Checklists for Validating an English Test in the Pre-Testing Stage

Jatupong Mora[1*]

[1] Dr, Lecturer, Western Languages Department, Faculty of Humanities and Social Sciences, Thaksin University, Songkhla Campus

***Corresponding author, E-mail:** mjatupong@tsu.ac.th

Abstract

In the pre-testing stage of an English test construction, validating the test is what the test constructors should do. This procedure makes sure that the test has some characteristics of an effective one before it is administered. This paper summarizes the concepts of test validation outlined by English test specialists and proposes checklists for English test constructors, English instructors, and English major student-teachers. They might use this ready-made checklists to validate their tests before administration. It is also hoped that the checklists could be a validating instrument and beneficial for those involving in English test construction to be able to construct effective English tests in the future.

Keywords: test validation, test qualities

Introduction

Validating an English test before administration is what professional English teachers should do. To validate the tests in this stage, what the test constructors or English test professionals can do is to examine some corresponding between the test blueprint and the test paper and to look through the test paper item by item (Weir, 2005: 222). It is more convenient for test reviewers or validators if they have a ready-made validating checklist (Oghabi, Pourdana, & Ghaemi, 2020). Most English test experts usually validate the test by examining six qualities of test in their checklists: validity, reliability, authenticity, appropriateness, impact, and practicality (Bachman & Palmer, 1996: 139-155; Genesee & Upshur, 1996: 241-244, and Brown, 2004: 30-38). This validating tool might be beneficial for the new English test constructors to realize what they should bear in mind if they want to produce effective English tests.

Test validation in the pre-testing stage

Validating an English test in the pre-testing stage focuses on investigating some correspondence between the test blueprint and the test paper and checking if the test items have some characteristics of being good items. To validate the test, Bachman & Palmer (1996: 139-155) and Genesee & Upshur (1996: 241-244) use almost the same categories of test quality. For Bachman & Palmer (1996: 139-155), any good test should contain six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Genesee & Upshur (1996: 241-244) also use six qualities to validate the test, but a few terms are different from Bachman & Palmer's: test content and purpose, appropriateness, practicality, user qualities, reliability, and validity. However, Brown (2004: 30-38) asks six questions to validate the test. According to Brown (2004: 30-38), an effective English test might be seen if the most answers to these six questions are 'YES'. These validating questions are 1) Are the test procedures practical?, 2) Is the test reliable?, 3) Does the procedure demonstrate content validity?, 4) Is the procedure face valid and "biased for best"?, 5) Are the test tasks as authentic as possible?, and 6) Does the test offer beneficial washback to the learner? The following sections outline the six qualities an effective English test should have in more details. Each quality is defined to have practical definition and to be easy to understand, and the sublists of each quality are identified to obviously see what the test constructors need to focus on when they are writing a test.

One important characteristic of any good tests is *validity* which can be defined as a quality of a test showing that 1) it measures language ability or sub-skills as defined by its objectives, 2) the test scores can be interpreted as the test takers' areas of language ability to be measured, or the scores can be used as inference of the test takers' target language use (Brown, 2004: 30-38; Bachman & Palmer1996: 139-155). In a classroom setting, test validity might be validated with six to seven sublists (Kadir, Zaim, & Refnaldi, 2019; Hien Huong, 2020). Kadir, Zaim, & Refnaldi (2019) developed six sublists to assess content validity of authentic assessment for speaking skill at junior high school: 1) The authentic speaking assessment is represented the learning objectives in general, 2) The authentic speaking assessment is represented the learning objectives in specific, 3) The authentic speaking assessment is assessed the basic competences needed to mastered, 4) The authentic speaking assessment is assessed the language functions needed to mastered, 5) The authentic speaking assessment is assessed the learning topics needed to mastered, and 6) The authentic speaking assessment is assessed genre-based texts needed to mastered. In addition, Hien Huong, (2020) investigated face validity of the institutional English test based on The Common European Framework of Reference at a public university in Vietnam by using the seven sublists: 1) Weightage for the test was appropriate, 2) Time allocation of the test was sufficient, 3) Language skills taught were sufficiently represented in the test, 4) Topics taught were sufficiently represented in the test, 5) Questions in the test was clear, 6) Instructions explaining what to do in each section of the test were clear, and 7) Marks allocated for each section of the test were stated clearly.

The test should contain *reliability* which refers to reliable or consistent test results. When administered to the same group of the examinees, the tests should have the same test results no matter how many times, when or in what setting they are tested (Hughes, 1989; Heaton, 1990). Any tests can be claimed to be reliable if they have these properties. First, the test instructions should be clear and understood by all candidates. The instructions of each test task or part are written in concise for examinees to be clear in the test rubrics. Second, the scoring method is relatively fair for all test-takers, especially in the listening and reading parts. Although the writing and speaking parts seem rather subjective in grading, the criteria for grading writing and speaking tests are designed carefully to minimize the subjectivity from test scorers (Chanthawee & Rungruang, 2020; Kelly 2020). In addition, test formats are familiar to all examinees. The environment and time of examination is suitable and appropriate to make the test-takers comfortable and can do exam at their full-potential (Bachman & Palmer, 1996; Genesee & Upshur, 1996, and Brown, 2004). In a classroom setting, test reliability might be investigated by these eight sublists: 1) All pages of the test can be read clearly to all, 2) All audio files can be open and are loud enough for all when administering in a large testing room, 3) Video input is equally visible to all, 4) Each of objective test items has only one correct answer, 5) There is a marking criteria to mark subjective test items, 6) The test rubric is clear to all test takers to inform what they have to do and everyone has a fair chance to give correct answers, 7) There are some example test items for the part which might look unclear for the test-takers, and 8) The test booklet and answer sheet have layouts that facilitate comprehension and responding.

The tests should have *authenticity* referring to corresponding degree of the test task characteristics to the features of target language use tasks (Bachman & Palmer, 1996). The test tasks in all parts should be neatly constructed to have high authenticity. For example, the listening parts should be corresponded to what the test-takers are going to listen in their future work or daily life. The reading parts so as the speaking and writing parts, should also contain what the test-takers have to use in their work or daily life. In a classroom setting (Brown, 2004), test authenticity might be reviewed with five questions: 1) Is the language in the test as natural as possible?, 2) Are items as contextualized as possible rather than isolated?, 3) Are topics and situations interesting, enjoyable, and/or humorous?, 4) Is some thematic organization provided, such as through a story line or episode?, and 5) Do tasks represent, or closely approximate, real world tasks?

The tests should have *appropriateness*, which refers to involving degree of the test takers' personal characteristics (areas of language knowledge, topical knowledge, and affective schemata) in accomplishing the test tasks. The tests will have high appropriateness if accomplishment of the test tasks result from their background knowledge, topical knowledge and effective schemata. All the test tasks should be designed with taking the candidates' background knowledge, topical knowledge, and effective schemata into consideration (Bachman & Palmer, 1996; Genesee & Upshur, 1996, and Brown, 2004). Therefore, the tests will certainly possess high appropriateness. In a classroom setting, test appropriateness might be examined with four sublists: 1) The test tasks contain the topical knowledge that all test-takers are familiar with, 2) The test tasks are related to the personal characteristics of the test takes (age, grade level, and educational background), 3) The language sub-skills tested are related to the level of knowledge of the test-takers, and 4) The test difficulty is challenging to all test-takers, not too easy or not too difficult.

Another quality any good tests should have is called *impact*. It refers to the benefits the test takers, educational systems, and society have after the test takers have taken those tests. First, the test takers might get benefits from doing this test. They can use some knowledge of taking the test in their working life. Furthermore, institutions also get some benefits from the test results. The program planning team can use the results to improve some change in the curriculum. The impact of the tests for the society may be when the test-takers graduate. The organizations or workplace will get quality workforce to be in their workplace and help develop their organization and society (Bachman & Palmer, 1996; Genesee & Upshur, 1996, and Brown, 2004). In a classroom setting, test impact might be examined with five sublists: 1) Test-takers have got some benefits from taking the test, 2) Test tasks measure language abilities that are included in the teaching materials, 3) Test tasks are similar to teaching and learning activities encountered in class, 4) The purpose of the test is consistent with the values and goals of the teachers and of the instructional program, and 5) Test tasks can reflect test-takers language abilities currently demanded by society and educational system.

The last quality of any good tests is *practicality* which can be defined as the degree of possibility that the test will be constructed and implemented successfully by using the available resources including human, material, and time (Bachman & Palmer, 1996; Genesee & Upshur, 1996, and Brown, 2004). In a classroom setting, test practicality might be examined in four areas. For example, Kadir, Zaim, & Refnaldi (2019) developed instruments for evaluating practicality of the authentic assessment for speaking skills at Junior High School; they investigated practicality in terms of time, media, test procedure, and scoring. Time was evaluated by two sublists: 1) The practical level of the authentic speaking assessment related to time duration in conducting the test, and 2) The practical level of the authentic speaking assessment related to time given to students in doing the test. Media was also reviewed by two sublists: 1) The practical level of the authentic speaking assessment related to the media needed in conducting it, and 2) The practical level of understanding media used in the authentic speaking assessment. Test procedure was examined by four sublists: 1) The practical level of the authentic speaking assessment related to understanding the procedure of conducting it in the class, 2) The practical level of the authentic speaking assessment related to the procedure of conducting it in the class, 3) The practical level of understanding instruction given in doing the authentic speaking assessment by students, and 4) The practical level of the authentic speaking assessment related to the procedure of having students doing the test. Scoring was assessed by three sublists: 1) The practical level of the authentic speaking assessment related to scoring instrument, 2) The practical level of the authentic speaking assessment related to scoring, and 3) The practical level of the authentic speaking assessment related to giving feedback.

Test validation checklists

The following checklist is adapted from the English test specialists' pretesting validation concepts: Bachman & Palmer's (Bachman & Palmer, 1996), Genesee & Upshur's (Genesee & Upshur, 1996), and Brown's (Brown, 2004). The checklist is designed to be more practical and easier to use for any classroom English teachers. The checklist includes six qualities: validity, reliability, authenticity, appropriateness, impact, and practicality.

In order to see what the overall pre-testing validation instrument looks like, all the lists above can be grouped and put in the following form. The validators might be the test constructors themselves if they want to recheck their classroom tests; however, for a high-stake test, for example, an admission test, a school proficiency test, or an exit exam, the test validators should be third-party people who are specialized in English language testing. They can check each item in this form 'Yes', 'Not sure', or No, depending on their agreement to each statement after looking through the test blueprint and the test paper.

Test validation form

Please make a tick (✓) in the box under the column 'Yes', 'Not sure', or 'No' to respond each statement after you have examined the test blueprint and the test paper.

Qualities of an effective English test		Yes	Not sure	No
Validity	1) The test purpose is clearly defined and the test content is in accordance with the test purpose.			
	2) The test objectives are clearly defined and the test content is in accordance with the test objectives.			
	3) Test contents are consistent with what the test-takers learned or experienced in their learning courses.			
	4) The test contains more than one section.			
	5) The test includes a variety of item types.			
	6) Each section is consistent with its objective and has appropriate weight of score.			
	7) The structure of the test is organized logically.			

Qualities of an effective English test		Yes	Not sure	No
Reliability	8) All pages of the test can be read clearly to all.			
	9) All audio files can be open and are loud enough for all when administering in a large testing room.			
	10) Video input is equally visible to all.			
	11) Each of objective test items has only one correct answer.			
	12) There is a marking criteria to mark subjective test items.			
	13) The test rubric is clear to all test takers to inform what they have to do and everyone has a fair chance to give correct answers.			
	14) There are some example test items for the part which might look unclear for the test-takers.			
	15) The test booklet and answer sheet have layouts that facilitate comprehension and responding.			
Authenticity	16) The language in the test is as natural as possible.			
	17) Test items are contextualized as possible rather than isolated.			
	18) The topics and situations are interesting, enjoyable, and/or humorous.			
	19) The tasks represent or are closely related to real-world tasks.			
Appropriateness	20) The test tasks contain the topical knowledge that all test-takers are familiar with.			
	21) The test tasks are related to the personal characteristics of the test takes (age, grade level, and educational background).			
	22) The language sub-skills tested are related to the level of knowledge of the test-takers.			
	23) The test difficulty is challenging to all test-takers, not too easy or not too difficult.			
Impact	24) Test-takers have got some benefits from taking the test.			
	25) Test tasks measure language abilities that are included in the teaching materials.			
	26) Test tasks are similar to teaching and learning activities encountered in class.			
	27) The purpose of the test is consistent with the values and goals of the teachers and of the instructional program.			
	28) Test tasks can reflect test-takers language abilities currently demanded by society and educational system.			

Qualities of an effective English test		Yes	Not sure	No
Practicality	29) The test can be constructed with the resources and budget available.			
	30) Students can complete the test reasonably with the test time frame.			
	31) The test can be administered smoothly without procedural difficulty.			
	32) The test is easy to score and can be completed in the teacher's time frame			

Suggestions: ………

………

………

………

………

………

………………………………………

Validator

Conclusion

The validation checklist proposed in this paper is designed for a convenient use for some new English test constructors. The checklist might contain sublists similar to or different from what some researchers used in their studies. However, the checklist has the same purpose that is to improve the test qualities. In term of validity, for example, item 2) *The test objectives are clearly defined and the test content is in accordance with the test objectives,* is in accordance with Kadir, Zaim, & Refnaldi's content validity sublists 1 and 2. In addition, item 3) *Test contents are consistent with what the test-takers learned or experienced in their learning courses,* covers the content validity sublists 3 to 6 in Kadir, Zaim, & Refnaldi's study (Kadir, Zaim, & Refnaldi, 2019), and is related to the face validity sublists 3 and 4 in Hien Huong's study (Hien Huong, 2020). Finally, item 6) *Each section is consistent with its objective and has appropriate weight of score,* is in line with the face validity sublist 1 in Hien Huong's study (Hien Huong, 2020). For reliability, item 8) *All pages of the test can be read clearly to all,* is related to the face validity sublist 5 in Hien Huong's study (Hien Huong, 2020). Item 9) *All audio* *files can be open and are loud enough for all when administering in a large testing room,* and Item 10) *Video input is equally visible to all,* covers two sublists about media in the practicality questionnaire in Kadir, Zaim, & Refnaldi's study (Kadir, Zaim, & Refnaldi, 2019). Item 13) *The test rubric is clear to all test takers to inform what they have to do and everyone has a fair chance to give correct answers,* is in correspondence with Hien Huong's face validity sublist 6 (Hien Huong, 2020). For practicality, item 30) *Students can complete the test reasonably with the test time frame,* is related to Hien Huong's face validity sublist 2 (Hien Huong, 2020), and covers two practicality sublists about time in Kadir, Zaim, & Refnaldi's study (Kadir, Zaim, & Refnaldi, 2019).

Validating an English test before it is used in administration is an important process that can help inexperienced testing practitioners know how to produce an affective English test and can also develop testing skills in their teaching professions. Test validators would be familiar with producing good English tests that contain six qualities: validity, reliability, authenticity, appropriateness, impact, and practicality. However, the proposed validation checklists might not be fit with all future uses, the users can adapt some sublists to meet their purposes and uses.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* New York: Pearson Education, Inc.

Chanthawee, C. & Rungruang, A. (2020). The use of English Regular Plural forms by Thai EFL Learners. *The Golden Teak: Humanity and Social Science Journal.* 26 (1). Retrieved from https://so05.tci-thaijo.org/index.php/tgt/article/view/240895

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Heaton, J. B. (1990). *Classroom testing.* London: Longman.

Hien Huong, N. (2020). Face validity of the institutional English test based on The Common European Framework of Reference at a Public University in Vietnam. *VNU Journal Of Foreign Studies, 36*(1). doi:10.25073/2525-2445/vnufs.4501

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kadir, J. S., Zaim, M., & Refnaldi, R. (2019). Developing instruments for evaluating validity, practicality, and effectiveness of the authentic assessment for speaking skill at Junior High School. *Proceedings of the Sixth of International Conference on English Language and Teaching (ICOELT 2018).* doi: https://doi.org/10.2991/icoelt-18.2019.14

Kelly, Melissa. (2020, February 11). Creating and Scoring Essay Tests. Retrieved from https://www.thoughtco.com/creating-scoring-essay-tests-8439

Oghabi, M., Pourdana, N, and Ghaemi, F. (2020). Developing and validating a sociocultural plagiarism questionnaire for assessing English academic writing of Iranian scholars. *Applied Research on English Language,* 9(2). doi: 10.22108/ARE.2019.118587.1485

Weir, C. J. (2005). *Language testing and validation.* New York: Palgrave Macmillan.