Assessment Literacy: From Theoretical Constructs to Test Design and Delivery

Md. Elias Uddin*

Abstract

Assessment literacy is a fundamental requirement for all the personnel involved in activities related to language assessment. The information gathered from assessment is, in the main, used in making impactful decisions about language learners, language programmes, curricula as well as educational institutions. Failure of assessment instruments to provide reliable and accurate information results in misdirected and inappropriate decisions to the detriment of all stakeholders of educational assessment. Hence, all assessment staff must have the knowledge of the theoretical and philosophical underpinnings as well as practical considerations in the field of assessment. Considering the importance of developing assessment literacy among language teachers, this paper intends to shed light on the key theoretical constructs of assessment and testing, stages of test design and successful delivery of tests. Additionally, it aims to provide the beginning language educators with the very basics of assessment in a language neither highly technical nor too specialized for them to understand. The contents of this paper will also cater for novice educators across disciplines.

Keywords: Assessment literacy; principles of testing; test design; test delivery

1. Introduction

Assessment is inextricably linked to language teaching and learning. Language teachers regularly assess learners in the classroom or elsewhere to identify learners' needs, record their progress, determine how they are performing as teachers and planners, and evaluate the effectiveness of their programmes (Frank, 2012). Alongside assessment in the language classroom or examination halls of educational institutions, there are international standardized language tests administered by international standardized test-providing bodies. Besides, university admission tests and recruitment examinations in different contexts have sections dedicated to assessing the language abilities of test-takers. The purpose behind all assessment activities is to glean sufficient and accurate information to be used in making decisions about language learners, language curricula, language testing institutions, and educational as well as recruitment policies. If the information gathered is inaccurate and unreliable, the decisions made based on it are sure to impact adversely on all stakeholders of testing including candidates, guardians, teachers, institutions, curricula, and the policy-making bodies like the ministry of education.

Hence in order for assessment to be able to provide reliable information, all involved in the assessment process must be assessment-literate to the required extent. Testers—

* Lecturer, Department of English, University of Dhaka

classroom teachers or others—must have the adequate knowledge about the key concepts and guiding principles of testing, development and administration of tests, their uses in decision-making, and the overall impact of assessment and testing on the pedagogy as well as wider society. Therefore, given the importance of developing assessment literacy in classroom teachers as well as others involved in assessment activities, this paper seeks to present an overview of assessment literacy by focusing attention on the key constructs of assessment, stages of test design, and successful delivery of tests. It also aims to equip the beginning language educators with the rudimentary knowledge of assessment in words they can easily understand.

Assessment Literacy

The concept of assessment literacy was introduced by Stiggins (1991). In his view, assessment- literate teachers or test-writers must have an adequate understanding of the key principles of sound assessment practices. They must know how to design, administer and score tests. They must also be knowledgeable about how to 'interpret data generated from a test to make useful modifications to teaching and to use assessment as a tool to improve students' learning' (Rogier, 2014). Assessment-literate stakeholders are also aware of the rights and needs of candidates as well as the psychological factors that affect test performance.

In other words, assessment literacy entails the knowledge of the theories, philosophies and practical uses of assessment instruments which guarantees optimum reliability, validity and efficiency. In the classroom setting, it would lead to student learning enhancement as well as more effective instructional practices informed by the feedback obtained through classroom assessment. Moreover, in any context including classrooms, assessment literacy tends to ensure the highest possible degree of fairness in measuring learner achievement and optimum efficiency on the part of all the personnel involved. A significant amount of efforts, time and money is spent for assessing learners' performance, teachers' practices, language programmes or institutions as well as test-providing bodies or authorities. As Stiggins (1999) has stated, classroom teachers tend to devote around 30-50% of their professional time to assessment-related activities. If the teachers lack the required level of assessment capability, they will not be able to maximize the learning potential of assessment activities.

For all these reasons, assessment literacy is not at all an option, rather an imperative for all assessment personnel. On their part, all language testers must be capable of using assessment to measure students' learning, monitor learners' progress and make necessary changes in their teaching practices. With a view to expatiating on the concept of assessment literacy, the following sections of this paper will explain the fundamental principles of assessment, the procedures of test construction, and some major considerations in test delivery.

2. Key Theoretical Constructs

2.1 'Assessment' and 'Test'

Assessment is a broader concept which subsumes a number of techniques, one of them being tests. It is 'a process for obtaining information for making decisions about students;

curricula, programs, and schools; and educational policy' (Brookhart & Nitko, 2015). This information on students' performance can be collected by using formal tests, assignments, projects, labwork, oral presentation, formal or informal observation, and so on. Conversely, a test is a tool which is formally and systematically used to elicit evidence of students' language abilities. The formal examinations in educational institutions and recruitment and university admission examinations are some examples of tests. Thus, there is an obvious risk in regarding assessment and test as synonymous. Notwithstanding the distinction between the two terms and the potential risk in using them synonymously, these two terms are often used interchangeably to indicate the action of measuring students' learning. These two terms have been used likewise throughout this paper.

2.2 Validity

Validity is the most complex theoretical construct in assessment literature, and there has been a profusion of disagreements about the definition and scope of validity. From a traditional viewpoint, a valid test 'measures accurately what it is intended to measure' (Hughes, 2003, p. 26), and three major forms of evidence for validity are related to content, criterion and construct. However, Messick (1989) has contended that content- and criterion-related evidences contribute to interpretation of test scores, and therefore are aspects of construct-related validity. Hence the term 'construct validity' has recently been used to refer to the overarching concept of validity (Hughes, 2003, p. 26).

Messick (1989) has defined validity as 'an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy or appropriateness of inferences and actions based on test scores or other modes of assessment' (p. 13). This view of Messick brings about a radical change in the common understanding of validity. Validity is now not a characteristic of a test; rather, it is 'the degree to which we are justified in making an inference to a construct from a test score' (Fulcher & Davidson, 2007, p. 12). That is, a test is valid if any predictions, indications and inferences or decisions based on a test score, for example, are justified.

Content validity: A test has content validity 'if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned' (Hughes, 2003, p. 26). If the items included in a test represent the whole range of content areas or learning objectives, it is said to have content validity. A test will lose its content validity if a portion of the specified content or learning objectives are underrepresented or not covered at all.

Criterion-related validity refers to the degree to which test results agree with the results provided by some independent and highly dependable assessment of the candidate's ability, where the independent assessment is the criterion-measure against which the test is validated (Hughes, 2003). Criterion-related validity is of two types: concurrent and predictive validity. When a test is given at the same time the independent assessment or the criterion measure is administered to validate the test, evidence of its concurrent validity is gathered. For example, two tests are designed to measure the achievement of 20 learning objectives of a language course: one larger with 20 or so items covering all 20 objectives and the other shorter with only 10 items, representative of all the course

objectives. If the scores of the candidates on the two tests tend to agree to a considerable extent, the shorter version of the test will be considered to have concurrent validity, and the longer version will be the criterion.

Predictive validity refers to the degree to which test scores can predict the future performance of test-takers. A 'cut score' on a placement test (e.g. university admission tests in Bangladesh) used to predict a candidate's ability to cope with the first-semester undergraduate courses concerns predictive validity of the test. If a candidate obtains the 'cut score' or above, it is likely that (s)he will be able to perform satisfactorily in the first semester undergraduate courses.

Construct validity means 'the concomitance between the test and the underlying teaching principles' (Basanta, 2012, p. 34). That is, the test must conform to the theoretical constructs that underlie the teaching-learning-testing network. The tasks on the test must be consistent with the theoretical constructs it claims to provide information about so that the scores of the test can be interpreted to indicate 'what is valued in performance on the test' (Fulcher & Davidson, 2007, p. 13). If a test is designed to measure the achievement outcomes of learners who were taught English following the CLT approaches, it must adhere to the principles of communicative language testing; otherwise, it will lose construct validity. A vocabulary quiz asking learners to write down the definition of words given will not be consistent with communicative language testing because it will not require them to consider using words in context which is an essential construct of communicative language use.

Validity in scoring: A test has to be scored validly. For example, if marks are deducted for spelling and grammatical errors in short responses written by candidates on a reading or listening test, the scoring becomes invalid because a reading test is supposed to assess candidates' reading skills only, not writing skills.

Face validity implies that a test must look as if it measures what it intends to measure; if it looks as such, it is said to have face validity. Mousavi (2009) has defined face validity as 'the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgement of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers' (p. 247). If a test does not look right to test-takers, it might cause fluctuations in the candidates' confidence and eventually affect their performance on the test. Candidates' perception of fair tests can be increased by using expected and well-constructed formats, uncomplicated items, clear instructions, and tasks related to course-work and with a reasonable level of difficulty (Brown & Abeywickrama, 2019, p. 38).

2.3 Reliability

Reliability is concerned with the consistency of test scores. A candidates' score on a reliable test would be more or less the same if (s)he were to take the same test on two different occasions in two different settings. Major factors that might affect the reliability of test scores include: 1) fluctuations in the learners (e.g. illness, fatigue, anxiety, stress, etc.), 2) subjectivity or mechanical errors in scoring, 3) inconsistent administrative procedures and assessment conditions, and 4) flaws in the test itself (Coombe & Hubbley, 2009; Brown & Abeywickrama, 2019).

First of all, candidates might not be able to perform according to their competence or preparedness because of sickness, fatigue, anxiety, stress or some other physical or mental problems, and hence their scores might not be reliable. Again, some candidates might perform better or worse due to variations in their test-wiseness. However, the test providers probably have very little to do in such cases. Secondly, the reliability of test scores might be affected by the scorers' 'lack of adherence to scoring criteria, inexperience, inattention, or even perceived biases' (Brown & Abeywickrama, 2019, p. 30). Human or mechanical errors can also hamper reliable scoring. Thirdly, inconsistency in test administration and inauspicious test environment might lead to unreliable test scores. Factors relevant to this source of unreliability include external noise, poor light in different parts of the examination hall, faulty instruments (e.g. record players, computers, etc.), variations in hall temperature, poor acoustics of the examination hall, condition of tables, chairs or desks, lax or over-strict invigilation, power failure, variations in photocopying or printing quality, and such-like. Finally, poorly designed tests containing faulty items, typos, unclear instructions, lack of focus, etc. are sure to provide unreliable scores. Decisions taken based on unreliable test scores will be wrong, which in turn will affect test validity.

2.4 Authenticity

Test tasks would be authentic if they reflect the tasks, situations and contexts of real life. According to Bachman and Palmer (1996), authenticity is 'the degree of correspondence of the characteristics of a given language test task to the features of a target language task' (p. 23). That is, it refers to the extent of similarity between what the activities on a test require candidates to do in a target language and what the native speakers of that particular language do with it in real–life situations. Test designers should include test tasks that best simulate real-world tasks. They can do so by using language that is as natural as possible, contextualized test tasks, meaningful, relevant and interesting topics, and activities that replicate real-life tasks (Brown & Abeywickrama, 2019).

2.5 Interactiveness

An interactive test offers test tasks that interact with the test takers. That is, the completion of the tasks requires the involvement of the candidates' individual characteristics that include their language ability, knowledge of the topics of test tasks, and affective schemata (Bachman & Palmer, 1996). The more interactive the test tasks are, the more valid the test will be.

2.6 Practicality

Practicality is related to test administration. A practical test is administration-friendly to a considerable extent. A test is considered practical if its construction, implementation and scoring do not involve much time or money (Basanta, 2012). Test practicality is concerned with issues in assessment like 'cost of development and maintenance, time needed to administer and mark the test, ease of marking, availability of suitably trained markers, and administration logistics' (Rogier, 2014, p. 5). If the delivery of a test exceeds the budgetary limits or requires such a long time that it is unmanageable for both candidates and administrators, it will be impractical. Again, if the scoring of the test requires an unusual proportion of time and effort on the part of the scorers, it will be impractical too.

2.7 Transparency

A test would be transparent only if the candidates have full access to detailed information about all aspects of the test. Test-takers should be aware of the test content (e.g. structures, vocabulary, topics, etc.), test technique and format (e.g. MCQ, essay, reports, role play, etc.), modes of answering (e.g. pencils or pens, paper-based or computer-based, word limits, etc.) and scoring procedures (e.g. marks allocation, scoring rubrics, deduction of marks for wrong answers, number of scorers, etc.). Any lack of information about the test might cause anxiety and diffidence in candidates, which can lead to their poor performance. Conversely, if they are well informed about the test, their anxiety decreases; their confidence increases; and thus they are likely to perform better.

2.8 Security

If a test, or part of it, is leaked out before being administered, the test loses reliability and validity. A well-designed test is reduced to nothing if it is leaked out before it is delivered. Therefore, the test-providers have to adopt airtight security measures to ensure test security.

2.9 Impact and Washback

Test impact refers to the effects of tests on society as a whole including candidates, teachers, test developers, educational systems, guardians, and so on (Hughes, 2003; Bachman & Palmer, 1996; McNamara, 2000). One aspect of impact is washback or backwash which refers to the beneficial as well as harmful effects that testing has on teaching and learning (Hughes, 2003; Alderson & Wall, 1993).

A test might positively or negatively influence what and how teachers teach in the classroom as well as what and how students learn. For example, if teachers teach with the entire focus on how they can help their students prepare for the test, a considerable portion of the course syllabus or curriculum might not receive adequate attention. Thus learning might suffer considerably. Further, the information gathered from the test might not present the real picture of the students' abilities. As a result, the washback will be harmful. In contrast, when 'testing and curriculum are based on clear course outcomes that are known to all students and teachers' (Rogier, 2014, p. 6), the test effects are considered to be beneficial. For example, if a test requires students to write essays, teachers and students will focus on writing essays integrating all skills needed. Thus, the test effects on both teaching and learning will be beneficial. To ensure beneficial backwash, teachers have to link teaching and testing to instructional objectives so that their tests reflect the goals and objectives of the courses, and test techniques match the types of activities used in teaching the content (Rogier, 2014).

2.10 Usefulness

According to Bachman and Palmer (1996), test usefulness is a comprehensive notion which refers to 'a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test' (p. 18). That is, a useful test must embody an appropriate balance of all the test principles. Therefore, when a test achieves all the test qualities to the greatest possible extent in a complementary combination, and not with the rejection or absence of any of the qualities, it is said to be useful.

3. Test Design

Equipped with an understanding of the key concepts in assessment, testers have to consider putting their theoretical knowledge into the actual practice of designing or developing useful tests. To do that, they have to carefully follow a set of procedures. Test development is the entire process of planning, writing and using a test. It involves teamwork and is done through a series of procedural steps which are not sequential or linear all the time, rather iterative at times. It is impossible for an individual to accomplish the stages in developing a test that embodies all desired test qualities. Again, not all the stages of test development are completed with equal investment of time, effort, rigor and formality. As Bachman and Palmer (1996) has pointed out:

At one extreme, with low-stakes tests, the processes might be quite informal, as might be the case if one teacher were preparing a short test to be used as one of a series of weekly quizzes to assign grades. At the other extreme, with high-stakes tests, the processes might be highly complex, perhaps involving extensive trialing and revision, as well as coordinating the efforts of a large test development team. This might be necessary if a test were to be used to make important decisions affecting a large number of people. (p. 89)

Ten steps in test development

The following section of this article will focus on the general procedures of developing a test, taking Hughes (2003) as the informing source. His ten-step procedures in language test construction will frame the basis for this brief discussion on test design.

Step 1: Stating the problem

First of all, test writers have to decide on what they are required to test and what purpose the test is going to serve. They have to consider what type of test they are going to construct (e.g. final or progress achievement, proficiency, placement or diagnostic). They should also consider how detailed the test results must be. Additionally, the possible wash back and constraints related to the availability of expertise, facilities and time in designing, delivering and scoring of the test are also to be taken into account.

Step 2: Writing test specifications or blueprints

Test specifications are written documents that contain comprehensive information required for the creation of a test. Before starting to write a test, the test developers have to prepare a test blueprint which would include detailed information on *test content and objectives* (e.g. tasks that candidates have to accomplish; text types, e.g. letters, essays, etc; length and difficulty level of texts; range of topics; vocabulary range; and so on); *test structure* (number of sections, number of items in each section, marks distribution, number of passages), *timing* (for the full test as well as individual sections), *test techniques* (MCQ, fill-in-the-gap, short answer questions, role play, etc), *criterial levels of performance* (e.g. 80% for A+, 75% for A, 40% for passing, etc.) and *scoring procedures* (scoring rubrics, number of raters, etc.). Using the test specifications, test-writers can ensure that they have included test tasks or items that represent the entire content of a course or the whole range of course objectives. Thus, the test can measure what it is supposed to measure.

Step 3: Writing and moderating items

Test items should be written following the specifications. Although it is not possible to include all the specified content areas into a single test, item-writers should carefully choose from the content so that the items represent the entire content. Then the items should be scrutinized by at least two moderators who have not taken part in writing the items. The items should then be modified addressing the weaknesses identified by the moderators. Using moderation checklists might make their job easier.

Step 4: Informal trialing of items on native speakers

After moderation, the test should be trialed informally on a group of native speakers of more or less the same age, education and general background as the potential candidates. Considering the feedback gained from this informal trial, test-writers should modify the test items where needed. Although this informal trialing of test items is indispensable for some standardized international benchmark tests like IELTS and TOEFL, it is not required in other cases, e.g. classroom quiz, for different reasons.

Step 5: Trialing of the test on non-native speakers

After the informal trial, the test is to be formally administered under operational conditions to a group of non-native speakers similar to the potential candidates for whom it is being developed. Through this field trial, the likely challenges in test delivery and scoring can be detected. Again, this type of trial might not be possible for the unavailability of non-native speakers with required characteristics as well as for security reasons.

Step 6: Analysis of results of the trials and modification of test items

The results of the trials are to be analysed, and required changes are to be made to test items as well as to administration procedures in the light of the feedback gathered.

Step 7: Calibration of scales

In case of using rating scales for speaking and writing skills, samples of candidates' spoken and written performance are to be collected and assigned to all the points on the rating scale. For example, sample of paragraphs or essays written by candidates can be assigned to letter grades like A+, A, A-, etc. so that the sample answers can be used as reference points by examiners.

Step 8: Validation

The final version of the test must agree with the specifications prepared at the very outset of its development. Experts have to check whether the test will be able to assess what it is supposed to assess, and whether all major content areas or learning objectives have been covered, and whether it will provide reliable information on candidates' language abilities. Validation is a crucial step in test construction, and it is a must for high-stakes tests.

Step 9: Writing handbooks or test manuals

Handbooks should be written for candidates, test users and testing staff so that they all can have detailed information on the purpose of the test, procedures of its construction, description of test candidates, and administration and scoring procedures. Although the

writing of test manuals is mandatory for high-stakes tests, it is often not needed in some contexts where timely oral instructions are considered sufficient.

Step 10: Training of staff

Adequate training should be provided beforehand to all staff involved in the process of testing including invigilators, raters, computer operators, exam hall attendants, and so on.

4. Test Delivery

The delivery of tests is as important as the planning, writing, moderation and validation of tests. A test prepared with utmost care and in the fullest possible conformity with the theoretical constructs might not be able to provide reliable and valid information about test-takers' performance if it is not administered in appropriate ways. As Douglas (2011) has stated, 'any of the elements of test administration can potentially lead to problems with reliability and cause our interpretations of test-takers' performance to be erroneous' (p. 54). Hence, the administration of any test requires careful thought and planning. However, the procedures to be followed in administering all tests are not equally complex. While the delivery of a large-scale high-stakes test is exceptionally complicated, that of a class test or quiz is rather easy or less complicated. Such variance in the level of complexity does not necessarily allow of any slightest laxity in administering a test on the part of the testers.

All-out attempts must be made to ensure that every requirement for the administration of a test has been fulfilled. To ensure optimum performance of the candidates, a sound professional attitude of all involved in test delivery might help to overcome all inherent or systemic limitations of the actual administration of tests. In order for a test to be successfully administered, adequate attention should be given to some important considerations related to test environment, personnel involved in giving the test, and the procedures to be followed from the beginning to the end of test delivery.

4.1 Test Environment

Creating and maintaining a testing environment conducive to candidates' optimum performance is of paramount importance. If the environment of the examination hall is distracting in some ways, the test-takers cannot concentrate properly, which hampers their performance on the test. Consequently, inferences made on the basis of the test results become erroneous at the expense of test reliability and usefulness. To ensure that the overall test environment is congenial for optimum test performance, attention and care should be paid to the following:

1. All materials (test papers, paper, pencils, etc.) and equipment (multimedia, loudspeakers, tape recorders, microphones, etc.) should be kept ready for use and in working order well before the test is administered.
2. The examination room should be quiet, and there should be good lighting, comfortable seating arrangements for test-takers, and adequate space between students, desks or tables.
3. Clocks should be visible to all candidates. It is to be ensured that test-takers can know the time whenever they need to without disrupting the concentration of others.

4.2 Personnel

The personnel involved in test administration, including administrators, examiners, inspectors, technical aides, and other support staff, play a crucial role in successful test delivery. Candidates' performance on a test might be hampered because of the negligence of test personnel in carrying out their respective responsibilities or their ignorance or unprofessional attitudes and behaviour in the examination hall. For example, if some invigilators talk loudly among themselves or shout at candidates trying to communicate with others, it will disrupt the candidates' concentration and thus affect their test performance. The administration personnel must be knowledgeable about their responsibilities and have the good intention to carry out their duties. The following are some important considerations related to test delivery personnel:

1. There should be 'an adequate number of people on hand to help seat test-takers, if there are large numbers, pass out test booklets and other materials, monitor the test-takers, and provide computer and other equipment support if necessary' (Douglas, 2011, p. 54).
2. The personnel should receive adequate training.
3. Detailed guidelines should be prepared for and provided to the administration personnel, and they should read and understand the instructions thoroughly so that they can perform their duties properly.
4. Those who will be required to use any equipment should familiarize themselves with its operation beforehand (Hughes, 2003, p. 216).

4.3 Administration procedures

The actual delivery of a test is required to follow a set of well laid-out procedures. A successful execution of these procedures facilitates optimum test performance. Consequently, reliable data can be gathered, and a valid interpretation of candidates' performance becomes possible. In order for the administration procedures to be smooth and effective, the following points should be carefully considered:

1. The identification of candidates should be done very carefully by checking relevant identification documents. It has to be ensured that candidates have been seated in their designated places and are using their own test papers or scripts, not those of others.
2. Test-takers should be required to report to the examination hall well before the test starts. Latecomers should be treated following the relevant instructions mentioned in the specification so that other candidates' concentration is not distracted. Latecomers should be allowed to enter the examination hall only up to the stipulated time, and not after that.
3. Adequate distance should be maintained between seats so that test-takers cannot communicate between themselves to pass information to each other.
4. Invigilators should read out to candidates the written instructions about what they are required to do, what they are allowed to do, and what they are not allowed to do during the test. Candidates should also be briefed about what consequences any irregularities on their part would result in.

5. Throughout the test, invigilators should monitor test-takers' behaviour without causing any distraction. Invigilators should behave with the test-takers with utmost care, respect and politeness. An impression should be created that all security procedures are meant to help candidates' concentrate to the fullest possible extent and thus maximize their performance on the test. Invigilators are not allowed to treat any candidates rudely or shout at them even when they are found to be cheating in any form; rather, the irregularity issues should be dealt with utmost care and silence so that the overall milieu remains fairly undisturbed.
6. Invigilators themselves should distribute test materials to each test-taker individually. Test-takers should not be made to distribute or pass test materials.
7. Test-takers should be instructed to provide required information (e.g. examination roll-number, date, and venue) on test papers. Invigilators should check whether the candidates wrote the details in designated places.
8. Invigilators should maintain the time strictly. All test-takers should start and stop writing at the same time.
9. Invigilators should make sure that once the test is over, all the test papers have been collected and counted, or all necessary files have been saved on computers. Only this being done, should the candidates be allowed to leave the examination hall.

4.4 Scoring and Communication of Results

Now that a good test has been administered with the fullest possible care and caution, and the test materials used by candidates have been collected, it is to be ensured that the test papers are rated properly, and the results have been communicated to candidates in due manner. Otherwise, the efforts employed in the previous phases will be all futile, and the test results will be unreliable. To ensure fair scoring of test papers and smooth communication of test results, the following points should be kept in mind:

1. Scoring keys should be prepared for selected response items, including MCQ, yes/no/not given, true/false, matching, ordering, information transfer, gap-fill, etc. In case of automated scoring, required programmes should be carefully installed on computers, and it should be checked whether they are functioning properly.
2. Scoring rubrics or criteria should be prepared, clearly noting down all language abilities to be assessed and their weight in rating.
3. Raters should be trained so that they can apply the scoring keys or rubrics consistently. In case of automated scoring, the computer operators should be given adequate training as well.
4. At least two raters should be used in rating extended writing and speaking tests, and their scores are to be averaged to arrive at the final score. The maximum difference between the scores given by two individual scorers is to be stipulated beforehand, and if the difference between two scores exceeds the stipulated limit, it should be referred to a third rater whose score would be averaged with one of the previous two scores which is closer to it. No rater should have any idea about the rating of others.
5. The final results should be recorded in the required documents, and communicated to the candidates within the scheduled time. Candidates should be allowed to lodge their

complaints if they think their results need to be further scrutinized. The concerned authorities should cooperate with the candidates in all possible ways to avoid any discrepancy that might put reliability of test results to question.

5. Conclusion

Assessment plays a vital role in building a bridge between teaching and learning. Language test developers must have a sound understanding of the key principles of assessment as well as the procedures of test construction and administration so that they can collect dependable evidence of test-takers achievement of learning contents or objectives. Without a solid understanding of the theoretical constructs of assessment, it is impossible for test-designers to develop a good test. However, even a good test will fail to provide reliable evidence of candidates' true abilities if it is not delivered in the proper way. After the test is given to candidates, the test papers or files are to be collected and scored reliably, and results should be communicated to the candidates through proper channels in due time. The absence of assessment-literate personnel in any of the stages will culminate in test results being unreliable, and test purpose remaining unattained. Having focused on the conceptual as well as pragmatic issues in assessment, this paper would hopefully foster some awareness about meaningful assessment literacy among the novice language teachers as well as other stakeholders of assessment and testing across disciplines.

6. References

Alderson, J. C. and Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14, 115-129.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Basanta, C. P. (2012). Coming to grips with progress testing: Some guidelines for its design. *English Teaching Forum*, 50(3), 37-40.

Brookhart, S. M. and Nitko, A. J. (2015). *Educational assessment of students* (7th ed.). Boston, MA: Pearson education, Inc.

Brown, H. D. and Abeywickrama, P. (2019). *Language Assessment: Principles and classroom practices*. Hoboken NJ, USA: Pearson education, Inc.

Coombe, C. and Hubbley, N. (2009). An Introduction to Key Assessment Principles. In Coombe, C., Davidson, P. and Lloyd, D. (Eds). *The Fundamentals of language assessment: A practical guide for teachers*. Dubai, UAE: TESOL Arabia Publications. 3-10.

Douglas, D. (2011). *Understanding language testing*. UK: Hodder Education.

Frank, J. (2012). The role of assessment in language teaching. *English Teaching Forum*, 50(3), 32.

Fulcher, G. and Davidson, F. (2007). Language testing and assessment: An advanced resource Book. Oxford and New York: Routledge.

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

McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.

McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.

Messick, S. (1989). Validity. In R. Linn (Ed), *Educational measurement* (pp. 13-103). New York, NY: Macmillan.

Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing* (4th ed.). Tehran, Iran: Rahnama Publications.

Rogier, D. (2014). Assessment literacy: Building a base for better teaching and learning. *English Teaching Forum,* 3, 2-13.

Stiggins, R. (1991). Assessment literacy. *Phi Delta Kappan*, 72(7), 534-539.

Stiggins, R. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), 23-7.