Assessment in English Language Teaching and Learning

Yu LIU
Xi’an University of Technology, Xi’an, Shaanxi, China
414453892@qq.com

Keywords: Assessment, Teaching, Learning.

Abstract. “Assessment is an integral aspect of the pedagogical process of designing lessons, implementing them, and evaluating their success.”[1]. Through assessment, teachers can get feedback from students about whether their teaching are effective or not and do they accomplish the aims and objectives that they set for the lessons. It works not only on teachers but also on students. This paper will examine some issues that we need to take into account when conduct an assessment with the comparison of the CEFR Global Scale and MASUS Procedure. With appropriate assessment, the English language teaching and learning can be more effective. Before the lessons, teachers can make sure that students are on the right level of class. During the lessons, teachers can get the detailed information of the students themselves. Whether they grasp a certain grammar point or whether they understand a particular activity etc. After the class, teachers can check whether students reached the aims and objectives of the programs through assessment and which level they are reached.

Introduction to CEFR Global Scale and MASUS Procedure

The CEFR Global Scale refers to the Global Scale of ‘Common European Framework of Reference for languages: Learning, Teaching, Assessment.’ Here, the Global Scale is not ‘language-specific’. That is to say, it can be used for any languages in the world and can be used to ‘compare achievement and learning across languages’. The CEFR Global Scale ‘provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc.’[2].

When using this to make an assessment, some issues arise. Firstly is about the broad nature of the framework. The CEFR Global Scale describes what a student can do at six specific levels, from basic level to proficiency level. Here, it uses the words like ‘effective’ etc. to identify different levels of students. ‘The CEF deliberately does not refer to grammar or structures. It is designed to describe how language users communicate and how they understand written and spoken texts.’[2], Therefore, there is no detailed information about grammar structures etc. Teachers can only make a general assessment of the students’ work due to limited information. Secondly, only one piece of written or spoken text of each student available for assessment is far more not enough under this kind of broad framework.

The MASUS Procedure is a diagnostic assessment. It is a forward-looking assessment and usually used like a needs analysis process. This assessment used to identify what students can and cannot do in a specific context. Just as Bonanno and Jones noted in their article, ‘The whole MASUS Procedure enables the creation of a literacy profile of a student cohort in terms of a set of specified criteria.’[3]. Through this assessment, we can know about the students’ detailed learning outcomes in terms of the whole structure of the text and the grammar or vocabulary, etc. This is more specific compared with CEFR Global Scale.

Compared with CFFR Global Scale, one of the issues arise when using the MASUS Procedure is that although there are ratings within each criterion but there is no specific explanation about each rating. Thus, teachers may judge the ratings subjectively. However, the specific describing of each criterion, such as the organization of genre and the grammar accuracy, avoid this kind of subjective to some extent when making assessment.
Issues Arising in Assessments

When we make assessment decisions, some issues will be arose, particularly validity, reliability, washback and practicality. Here, I will discuss the influence of these issues on CEFR Global Scale and MASUS Procedure as well as their effects on designing and conducting a written or spoken test, etc.

Both of these two assessment instruments are criterion-referenced assessment. It ‘depend(s) upon an absolute standard of quality.’ That is to say, each student should reach a certain level of competence which is ‘independent of reference to the performance of others.’[4]. In addition, one important factor of criterion-referenced tests is that it is ‘not high discrimination, but to represent a continuum of relevant tasks.’[4].

Validity

According to Gipps (1994), validity in criterion-referenced assessment is mainly content validity. ‘Content validity is defined as any attempt to show that the content of the test is a representative sample from the domain that is to be tested.’[5]. In terms of the CEFR Global Scale, it is not limited to any one course, curriculum, or single skill in the language. It assesses students from a broader perspective without specific criterions. Thus, it surely has content validity weaknesses. However, the MASUS Procedure focuses on the specific criteria, such as the structures of the text and the grammar accuracy, etc. Therefore, it presents the content validity of the assessment.

‘When considering criterion-oriented validity, the tester is interested in the relationship between a particular test and a criterion to which we wish to make predictions.’[5]. It can be divided into two types: predictive validity and concurrent validity.

‘Predictive validity is the term used when the test scores are used to predict some future criterion, such as academic success.’[5]. The CEFR Global Scale was used as an entry to a course or a program. It describes what students will be able to do in the future, such as students ‘can describe experiences and events etc.’[2]. Therefore, if it is well used, the predictive validity will be high. However, the MASUS Procedure assesses the students’ performance at the end of a unit of work. It just summaries the students’ strengths and weaknesses and builds learning objectives for the future lessons while cannot predict what students actually will be able to do in the future. Thus, it is lack predictive validity to some extent.

As for concurrent validity, the CEFR Global Scale is too broad to assess. We cannot determine the concurrent validity of CEFR Global Scale just depend on one short piece of students’ written or spoken test. For MASUS Procedure, it gives us the detailed criteria to assess and also the ratings of each assessment aspects. Thus, it did better on concurrent validity than CEFR Global Scale.

When it comes to construct validity, ‘every issue in language learning and teaching involves theoretical constructs. Tests are, in a manner of speaking, operational definitions of such constructs in that they operationalize the entity that is being measured.’[1]. The MASUS Procedure assesses students on the organization of genre, organization of discourse, grammar accuracy and qualities of presentation. The justification of these four aspects lies in a theoretical construct that claims those aspects as major components of writing or speaking proficiency. Thus, the construct validity of MASUS Procedure is very high. If there is one aspect that does not involved in MASUS Procedure, then its construct validity will be doubted. For CEFR Global Scale, the broad nature of it determines its poor construct validity.

In terms of face validity, if a test looks like it tests what it is designed to test, then it has face validity. According to Brown (2007), ‘Face validity is almost always perceived in terms of content,’ thus a lack of content validity and the broad nature determines that the CEFR Global Scale has poor face validity.[1] However, from the detailed assessment criteria of the MASUS, we can see that it certainly has face validity. For example, the criteria of the organization of genre and discourse use as well as the grammar and vocabulary etc. These are all what we need in English language teaching and testing.
Reliability

According to Brown (2007), a reliable test is consistent and dependable. A number of sources of unreliability may be identified:

- The test itself (its construction), known as test reliability
- The administration of a test
- The test-taker, known as student-related reliability
- The scoring of the test, known as rater (or scorer) reliability [1]

For CEFR Global Scale, it is too broad to determine its test reliability. For MASUS Procedure, there are specific criteria to assess, but the ratings of each criterion are a little bit vague. This may have a bad effect on MASUS Procedure’s test reliability. For both of them, the test administration reliability for example may affect by the surrounding environment of the examination room. And it can also affect by the student themselves, which is called student-related reliability. The test-takers may feel nervous or have an illness that day etc. In addition, both CEFR Global Scale and MASUS Procedure are assessed by raters or scorers, which is too subjective. Even though MASUS Procedure gives the clear criteria for assessment, then how to exactly define ‘effective’, ‘appropriate’ etc. Therefore, both of them lack rater (or scorer) reliability. For improving rater reliability, we can do from two aspects. That is inter-rater reliability and intra-rater reliability. For the CEFR Global Scale, it is much easier to improve on inter-rater reliability with each student’s work assessed by two or three raters. For MASUS Procedure, it also can be improved on inter-rater reliability, but the improving on intra-rater reliability is much easier. The teachers can assess the students’ work by two or three times in different period of times. Moreover, as Gipps (1994) noted in his article, ‘improved reliability may ensue from tightening up the specifications and targeting marking in relation to grade descriptions.’ [4].

Washback

In simple terms, ‘the effect of testing on teaching and learning is known as backwash.’ [6]. It can be either positive or negative.

According to Council of Europe Modern Languages Division (2001), ‘the Common European Framework provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc.’ Therefore, the CEFR Global Scale has an impact on ‘learners, teachers, course designers, examining bodies and educational administrators.’ [2]. First for learners and teachers, this framework can help students to do self-assessments and be aware of what levels they are present in. It also can help teachers have a general sense of the aims and objectives for a particular course or program they will conduct if the CEFR Global Scale was used for an entry of a course or a program. For the ‘course designers, examining bodies and educational administrators’, although this framework is very broad, it also have an effect on these entities to fit their texts to the six bands of this framework and this in turn give them more freedom to design their texts and assessments. This is what van Lier called ‘washforward’ [6, van Lier quoted in Bailey 1996, p. 263].

As for MASUS Procedure, washback have a more direct effect on students and teachers due to its specification and conducted form. Before the test, the teachers can plan their lessons according to the detailed assessment criteria and students in the same have a guide on their learning. After the test, students can find out their strengths and weaknesses and then work on them and teachers can set a series of future leaning objectives according to the students’ strengths and weaknesses. Just as Hughes noted, ‘The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work.’ [6, Hughes quoted in Bailey 1996, p. 262].

For promoting beneficial washback when designing and conducting this kind of written or spoken test, etc., there are some factors that we need to take into consideration. The first one is language learning goals. The assessment should make sure that it focuses on the course or
program’s goals so that teachers and students can work on it during their lessons. The second one is authenticity. According to Doye, ‘Absolute congruence would exist when tasks in the test situation and in the corresponding real-life situation would actually be identical. In this extreme case the test situation and the tasks in it are called authentic.’ [6, Doye quoted in Bailey 1996, p. 269]. That is to say, the assessment should reflect the real-life and be the interactive communicative operations but not the traditional routine exams. The third one is the learner autonomy and self-assessment. In this aspect, CEFR Global Scale and MASUS Procedure all do well, one from a broader view and one from a detailed view. The last one is the score reporting. To promote beneficial washback, “assessment information must be ‘detailed, innovative, relevant and diagnostic’ and that it must ‘address a variety of dimensions rather than being collapsed into one general score’.” [6, Shohamy quoted in Bailey 1996, p. 271]. For this, MASUS Procedure does better than the CEFR Global Scale as it gives the detailed assessment criteria for teachers and students.

Practicality

Except validity, reliability and washback, practicality is also an important factor that we need to take into account when designing and conducting an assessment. When judge whether a test is practicality or not, we need to take into consideration of the ‘financial limitations, time constraints, ease of administration, and scoring and interpretation.’ [1]. In addition, ‘the extent to which a test is practical sometimes hinges on whether a test is designed to be norm-referenced or criterion-referenced.’ [1]. The criterion-referenced test gives clear feedbacks to students and teachers due to its specific criteria. The MASUS Procedure does well on this and also it conducts on small groups. Thus it is more practicality than the CEFR Global Scale.

Conclusion

Assessment is a complex and significant process for both teaching and learning. ‘Without an assessment component in every lesson, every unit, and every course, we couldn’t determine the attainment of objectives and goals.’ [1]. The assessment covers the whole process of teaching and learning. Thus, we should make sure that it manages well to produce valid and reliable results so that it can in turn have a positive effect on teaching and learning. In addition, when we design and conduct assessment, we should take into consideration of its practically and make sure that it connects with the real world. Moreover, for different assessment types, we should choose different criteria that appropriate them. If we do well on all of these, then the role of the assessment will be maximize and the course designers, administrators, teachers and students etc. will benefit a lot.

Acknowledgement

This research was financially supported by the National Science Foundation.

References


