NATIONAL EXAMINATION OF ENGLISH IN INDONESIA: A VALIDITY AND RELIABILITY-BASED ELUCIDATION[*]

By: *Chothibul Umam*[**]
chothib_umam@yahoo.co.id

Abstract: The administration of national examination in Indonesia has caused much controversy. Some people think that the use of centralized tests is in conflict with the Law No. 20/2003 on national educational system. Some others argue that the law needs national monitoring of the students' achievement on standard competencies and controlling the quality of education through nationwide evaluation. Eventhough, those who agree with the administration of national examinations still question the validity and reliability of the tests. Validity of a test has traditionally been defined as 'the degree to which the test actually measures what is intended to measure'. Meanwhile, reliability refers to the consistency of measurement – that is, to how consistent test scores or other evaluation results are from one measurement to another. This paper tries to review the senior-high-school national examination of English on the basis of the traditional perspectives of test validity typology which includes content validity, constructs validity, and criterion-related validity as well as reviews it based on language test reliability viewpoint. Some suggestions are provided as well if the government should continue to use the national examination as a tool to assess students' performance.

Key words: national examination, validity, reliability

National Examination of English in Indonesia

The Regulation of the Ministry of National Education No. 23/2006 specifies the standard of competencies of senior-high-school students who learn English as follows: (a) the students understand oral formal and informal interpersonal and transactional discourses in the form of recount, narrative, procedure, descriptive, news item, report, analytical and hortatory exposition, spoof, explanation, discussion, and review, in terms of daily contexts, (b) the students orally express formal and informal interpersonal and transactional discourses in the form of recount, narrative, procedure,

[*]) The paper had been published at Jurnal UNIVERSUM STAIN Kediri, Vol. 5 (1) 2011 p. 1-14.
[**]) A faculty member at the department of English Language Education, Faculty of Tarbiyah, The State College of Islamic Studies (STAIN) Kediri, Indonesia.

descriptive, news item, report, analytical and hortatory exposition, spoof, explanation, discussion, and review, in terms of daily contexts, (c) the students understand written formal and informal interpersonal and transactional discourses in the form of recount, narrative, procedure, descriptive, news item, report, analytical and hortatory exposition, spoof, explanation, discussion, and review, in terms of daily contexts, (d) the students write formal and informal interpersonal and transactional discourses in the form of recount, narrative, procedure, descriptive, news item, report, analytical and hortatory exposition, spoof, explanation, discussion, and review, in terms of daily contexts. These competencies include all the four language skills (listening, speaking, reading, and writing).

The 2010/2011 national examination of English consists of 50 multiple-choice items: 15 listening comprehension items (understanding dialogues, giving responses, and understanding monologues) and 35 reading comprehension items (understanding written dialogues, advertisement, and reading passages). The examination assesses two language skills (listening and reading) only. The ministry of national education assumes that speaking and writing skills will be assessed by school teachers themselves. However, as speaking and writing skills are not represented in the examination, teachers may simply not teach the language skills (speaking and writing), and students may not learn the skills. Shohamy (2005:107) states that centralized tests are capable of dictating the teachers what to teach and what test-takers will study. Teachers focus on teaching language skills will be tested and emphasize the material that is to be included on the test. If due to some

reasons the examination could just assess listening and reading skills only, then the government should redefine the objectives of teaching English to high-school students.

Basically, learning a language aims at developing 'the four levels of literacy, namely performative, functional, informational, and epistemic levels' (Wells, 1987 in Alwasilah, 2006:109), which respectively refer to the ability to read and write, the ability to use the language in everyday communication, the ability to access knowledge, and the ability to transform knowledge. Alwasilah (2006) proposes that the four levels of literacy are taught in stages in accordance with the levels of education: the first level of literacy is taught to elementary-school pupils, the second level to junior-high-school students, the third level to senior-high-school students, and the fourth level to university students. Therefore, the objectives of teaching English to senior-high-school students can be limited to the ability to access knowledge in English.

High-school centralized tests have been administered in Indonesia since 1980. They were called EBTANAS (*Evaluasi Belajar Tahap Akhir Nasional* or National Final Evaluation of Students' Learning) from 1980 to 2001, and then UAN (*Ujian Akhir Nasional* or National Final Examination) in 2002. They have later been named UN (*Ujian Nasional* = National Examination) since 2005. The national examinations have caused much controversy. Some people think that the administration of national examinations is in conflict with the Law No. 20/2003 on national education system. Article 58 of the law states that teachers evaluate their students in

terms of the learning process, progress, and remedy. However, some others argue that articles 35 and 57 of the law respectively requires national monitoring of the students' achievement levels of standard competencies and controlling the quality of education through nationwide evaluation (see Furqon, 2004). Some educational activists recommend the government to consider the unbalance quality of schools nationwide, including poorly skilled teachers, and improper facilities in a number of regions before the government keeps pressing ahead with the nationwide examination system (*The Jakarta Post*, 26 June 2006).

However, those who agree to the government's decision on National Examination still question the validity and reliability of the national examinations. Validity of a test has traditionally been defined as 'the degree to which the test actually measures what is intended to measure' (Brown, 1996:231). Next to validity, Grondlund (1985: 93) defines reliability as 'the consistency of measurement – that is, to how consistent test scores or other evaluation results are from one measurement to another.' This paper reviews the senior-high-school national examination of English on the basis of the traditional perspectives of test validity typology which includes content validity, constructs validity, and criterion-related validity as well as reviews it based on language test reliability point of view. Some suggestions are provided if the government should continue to use the national examination as a tool to assess students' performance.

Validity-Based Analysis

Validity evidence of the assessment results can be collected from the test (the assessment instrument) being used and other related data (criterion-related validity evidence). We can collect construct and content validity evidence from the test (the assessment instrument) being used and we can collect concurrent and predictive validity evidence from criterion-related validity evidence. The main factor affecting the validity of language skill assessment result is the appropriateness of the procedures of the assessment (the appropriateness of the choices of instrument). For example, an assessment of speaking skill using a paper and pencil test that requires the examinees to show their speaking skill by writing and based on the writing the speaking skill is estimated will result in the speaking score with low validity (weak construct-validity evidence).

According to Weir (2005:14), construct validity is a function of the interaction of two aspects of validity. The first, it refers to the extent to which a test is constructed on the basis of general theories concerning the language processing which underlies the various operations required in real-life language use. It is usually called as theory-based validity. The second, it refers to the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample. It is called as context validity. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting.

A test should, therefore, always be constructed on an explicit specification which addresses both the cognitive and linguistic abilities involved in activities in the language use domain of interest, as well as the context in which these abilities are performed. There are two major threats to construct validity: construct under-representation and construct irrelevance (Messick, 1989). Test developers need to ensure the constructs elicited are precisely those intended to and that these are not contaminated by other irrelevant variables. If important constructs are under-represented in a test, this may have an adverse backwash effect on the teaching that precedes the test.

Another validity evidence of the assessment results that can be collected from the test is content validity. Content validity is important when we wish to describe how an individual performs a domain of tasks that the test is supposed to represent. A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned. It is obvious that grammar test, for instance, must be made up of items testing knowledge or control of grammar. The test would have content validity only if it included proper sample of the relevant structure. Just what are the relevant structures will depend, of course, upon the purpose of the test.

Based on the principles of construct and content validity above, we should question the validity of the English test score on the students' certificate when they have graduated from the school. English consists of four skills (listening, speaking, reading, and writing) but the examination assesses

two language skills (listening and reading) only. The 2010/2011 national examination of English consists of 50 multiple-choice items: 15 listening comprehension items and 35 reading comprehension items.

Besides, there is a question about the test development. According to SEAMEO Library (2001), test items are solicited at the district and provincial levels throughout the country. Teachers from selected schools are invited to become part of item writing teams. Each team produces 50 to75 items for one national examination. These items are then sent to Jakarta, and selected and reviewed by the National Examination Committee. This procedure tends to give the districts and provinces a sense of involvement. However, in terms of credibility and practicality, the national examinations should be developed by professional test-developers. The construct validity of a test does not lie in the sense of involvement, but in the representativeness and relevance of samples of abilities or skills being measured.

Criterion-related validity refers to the extent to which test scores correlate with a suitable external criterion of performance with established properties. The validity is the degree to which the first test is seen as related to the established criterion. By showing this relationship, one feels more confident in claiming the test as a valid measure of the same thing that was measured by the criterion test (Hatch & Farhady, 1982:251). There are two types of criterion-related validity: concurrent validity and predictive validity. Concurrent validity looks for 'a criterion which we believe is also an indicator of the ability being tested' (Bachman, 1990:248). Test scores could be correlated with another measure of performance, usually older, longer,

established test, taken at the same time or teachers' rankings of students, or even student self-assessment. Predictive validity is concerned with making certain predictions about students' future performance on the basis of test results. Predictive validity can be established by correlating language performance against later job/academic performance.

The experience of a friend as an English teacher at one senior high school gives an example of low concurrent validity evidence of the National examination result. In 2009/2010, he found awkwardness in the result of National Examination of English when he conducted a local observation in the school where he taught. He compared the result of the students' score in try-out examination and their score of National examination. The result of try-out examination of English in this school stated that only 15 percent of the students passed the exam but they passed 100% with very satisfying result when they answered the questions in the National Examination of English, whereas, the level of difficulty between the questions of try-out examination and the questions of National exam were relatively same. In conclusion, the English score of National Examination in this school has low concurrent validity evidence.

Related to predictive validity evidence of the National examination result, there has no research on the predictive validity of the national examination of English so far. Abdul-Hamied (1993) conducted a national research on English language teaching in 358 senior high schools in 26 provinces, and he found out that the results of the national examination of English were discouraging: 66.7% of the students had the scores below 6.0.

In 2006 among the provinces with the lowest percentage of students passing the national examinations were North Maluku (72.57%), East Nusa Tenggara (75.37%) and South Kalimantan (77.37%). (*The Jakarta Post*, 26 June 2006). The cut-score for passing or failing the national examinations was 4.26 in 2007 and 5.25 in 2008. However, according to the Ministry of National Education (2007:141), the senior secondary national examination scores for English have risen from 4.8-5.3 in 2004 to 6.9-8.0 in 2006.

Reliability-Based Analysis

Reliability of the result of language skill assessment refers to the preciseness of the language skill assessment result in representing the actual level of the skill of the examinees. The result of a language skill assessment has high reliability if the result precisely represents (or is very closed to, or is not too far away from, or gives good estimate of, or does not overestimate or underestimate) the true level of the skill being assessed. In other words, if the language skill assessment result is too far away different from the true level of the skill being assessed, then the assessment result has low reliability. The distance between the true level of the skill and the assessment result determines the degree of reliability. The bigger the distance is between the language skill assessment result and the actual level of the skill being assessed, the lower the reliability of that assessment result is. The distance is between the language skill assessment result and the actual level of the skill being assessed represents errors of the assessment result. The bigger the errors in the assessment result are, the bigger the distance is between the

assessment result and the actual level of the skill being assessed, and the lower the reliability of that assessment is (Lathief, 2001:217).

Some language testing experts define reliability as referring to consistency of the scores resulted from the assessment. However, consistency is not the meaning of reliability. Consistency is an important indicator for reliability, meaning that if an assessment result is (or the test scores are) consistent from one assessment to another, then the assessment result has (or the test scores have) high reliability.

Reliability concerns the extent to which test results are stable over time, consistent in terms of the content sampling and free from bias. Ebel and Frisbie (1991) emphasize some factors which affect the reliability of a test. Those are: 1) not the examinees' best performance, 2) not the raters' most objective judgment, 3) the assessment instrument being too short, 4) the assessment instrument content being too heterogeneous, 5) the assessment question being too easy or too difficult, 6) the type and the quality of assessment instrument, 7) cheating in the assessment, and 8) uncomfortable place and time.

Instability of test scores resulted from poor test administration in which there is an opportunity to 'cheat', for instance, would influence the reliability. Low reliability of assessment result means that the score resulted contain big errors and so give poor estimates for (overestimate or underestimate) the true level of the skill being assessed. If we suspect that the test is done in poor administration, for instance, and therefore the students can freely cheat each other, then we are questioning the objectivity

of the test result. Questioning the objectivity of the test result means believing that there have been errors in the assessment; some scores are believed to be too high and some too low from the actual level of the skill being assessed. Questioning the objectivity of the scores means that the reliability of the scores is low.

In test administration, proctors are an important factor. Proctors should make sure that there is no cheating in the test. However, it has been the 'public scret' that in some cases of the national examination administration, proctors who are also teachers, 'help' students by giving the answer key. Besides, at the district or regional level there is a 'succeeding team' which 'corrects' the students' answer sheets (*Koran Tempo*, 4 Februari 2005). Teachers and administrators often view the national examinations not only as testing the language performance and achievement levels of their own students but also as assessing or testing their own performances. With regard to responses scoring, in 2004 there was some controversy over the use of score conversion tables which attemptted to help slow students but was disadvantegous to bright students (*Tokoh Indonesia*, http: //www.tokohindonesia.com/ majalah/22/kilas-un.shtml). The test results derived from a poor scoring system, and the examinations were not managed by an authorized testing institution (*Pustaka Mawar*, 5 December 2007) Again, this is another form of 'cheating'.

A test is likely to have a backwash effect. Backwash is defined as the effect of a test on teachers, learners, parents, administrators, textbook writers, instruction, classroom practice, educational practices and beliefs,

and curricula. Backwash may refer to both intended positive or beneficial effects and to unintended harmful or negative effects (Bachman & Palmer, 1996). The negative effects of the national examinations are, for instance, students' committing suicide and vandalizing after the announcement of the national examination results for junior and senior high school students (*The Jakarta Post*, 25 June 2007), and teachers' and administrators' improper attempts to 'help' their students as already mentioned earlier. The government should, however, encourage beneficial effects by first improving the quality and administration of the national examinations in order to obtain reliable data of students' performance, and then on the basis of the data, taking appropriate measures to improve the quality of education in Indonesia.

Suggestions

If the government should continue to use the national examination as a tool for assessing students' performance, there is a need to do the following:

1) Let professional test-developers develop the national examination of English so that there will be no question about the construct validity of the test. The test should not be a compilation of teacher-made selected items. It could be developed by an independent educational testing institution, or in the case of English test, the TEFLIN (Teachers of English as a Foreign Language in Indonesia) organization may be asked to develop the test.

2) Manage the administration of the national examination at schools properly so that there will be no cheating from students, teachers, and administrators. Cheating affects the reliability of the test results. Proctors should be teachers from other schools, and the persons in charge of the test administration at schools should also be the principals from other schools. There should be no opportunity to let students' answer sheets 'stay for some time' at schools or regional educational offices. High-school test administrators can learn from the university entrance test administration.
3) Conduct research on the criterion-related validity of the national examination in order to convince the test users that the test is a valid measure. Use TOEFL (Test of English as a Foreign Language) or IELTS (International English Language Testing System) to find out the concurrent validity of the test.
4) Monitor the backwash effects of the national examination on students, teachers, parents, and administrators. If the national examination is considered as a high-stake test, the government should anticipate the detrimental effects of the test. The government should also use the test results to improve the quality of education by upgrading teachers and providing appropriate school facilities.

In conclusion, the use of high-stake tests, i.e. national examinations, demands the government to provide validity evidence of the instrumental value of the tests. It is the right of all test users to ask for evidence that demonstrates the tests are doing the jobs that they are supposed to be doing.

The government could convince the test users by developing the tests professionally, administering the tests properly, providing an appropriate scoring system, conducting adequate research on the correlation of students' performances on the tests with trustworthy external measures, and responsibly dealing with the backwash effects of the tests on stakeholders.

References

Abdul-Hamied, Fuad. 1993. *"Laporan Nasional Hasil Pengawasan dan Pemeriksaan Tema Pengajaran Bahasa Inggris di SMA Negeri."* Jakarta: Inspektorat Jenderal, Depertemen Pendidikan dan Kebudayaan.

Alwasilah, A. Chaedar. 2006. *Pokoknya Sunda: Interpretasi untuk Aksi*. Bandung: PT Kiblat Buku Utama.

Bachman, L. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

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

Brown, James Dean. 1996. *Testing in Language Programs*. Upper Saddle River, New Jersey: Prentice-Hall Regents.

Ebel. R.L. and Frisbie, D.A. 1991. Essentials of Educational Measurement. Englewood Cliffs, New Jersey; Prentice Hall, Inc.

Furqon. 2004. *"Masih Perlukah Ujian Akhir Nasional?"* Retrieved 8 January 2008 from http://www.pikiran-rakyat.com/cetak/1204/23/0804.htm

Gronlund, NE. 1985. *Measurement and Evaluation in Teaching.* New York: MacMillan Publishing Company.

Hatch, Evelyn and Hossein Farhady. 1982. *Research Design and Statistics for Applied Linguistics*. Rowley, Massachusetts: Newbury House Publishers, Inc.

Koran Tempo, 4 Februari 2005. *"Kontroversi Ujian Nasional."* Retrieved 8 January 2008 from http: //www.antikorupsi.org/ mod. php? mod= publisher&op= viewarticle&artid= 3764.

Latief, M. Adnan. 2001. *Reliability of Language Skills Assessment Results.* Malang: Jurnal Ilmu Pendidikan. Agustus.

Messick, S. 1989. "Validity." In R.L. Linn (Ed.), *Educational Measurement*. Third Edition. New York: Macmillan, pp. 13-103.

Ministry of National Education. 2006. *Peraturan Menteri Pendidikan Nasional No. 23 Tahun 2006 tentang standar Kompetensi Lulusan*. Jakarta: Ministry of National Education, the Republic of Indonesia

Ministry of National Education. 2007. *The 2007 EFA Mid Decade Assessment Indonesia*. Jakarta: EFA Secretariat, Ministry of National Education, the Republic of Indonesia

Pustaka Mawar. 5 December 2007. "UN Banyak Kelemahan, Sisi Metodologis Harus Diperbaiki." Retrieved 8 January 2008 from http: //pustakamawar. wordpress.com/ 2007/12/05/un-penjamin-mutu.

SEAMEO Library. 15 August 2001. "Indonesia". Retrieved 8 January 2008 from http://www. seameo.org /vl /library/ dlwelcome/ publications/ ebook/ exam/ 2ndindo.htm.

Shohamy, Elana. 2005. "The Power of Tests Over Teachers: The Power of Teachers Over Tests." In D.J. Tedick (Ed.), *Second Language Teacher Education: International Perspectives*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., pp. 101-111.

The Government of Indonesia Republic, 2003. *Undang-Undang No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional.* Jakarta: The Government of Indonesia Republic.

The Jakarta Post. 26 June 2006. "Lawmakers, activists criticize national examination system." Retrieved 8 January 2008 from http://www.kabar-irian.com/pipermail/ kabar-indonesia/2006-June/007792.html.

The Jakarta Post. 25 June 2007. "Gov told education overhaul needed." Retrieved 8 January 2008 from http://pendidikan.net/

Tokoh Indonesia. 2007. *"Kilas Balik Ujian Akhir Nasional".* Retrieved 8 January 2008 from http://www.tokohindonesia.com/majalah/22/kilas-un.shtml.

Weir, Cyril J. 2005. *Language Testing and Validation: An Evidence-Based Approach*. New York: Palgrave Macmillan.