A STUDY ON THE VALIDITY OF VSTEP WRITING TESTS FOR THE SAKE OF REGIONAL AND INTERNATIONAL INTEGRATION

Nguyen Thi Ngoc Quynh*

Center for Language Testing and Assessment, VNU University of Languages and International Studies, Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 03 April 2018

Revised 26 July 2018; Accepted 27 July 2018

Abstract: In the context of rapid regional and international integration, particularly the official establishment of the ASEAN Economic Community in 2015, English capacity has become essential for Vietnamese people to create their competitiveness in employment, education and other opportunities. In the reform of English education and assessment in response to this demand, VSTEP tests were developed and introduced by the Ministry of Education and Training as national English assessment instruments. VSTEP tests are meant to be alternative to the existing expensive international standardised English tests (e.g. IELTS, TOEFL). But this requires VSTEP developers to take action to assure test validity. They also need to accumulate and disseminate evidence of validity of the tests to gain international recognition. By doing so, they have taken meaningful action to contribute to the nation's international and regional integration. The paper highlights the commitment of ULIS-VNU as a VSTEP developing institution in this mission. It reports a recent VSTEP validation study as an example of this commitment[1].

Keywords: VSTEP, test validation, scoring validity, assessing writing, English for ASEAN integration

1. Background

Regional and international integration is nothing new to all Vietnamese citizens. The impact and evidence of this process can be seen in every corner of the country, ranging from the presence of foreigners who come to Vietnam for various purposes with their increasing number and greater access to all parts of the land, to the increasing number of Vietnamese labourers from different areas and professions in the country to work overseas such as in Japan, Korea, and Taiwan. Particularly, with the official establishment of the ASEAN Economic Community (AEC) in 2015, the connotation of regional and international integration becomes more pressing for Vietnamese government, its every sector and any ordinary person. On the one hand, the establishment of the AEC is a major milestone in the regional economic integration agenda in ASEAN, offering opportunities in the form of a huge market of US$2.6 trillion and over 622 million people[2]. In 2014, AEC was collectively the third largest economy in Asia and the seventh largest in the world. On the other hand, this means that citizens of one ASEAN nation can go and work in another. Employment now becomes more competitive not only within one nation's borders, but in the whole region. In this context, English capacity

* Tel.: 84-904322142, Email: ngquynh@gmail.com

[1] This study was completed under the sponsorship of the University of Languages and International Studies (ULIS-VNU) in the project N.16.22

[2] http://asean.org/asean-economic-community/

plays a critical role, and language is often an assumption in ASEAN documents (Dudzik & Nguyen, T.N.Q, 2015). Article 34 of the ASEAN Charter designates English as the working language of ASEAN (ASEAN, 2008). The ASEAN Socio-Cultural Community explicitly states English language capacity-building in its blueprint, along with educational investment, life-long learning, human resource training and capacity-building, and applying technology (ASEAN, 2007).

In response to growing regional and international demand for foreign languages, the government of Vietnam issued a decision to "thoroughly renovate the tasks of teaching and learning foreign languages within the national education system" in order to produce graduates who "gain the capacity to use a foreign language independently" (Government Decision 1400 I.1, 2008, p. 1). Decision 1400, entitled *Teaching and Learning Foreign Languages in the National Educational System*, Period 2008-2020, gave birth to the National Foreign Language 2020 Project (NFLP 2020). Major goals of this project are to reform the teaching, learning and assessment of foreign languages, especially English in the education sector.

To date, two of the most significant achievements of the NFLP 2020 have been the development of the Vietnam's Framework of Foreign Language Competency aligned with the Common European Framework of Reference (CEFR-VN) and the locally-produced standardised English proficiency tests, so-called VSTEP. As for the former, instead of the six levels A1, A2, B1, B2, C1 and C2 as described in the CEFR introduced by the European Union, the CEFR-VN consists of levels 1 to 6 with similar descriptors of competences to the CEFR, but with adaptation to match the features of English context and use in Vietnam. The latter, Vietnamese Standardized Test of English Proficiency (VSTEP), is a test of general English proficiency developed based on the Common European Framework of Reference. Two VSTEP test formats, one measuring levels 3-5 and the other measuring level 2 according to the CEFR-VN, have recently been issued by Vietnam's Ministry of Education and Training as national test instruments for English assessment. The test consists of sections assessing reading, writing, speaking, and listening, with all four sections taken by all test takers. In fact, VSTEPs are the first ever locally-produced standardised English proficiency tests in Vietnam.

One major goal, among several others, of these initiatives is to assure fairness in English assessment in Vietnam, both because they are made suitable for Vietnamese learners and the context of English education and use in Vietnam, and because they are of lower cost and thus more accessible for the majority of English learners in the country (Nguyen. T.N.Q. & Do. T.M., 2015). At least the latter is evident. Since the arrival of the VSTEP, a great number of Vietnamese people, not limited to the education sector, have been assessed on their English proficiency against the CEFR-VN, aligned to the CEFR. Let alone at the University of Languages and International Studies, Vietnam National University, Hanoi (ULIS-VNU), about 8,000 people took the VSTEP test in the year 2016.

However, a big challenge in the development of VSTEP tests is to assure their quality so that their test scores are valid and meaningful indicators of Vietnamese learners' English ability levels as compared to international standards. That is, a level-3 learners according to the CEFR-VN should be equally proficient in English to those

identified at B1 level based on the CEFR, level-4 to B2 level, and level-5 to C1 level. It is highly important that VSTEP test developers in Vietnam take necessary quality assurance measures because such credibility and validity of their test scores are essential so that the international and regional public, such as employers who want to recruit Vietnamese labour, could trust the levels of English capacity of their Vietnamese counterparts reported by these test results.

In the following sections, the paper discusses the aspects of test validity, and the research taken to examine the validity of the VSTEP.3-5 tests at the ULIS-VNU as a commitment to gain international recognition of its VSTEP test scores.

2. Test validity

Validity has been regarded as 'indisputably necessary for any serious test' (Hughes, Porter & Weir, 1988: 4). In order to tell whether a test is 'good' or not, one often examines whether it is valid or not. Despite its universally agreed importance in testing and assessment, validity has been perceived as of a range of various concepts, not necessarily the same among test developers and researchers. Lissitz (2009) provided a collection of different perspectives of the concept of validity in language testing and assessment.

The most general and classic concept of test validity is the degree in which a test is truly measuring what it is intended to measure (Kelly, 1927; Lado, 1961; Cronbach, 1971; Henning, 1987; Davies, 1990; Hatch & Lazaraton, 1997). This view focuses primarily on the test itself, and such concepts as content validity and construct validity are of central attention. This unitary view of validity has attracted a lot of critiques and been modified by other theorists.

Messick's (1989) unified view of validity defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p.13). This view sees construct validity as the superordinate category for different test validities.

However, according to Weir (2005), "validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e., test scores, which are an accurate representation of a candidate's level of language knowledge and skills" (p.12). In such a view, validity is not an attribute of the test itself, but rather in the scores on a particular administration. Over time, if various versions of a test or administrations of the same test provide similar results, then synthetically a case may be made for that test being valid over time and across versions and population samples (Weir, 2005:13).

Modern theorists now seem to have reached a consensus that "validity is multifaceted, and different types of evidence are needed to support any claims for the validatity of scores of a test" (Weir, 2005: 13). None of the evidences by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores (Bachman, 1990: 237).

In addition, it is also widely agreed that validity is a matter of degree, not all or none (Messick, 1989; Weir, 2005). This should be viewed as a relative concept. For example, in terms of content coverage against test specification, one test may present some other aspects of content described in the test specification from what another test does. Different tests may also differ in their claims to validity and across different types of validity

(Messick, 1989, 1998; Weir, 2005; Shaw & Weir, 2007). For example, version X of a writing test may be strong in theory-related validity and cover more content specified in the test specification, but its rater reliability is weak; while version Y covers less content, but is rated with higher rater reliability coefficient.

However, now comes the question of what types of evidences and for what types of validity test validators should collect in order to claim the validity of a test. In this regards, Weir (2005) presents a socio-cognitive framework for test validation. He distinguishes theory-based validity, context validity, scoring validity, and the two external validities, criterion-related validity and consequential validity of a test. He also suggests the types of *a priori* (i.e. taken before the test administration) and *a posteriori* (i.e. taken after the test event) validity evidence.

Weir's (2005) socio-cognitive framework has been further developed by himself and his colleagues to provide validation porfolios specific to tests of different macro skills: listening, reading, speaking, and writing. These models are very practical and useful for test developers.

3. VSTEP validity evidence for international recognition: an emphasized mission

The first section of this paper stresses the demand for English capacity of Vietnamese people in the new era of international and regional integration. For various types of purposes such as overseas employment and education, Vietnamese are to be able to provide the information and evidence of their English ability to gain their competitiveness. As discussed in the first section, instead of taking expensive English proficiency tests provided by international language testing organisations such as IELTS, TOEFL or Cambridge main suite tests (i.e. PET, FCE, CAE), Vietnamese people can now choose to take VSTEP, which is of lower cost and tailored to English use and English teaching and learning context in Vietnam, to gain certification of their English proficiency (Nguyen. T.N.Q. & Do. T.M., 2015). Thus, it is justifiable to say that the development of VSTEP tests has both technical, practical, and humanitary meanings. These are new EFL tests that are made in Vietnamese context, by Vietnamese experts and for Vietnamese English users.

However, as stated above, a challenge for VSTEP developers is to assure its validity. And what's more important is that they should be able to provide sufficient evidence of its validity. According to Messick (1992), it is a responsibility of test developers to provide evidence of validity of their test. However, the fact is that not many of test developers in the world have done so. They may have thousands of different reasons for not doing so, including constraints in terms of technical and financial resources for validation work, or simply their lack of commitment in providing transparent information of their tests. However, as for Vietnamese test developers of VSTEP, this is identified as an emphasized mission. Within the national borders, the use and recognition of VSTEPs are assured by the legal documents by Vietnamese governmental and education sectoral agencies. However, when it comes to regional and international contexts, sufficient evidence of VSTEP validity is essential to the recognition of international and regional counterparts. Foreign stakeholders, i.e. foreign employers, education institutions, researchers, etc., have their own right and choice whether to trust VSTEP test scores and their equivalent levels of English proficiency reported based on these scores or not. For their recognition consideration and decision, they deserve the right to know

about how the tests are developed, delivered, and scored, and whether these tests are up to international quality standards and can be used alternatively to the already widely-known and recognised international tests (e.g. IELTS, TOEFL).

Fully understanding the above responsibility, as one of the biggest VSTEP test developer nationwide, ULIS-VNU is committed to providing transparent and credible information of VSTEP validity, as a concrete action to gain regional and international recognition of their test results with an ultimate goal to contribute to the national reform of English education and assessment led by the NFLP 2020, and the integration of Vietnam to the world. A series of validation studies have been being conducted by ULIS researchers and in collaboration with international experts. The following section is a report on a recent study on the scoring validity of the VSTEP Writing Test as an example of ULIS-VNU effort and commitment in carrying out this mission.

4. Example: A study on scoring validity of VSTEP Writing Test[3]

4.1. Overview of the study

The aim of the reported study was to examine the consistency of scoring in the VSTEP Writing Test. This test consists of two tasks: Task 1 asks test takers to write a correspondence of at least 120 words for an intended purpose that is described in the task; Task 2 asks test takers to write an essay of at least 250 words about a given topic. The test is 60 minutes long, with the recommended time allowance for Task 1 of 20 minutes and Task 2 40 minutes.

The study was to answer the two following research questions:

1. How consistent are the ratings of the VSTEP writing test?

2. How do various facets of the rating process contribute to score variation in writing?

4.2. Methodology

4.2.1. Data collection

The study examined three forms of the VSTEP writing tests administered in late 2015 (hereinafter called as Test A) (with 546 participants), mid 2016 (1212 participants) (Test B), and late 2016 (476 participants) (Test C).

Each task of the test was rated separately with a four-subscale analytic rubric: task fulfillment, organization, vocabulary and grammar. Raters provided scores for every subscale for each task out of 10. The score for each task was the average of four subscales and calculated out of 10. An exception was with Test A form, in which Rater 1 provided analytic scores, but Rater 2 only provided holistic scores of each task. The composite score of the Writing test was calculated as follows: (Task 1 + Task 2 x 2)/3. This final score was rounded to 0.5.

As a measure to improve scoring validity, at ULIS-VNU, every student paper is double-rated by two raters at different times. If any ratings are disparate by 1 point or more (out of 10), the paper is marked by a third, lead rater. In addition, 15% of the papers are remarked by third rater for quality assurance. This rating procedure was applied at all the three studied test forms.

[3] This study examines VSTEP.3-5 tests under the sponsorship of the University of Languages and International Studies (ULIS-VNU) in the project No. N.16.22. This was jointly conducted with a language testing specialist sponsored by the Regional English Language Office, US Embassy in Hanoi.

4.2.2. Data analysis

A review of literature shows that in estimating the scoring consistency of rated tests, there are a number of methodological options, none of which presents entirely satisfactory results by itself. Inter-rater correlations, for example (see, e.g., Bachman, 2004; Carr, 2010), while perhaps the simplest and most commonly used approach, tell nothing about the effects of other aspects of the testing process, such as differences in task difficulty or test takers' language ability. Generalizability theory (see Brennan, 2001; Shavelson & Webb, 1991), in contrast, tells us how such aspects of the testing process contribute to score variation, and to dependability, but yields no information on the ability of individual test takers, the severity or leniency of individual raters, or the difficulty of specific tasks. Finally, the many-facet Rasch model (see, e.g., Bond & Fox, 2001; Linacre, 2014; McNamara, 1996) does provide information at the individual level, but without the information at the facet level and the clearly interpretable estimates of overall consistency provided by generalizability theory.

Therefore, this study adopts the triangulation approach employed in other studies (e.g., Bachman, Lynch, & Mason, 1995; Lynch & McNamara, 1998) of combining many-facet Rasch measurement with generalizability theory, while adding consideration of inter-rater score correlations as an additional source of information on scoring consistency.

Specifically, for each of the three test forms, the following statistical analyses were conducted: (1) the descriptive statistics for every subscale of the writing scores through SPSS; (2) Cronbach's alpha and correlations also through SPSS; and (3) generalizability theory (mGENOVA), many-facet Rasch (FACETS).

4.3. Results

4.3.1. Descriptive statistics

Tables 1-5 show the descriptive statistics for every subscale and the composite of the writing scores of the three test forms.

Table 1. Descriptive statistics – Task fulfilment

	Test A (n=546)	Test B (n=1212)	Test C (n=476)
Mean	6.2	5.1	4.7
Median	6.5	5	4.5
Mode	7	4	--[b]
SD	1.6	1.5	1.3
Skewness	-0.6*	0.3*	-0.1
SES	0.1	0.1	0.1
Kurtosis	0.2	-0.2	-0.1
SEK	0.2	0.1	0.2

Table 2. Descriptive statistics – Organisation

	Test A (n=546)	Test B (n=1212)	Test C (n=476)
Mean	6	4.8	4.5
Median	6	4.5	4.5
Mode	7	4	4
SD	1.7	1.5	1.2
Skewness	-0.5*	0.5*	0.1*

SES	0.1	0.1	0.1
Kurtosis	-0.2	-0.2	0.1
SEK	0.2	0.1	0.2

Table 3. Descriptive statistics – Vocabulary

	Test A (n=546)	Test B (n=1212)	Test C (n=476)
Mean	5.8	4.6	4.3
Median	6	4.5	4
Mode	7	3	4
SD	1.6	1.5	1.2
Skewness	-0.6*	0.5*	0.1
SES	0.1	0.1	0.1
Kurtosis	-0.1	-0.3*	0
SEK	0.2	0.1	0.2

Table 4. Descriptive statistics – Grammar

	Test A (n=546)	Test B (n=1212)	Test C (n=476)
Mean	5.7	4.5	4.2
Median	6	4.3	4
Mode	7	3	4
SD	1.7	1.5	1.2
Skewness	-0.6*	0.5*	0.2
SES	0.1	0.1	0.1
Kurtosis	-0.5*	-0.4*	0.1
SEK	0.2	0.1	0.2

Table 5. Descriptive statistics - Composite

	Test A (n=546)	Test B (n=1212)	Test C (n=476)
Mean	5.9	4.7	4.4
Median	6.1	4.5	4.3
Mode	6.8	--b	3.5
SD	1.6	1.5	1.2
Skewness	-0.6*	0.5*	0
SES	0.1	0.1	0.1
Kurtosis	-0.1	-0.3*	0.1
SEK	0.2	0.1	0.2

It can be seen from Tables 1-5 that the means of the test scores were very similar across all the subscales and to those of the composite scores of all the three test forms.

4.3.2. Traditional reliability analyses

The Cronbach's alpha and inter-rater correlations (by subscale and for composite) were calculated for the three

test administrations. However, for the Test A , only Rater 1 provided analytic scores while Rater 2 provided holistic scores. Therefore, the Cronbach's alpha was only calculated for rater 1 only, while the inter-rater correlation was not done for this test administration. Results can be seen from Tables 6 and 7 below.

Table 6. Cronbach's alpha of the three test administrations

Test administration	Cronbach's alpha
Test A (rater 1 only)	0.973
Test B	0.988
Test C	0.974

The Cronbach's alpha for all three test administrations were all high and close to 1.0. These mean that in all three administrations, the internal consistency of the ratings across subscales and with the composite scores was very strong.

Table 7. Inter-rater correlations

Score	Test B (n=1212)	Test C (n=476)
Task Fulfilment	.889**	.906**
Organisation	.905**	.912**
Vocabulary	.900**	.922**
Grammar	.903**	.903**
Composite	.942**	.946**

It can be seen that all the calculated inter-rater correlations for Test B and Test C were strong and significant. These show that the raters scored very consistently across all four subscales (i.e. marking criteria) and with the composite in both VSTEP administrations.

4.3.3. Generalizability theory

Again, due to different rating systems between Rater 1 and Rater 2 for Test A forms, two different G-theory analyses were done for the scores by these raters using the $p^{l}x\ t^{l}$ model, which means all test-takers take all tasks (items) and all tasks are scored in all of the rating categories (subscales). In addition, variance components results for holistic (task means for Rater 1 and holistic scores for Rater 2) writing scores were also calculated using the *p x r x t* model.

Table 8.1. Variance components results for Rater 1 analytic writing scores using the $p^{l}x\ t^{l}$ model for Test A

Source of variance	Task fulfillment		Organization		Vocabulary		Grammar	
p	2.00601	70.10%	2.52119	76.30%	2.28469	79.20%	2.76229	82.00%
t	0.00315	0.10%	0.00524	0.20%	0.00737	0.30%	0.00599	0.20%
pt, e	0.85308	29.80%	0.7759	23.50%	0.59245	20.50%	0.59932	17.80%
Total	2.86224	100.00%	3.30233	100.00%	2.88451	100.00%	3.3676	100.00%

As seen above, the major source of variance in Rater 1's analytic scores of Test A across all four subscale was test-taker variability with the percentages for *p* were all above 70%.

Table 8.2. Variance components results for Rater 2 writing scores: $p^{l}x\ t^{l}$ model

Source of variance	Composite	
p	2.26824	80.40%
t	0.0012	0.00%
pt, e	0.55167	19.60%
Total	2.82111	100.00%

Table 8.2 shows that Rater 2's holistic scores of Test A writings also varied mostly due to the differentiation among test takers (80.40%).

Table 8.3. Variance components results for holistic (task means for Rater 1 and holistic scores for Rater 2) writing scores: p x r x t model

Source of variance	Composite	
p	2.20644	94.40%
r	-0.00017	0.00%
t	0.00148	0.10%
prt, e	0.13079	5.60%
Total	2.33871	100.10%

As seen from Table 8.3, test-takers' variance accounted for 94.40% the cause of the variation between Rater 1's task mean scores and Rater 2's holistic scores in Test A.

For further examination of the Test A, index of dependability (Φ) results were calculated, using $p^{l}x\ t^{l}$ model for Rater 1 scores, *p x t* model for Rater 2 holistic scores, and *p x r x t* model for holistic (task means for Rater 1 and holistic scores for Rater 2) writing scores.

Table 9.1. Index of dependability (Φ) results for Rater 1 writing scores from the Test A administration: $p^{l}x\ t^{l}$ model

Subscale	Φ
Task fulfillment	0.824
Organization	0.866
Vocabulary	0.884
Grammar	0.901
Composite	0.893

Table 9.2. Index of dependability (Φ) results for Rater 2 writing scores: *p x t* model

Scale	Φ
Rater 2 (holistic)	0.891

Table 9.3. Index of dependability (Φ) results for holistic (task means for Rater 1 and holistic scores for Rater 2) writing scores: p x r x t model

Scale	Φ
Holistic scores	0.876

Tables 9.1 - 9.3 all show that both Rater 1's scores (across subscale and composite) and Rater 2's were all highly dependable (all above .80).

As for the remaining two administrations Test B and Test C, $p^{l}x\ r^{l}\ x\ t^{l}$ model was used to calculate the variance components and index of dependability.

Table 10.1. Variance components results for Test B: $p^{l}x\ r^{l}\ x\ t^{l}$ model

Source of variance	Task fulfillment		Organization		Vocabulary		Grammar	
p	1.86524	75.20%	2.08608	76.80%	2.08366	78.40%	2.13199	79.20%
r	0.00385	0.20%	-0.00003	0.00%	-0.00004	0.00%	0.00081	0.00%
t	-0.00011	0.00%	0.00755	0.30%	0.00628	0.20%	0.00081	0.00%
pr	0.08443	3.40%	0.06851	2.50%	0.07966	3.00%	0.08541	3.20%
pt	0.28353	11.40%	0.30433	11.20%	0.27013	10.20%	0.23847	8.90%
rt	0.00209	0.10%	0.00108	0.00%	-0.00012	0.00%	0.0005	0.00%
prt, e	0.2409	9.70%	0.24995	9.20%	0.2165	8.20%	0.23465	8.70%
Total	2.48004	100.00%	2.7175	100.00%	2.65623	100.00%	2.69264	100.00%

Table 10.1 shows that test-takers' variability accounted for more than 75% the cause of Test B's score variation for all four subscales.

Table 10.2. Dependability and reliability results for writing scores for Test B administration: $p^l x r^l x t^l$ model

Subscale	Φ	Eρ²
Task fulfillment	0.883	0.884
Organization	0.892	0.893
Vocabulary	0.9	0.901
Grammar	0.906	0.906
Composite	0.914	0.914

The above table shows that raters' writing scores for Test B were highly dependable and reliable for all four subscales and the composite, with the dependability and reliability results all above .80.

Table 11.1. Variance components results for Test C: $p^l x r^l x t^l$ model

Source of variance	Task fulfillment		Organization		Vocabulary		Grammar	
p	1.15549	48.40%	1.1352	54.00%	1.08188	55.70%	1.06374	55.00%
r	0.00139	0.10%	0.00004	0.00%	-0.00028	0.00%	-0.00054	0.00%
t	0.17594	7.40%	0.07343	3.50%	0.07666	3.90%	0.08241	4.30%
Pr	0.0317	1.30%	0.0362	1.70%	0.01603	0.80%	0.03205	1.70%
Pt	0.7978	33.40%	0.67132	31.90%	0.59088	30.40%	0.57463	29.70%
Rt	-0.00035	0.00%	-0.00024	0.00%	0.00024	0.00%	0.00065	0.00%
prt, e	0.22633	9.50%	0.18577	8.80%	0.17742	9.10%	0.18016	9.30%
Total	2.38865	100.10%	2.10196	99.90%	1.94311	99.90%	1.93364	100.00%

It can be seen from Table 11.1 that for Test C administration, person (i.e. test-taker variability) was the most important source of score variation across all four subscales, but the percentages were around 50%, lower than Test A and Test B. However, the second major source is the interaction between test-takers and the test tasks (person x task) (accounting for about 30% of the variance for all four subscales), with the remaining sources being all small.

Table 11.2 shows that raters' writing scores of Test C were quite dependable and reliable with the results being all at around .70.

Table 11.2. Dependability and reliability results for writing scores for the Test C administration: $p^l x r^l x t^l$ model

Subscale	Φ	Eρ²
Task fulfillment	0.674	0.71
Organization	0.722	0.74
Vocabulary	0.737	0.757
Grammar	0.732	0.753
Composite	0.736	0.761

4.3.4. Rasch analyses

Rasch analyses were conducted to investigate the inter-relation between task type, test date and subscale. Tables 12.1, 12.2 and 12.3 show the results.

Table 12.1. Measurement report for task type (n=2328)

Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	TaskType
-.01	.01	.98	-2.2	.97	-2.3	1.03	.47	.47	1	CORRESP
.01	.01	1.02	2.1	1.02	1.9	.97	.47	.47	2	ESSAY
.00	.01	1.00	.0	1.00	-.2		.47		Mean (Count: 2)	
.01	.00	.02	2.2	.02	2.1		.00		S.D. (Population)	
.02	.00	.03	3.1	.03	3.0		.00		S.D. (Sample)	

Table 12.2. Measurement report for test date (n=2328)

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	TestDate
24595	4064	6.05	5.28	-.12	.01	.97	-1.6	.99	-.6	1.00	.57	.56	1	Test A
99106	20910	4.74	4.89	.05	.00	1.03	2.9	1.02	2.3	.99	.38	.38	2	Test B
33495	7612	4.40	4.85	.07	.01	.94	-3.8	.94	-3.8	1.04	.18	.21	3	Test C
52398.7	10862.0	5.06	5.01	.00	.01	.98	-.9	.98	-.8		.37		Mean (Count: 3)	
33226.3	7251.2	.71	.19	.08	.00	.04	2.8	.03	2.5		.16		S.D. (Population)	
40693.8	8880.8	.87	.24	.10	.00	.05	3.4	.04	3.1		.20		S.D. (Sample)	

Table 12.3. Measurement report for subscale (n=2328)

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Subscale
41905	8147	5.14	5.35	-.15	.01	.96	-2.8	.96	-2.4	1.02	.45	.44	1	TF
39573	8147	4.86	5.04	-.02	.01	1.03	2.1	1.03	1.9	.96	.45	.45	2	Org
38219	8146	4.69	4.87	.06	.01	.99	-.8	.98	-1.2	1.02	.46	.46	3	Vocab
37499	8146	4.60	4.77	.10	.01	1.02	1.4	1.02	1.0	.99	.46	.46	4	Gramm
39299.0	8146.5	4.82	5.01	.00	.01	1.00	.0	1.00	-.2		.46		Mean (Count: 4)	
1678.7	.5	.21	.22	.10	.00	.03	2.0	.03	1.8		.01		S.D. (Population)	
1938.4	.6	.24	.25	.11	.00	.03	2.3	.03	2.0		.01		S.D. (Sample)	

It can be seen that the measurement results were all close to zero in all three above Rasch analyses.

4.4. Discussion

4.4.1. Research question 1: How consistent are the ratings of the VSTEP writing test?

With the exception of Test C, writing scores exhibited very high dependability levels. As for the Test C test form, the writing scores were less dependable than the desirable rate of .80 (.736), and a bit lower than the other two administrations, yet still acceptable.

Task 2 (essay) scores generally tended to be close to Task 1 (correspondence) scores, but for Test C, they were markedly lower (.5 points for mean composite, .6 for median composite). For this administration only, the task facet contributed a noticeable proportion of total score variance (7.4%); person-task interaction was 33.4%.

This shows that there may have been something confusing about the prompt (which was expected to be a little on the easy side). However, the inter-rater correlations for writing scores were all very high, for both individual subscales (≥ .889) and composite scores (≥ .937).

4.4.2. Research question 2: How do various facets of the rating process contribute to score variation in writing?

Five facets were considered: person (i.e. test taker), rater, task type, test date, and subscale.

Person: It was clear from the results that person (i.e. test-taker) majorly contributed to score variation in all three test administrations. The person-task interaction effect made a noticeable contribution to total score variance in the three administrations for which it could be estimated (29.8%, 11.4%, 33.4%).

Rater: The rater facet contributed very little to total score variance. In addition, rater-person and rater-task effects were also minimal.

Task type: The task facet contributed very little to total score variance, aside from tasks in Test C ($\sigma 2t = 7.4\%$). In the Rasch analyses, the correspondence and essay tasks were both very close to 0 in difficulty. Therefore, based on both G theory and Rasch results, task type does not seem to affect scores in any important way, although individual prompts may do so.

Test date: In the Rasch analyses, test date difficulty ranged from -.12 to .07, very close to 0, so it was not an important contributor to scores.

Subscale: Subscales all had difficulty measures from -.15 to .10, so this facet was also not important to total scores.

Clearly, overall the VSTEP Writing Test is demonstrating sufficient score dependability for high-stakes decisions. The dependability levels of the writing test also generally display high dependability, sufficient to support important decisions.

5. Conclusion

The above report of a validation study of VSTEP tests is an example of how VSTEP developers at ULIS-VNU are endeavoring to contribute to the process of regional and international integration of Vietnam. As elaborated earlier, it is critical that these made-in-Vietnam English proficiency tests be proven to be valid and reliable so as to be recognized not only by Vietnamese people, but also by stakeholders in the broader world. Such recognition would pave the way for Vietnamese learners of English to use their VSTEP test results (instead of expensive international tests) as proof of their English capacity, either for educational, employment or any other relevant purposes in international contexts. It is believed that the accumulation of such empirical evidence will help to convince the world of the quality and seriousness of VSTEP tests. These are also concrete and meaningful contributions of ULIS-VNU and its VSTEP development team to the nation's will for international and regional integration.

References

ASEAN (2007). *ASEAN Socio-Cultural Community Blueprint*. Jakarta, Indonesia: ASEAN Secretariat. Retrieved September 1, 2017 from http://www.asean.org/communities/asean-socio-cultural-community.

ASEAN (2008). *ASEAN Economic Community Blueprint*. Jakarta, Indonesia: ASEAN Secretariat. Retrieved September 1, 2017 from http://www.asean.org/archive/5187-10.pdf.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 239-257.

Bonk, W. J. & Ockey, G. J. (2003). A Many-Facet Rasch Analysis of the Second Language Group Oral Discussion Task. *Language Testing, 20*(1), 89-110.

Dudzik, L. Diana. & Nguyen Thi Ngoc Quynh (2015). Vietnam: Building English competency in preparation for ASEAN 2015. In Stroup, R. & Kimura, K. (Eds.), *ASEAN integration and the role of English language teaching.* Phnompenh: IDP Education (Cambodia) Ltd.

Koizumi, R., In'nami, Y., & Fukazawa, M. (2016). *Multifaceted Rasch Analysis of Paired Oral Tasks for Japanese Learners of English*. In Q, Zhang (Eds), Pacific Rim Objective Measurement

Symposium (PROMS) 2015 Conference Proceedings. Springer, Singapore.

Lissitz, R. W. (2009) (ed.) *The concept of validity: revisions, new directions, and applications.* Charlotte, NC: Information Age Publishing Inc.

Lynch, B. K. & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

Messick, Samuel (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, Samuel (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1-3), 35–44.

MOET (2008). *Government Decision 1400: Teaching and Learning Foreign Languages in the National Educational System, Period 2008-2020.* (English Translation). Hanoi, Vietnam.

MOET (2014). *Circular 01 on the issuance of Vietnam's Foreign Language Competency Framework*. Retrieved September 1, 2017 from https://thuvienphapluat.vn/van-ban/Giao-duc/Thong-tu-01-2014-TT-BGDDT-Khung-nang-luc-ngoai-ngu-6-bac-Viet-Nam-220349.aspx

MOET (2015). *Decision 729 on the issuance of Vietnamese Stardardised Test of English Proficiency levels 3-5*. Retrieved September 1, 2017 from https://thuvienphapluat.vn/van-ban/Giao-duc/Quyet-dinh-729-QD-BGDDT-2015-de-thi-danh-gia-nang-luc-su-dung-tieng-Anh-tu-bac-3-den-bac-5-267956.aspx

Nguyen Thi Ngoc Quynh & Do Tuan Minh (2015). Developing a made-in-Vietnam standardized test of English proficiency for adults – The status quo and future development. Plenary paper presented at the TESOL Symposium on *English language innovation, implementation & sustainability*, Da Nang, Vietnam.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke: Palgrave Macmillan.

NGHIÊN CỨU TÍNH GIÁ TRỊ CỦA BÀI THI VIẾT VSTEP NHẰM ĐÓNG GÓP CHO QUÁ TRÌNH HỘI NHẬP KHU VỰC VÀ QUỐC TẾ

Nguyễn Thị Ngọc Quỳnh

Trung tâm Khảo thí, Trường Đại học Ngoại ngữ, ĐHQGHN,
Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Tóm tắt: Trong bối cảnh hội nhập khu vực và quốc tế mạnh mẽ, đặc biệt là việc thành lập chính thức của Cộng đồng Kinh tế ASEAN năm 2015, năng lực tiếng Anh đã trở nên thiết yếu đối với người Việt Nam để tạo ra tính cạnh tranh trong tuyển dụng, giáo dục và những cơ hội khác. Để đáp ứng nhu cầu này, trong công cuộc đổi mới dạy học và kiểm tra đánh giá tiếng Anh, các định dạng đề thi VSTEP đã được xây dựng và ban hành bởi Bộ Giáo dục và Đào tạo, được xem như là các công cụ kiểm tra đánh giá tiếng Anh quốc gia dành cho người Việt nhằm thay thế cho các đề thi đánh giá năng lực tiếng Anh chuẩn quốc tế rất đắt đỏ hiện nay như IELTS, TOEFL... Tuy nhiên, điều này đòi hỏi những đơn vị xây dựng đề thi VSTEP phải có những biện pháp cụ thể để đảm bảo tính giá trị của đề thi. Họ cũng cần phải thu thập và quảng bá những minh chứng về tính giá trị của các đề thi này để đạt được sự công nhận của quốc tế. Đây là nhiệm vụ có ý nghĩa quan trọng, góp phần vào công cuộc hội nhập khu vực và quốc tế của Việt Nam. Thông qua một nghiên cứu xác trị đề thi VSTEP được thực hiện gần đây, bài viết nhấn mạnh cam kết của Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội với tư cách là đơn vị phát triển VSTEP trong nhiệm vụ này.

Từ khoá: VSTEP, xác trị đề thi, tính giá trị của quá trình chấm thi, đánh giá kỹ năng viết, năng lực tiếng Anh cho hội nhập ASEAN