability.

Other random factors. There are some kinds of factors, which may also affect the examinees by unsystematic, or random factors. These include unpredictable and largely temporary conditions, such as his mental alertness or emotional state. In the testing room, the location, the light, the temperature, even the humidity, all of these will have a positive or negative effect on the examinees. The emotion of the examinees may be influenced much more while in testing, for they need to concentrate on the testing more devotedly.

Even some test takers complain that their invigilators walk to and fro, this action made them be more nervous and can not concentrate on the testing.

The examinees may affect each other. For instance, if one of the examinees coughs or makes other noise in the listening test, the others may be influenced a lot, their listening ability can't be shown on the test results. In addition, if the one shakes his body to make himself calm, at the same time, the one in the back seat may be distracted by it; if the one on your left or right does pencil spinning all the time, then you may be annoyed while he just wants to find some ways to widen his thinking.

Such cases may vary in different situations and have different results for different examinees. What we can do is just to try best to give the examinees better testing conditions to get a higher reliability.

The way it is scored. We can suppose that a test with high reliability in test validity, test design and test administration is not reliable at all if the way it is scored involves too many personal or objective factors. So the way the test is scored is the key factor to affect the reliability. As to the scoring reliability, it may appear to be a recommendation to use multiple-choice items which involves completely objective scoring. However, this format sometimes cannot evaluate the test takers' communicative abilities which is the main aim of the most language tests, so we have to adopt the subjective tests to achieve the validity. The fact that we should face is that there are so many parts involving the objective scoring, such as the translation part and writing part, then how to improve the scoring reliability?

Provide a detailed scoring key. "For high scorer reliability the key should be as detailed as possible in its assignment of points. It should be the outcome of efforts to anticipate all possible responses and have been subjected to group criticism. Of course, this advice applies only where responses can be classed as partially or totally 'correct', not in the case of compositions, for instance." (Arthur Hughes: 41) All the scores should be familiar with the scoring principles.

Train scorer. This is especially important where scoring is most subjective. The administrator of scoring should analyze the patterns of scoring and the scoring key. Sometimes, certain amount of tests can be scored in advance where necessary. Even we can choose some authority to do it in advance then decide the scoring principles. Every scorer should be trained in scoring; anyone who hasn't trained to score accurately should not be assigned to score.

Employ multiple, independent scoring. In order to achieve higher reliability, we can employ multiple and independent scoring. The test papers can be scored separately by two independent scorers. Neither scorer should know how the other had scored the same test paper. They the test paper

can be passed to the third one, who compares the two sets of scores and investigates discrepancies. If necessary, we can employ more independent scorers to do this to get high reliability.

Identity the test takers by number, not by name. Studies show that even where the test takers are unknown to the scorers, the name on the test paper will make a significant difference to the scores given. For example, a scorer may be influences by the gender or nationality of a name into making predictions which can affect the scores given. The identification of test takers only by number will reduce such effects greatly.

Monitoring the scoring. Sometimes, the scorers' objective factors cannot be avoided, so the monitoring to the scorers can be important. The scoring group should form a monitoring group to recheck the scores and analyze the trends. Besides, for a test, not only the test takers but also the scorers should adjust psychology to be in a good mood. In this mood, the scorers can play their roles in scoring and get high reliability in scoring. That is the psychology factor.

Objective factors. The objective factors cannot be neglected. If the scorers are short of relaxing and have a heavy work, they will be tired in health and psychology, then they may be careless and lack of concentration. These elements will affect the reliability of the scoring of these kinds of scorers. If the scorers are not in a good health or mood, maybe it is necessary to suggest them to stop the scoring and have a rest.

Conclusion

I would reiterate that our primary interest in using language tests is to make inferences about one or more components of an individual's communicative language ability. A major concern in the design and development of language tests, therefore, is to minimize the effects of the test method, personal attributes that are not part of language ability, and random factors on the test performance.

Meanwhile, the language tests are just an indirectly way to evaluate the test takers' language abilities, not the direct one and not an absolutely valid method. However, the testing reliability is affected by so many factors from the test design to the test administration, from the test takers to the scorers, every process should be treated with caution. It is difficult to handle and difficult to realize the real high reliability. So the teachers should treat kinds of tests in rational attitudes, make full use of it to serve the teaching and learning.

References

[1] Lyle F. Bachman, *Fundamental Considerations in Language Testing* (Oxford University Press, Oxford, England, 1997), p.26-35.

[2] Lyle F. Bachman and Adrian S. Palmer, *Language Testing in Practice* (Oxford University Press and Beijing, Oxford, England, 1997), p. 28-57.

[3] Arthur Hughes, *Testing for Language Teachers* (Cambridge University Press, Cambridge, 1989), p. 26-47.

[4] Xiaoju Li, *The Science and Art of Language Testing* (Hunan Education Publishing House, Changsha, China, 1997), p. 24-47.

[5] Runqing Liu, *Language Testing and its Methods* (Foreign Language Teaching and Research Press, China, 2001), p. 206-223.

[6] Meihong Zhu, Hubei Social Sciences, (2003) No.3, p. 81-82.