AN INVESTIGATION INTO THE CONTENT VALIDITY OF A VIETNAMESE STANDARDIZED TEST OF ENGLISH PROFICIENCY (VSTEP.3-5) READING TEST

Nguyen Thi Phuong Thao*

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

Received 07 March 2018

Revised 26 July 2018; Accepted 31 July 2018

Abstract: This paper investigated the content validity of a Vietnamese Standardized Test of English Proficiency (VSTEP.3-5) Reading test via both qualitative and quantitative methods[1]. The aim of the study is to evaluate the relevance and the coverage of the content in this test compared with the description in the test specification and the actual performance of examinees. With the content analysis provided by three testing experts using Bachman and Palmer's 1996 framework and test score analysis, the study results in a relatively high consistency of the test content with the test design framework and the test takers' performance. These findings help confirm the content validity of the specific investigated test paper. However, a need for content review is raised from the research as some problems have been revealed from the analysis.

Keywords: language testing, content validity, reading comprehension test, standardized test

1. Introduction

In foreign language testing, it is crucial to ensure the test validity – one of the six significant qualities (along with reliability, authenticity, practicality, interactiveness and impact) for test usefulness (Bachman & Palmer, 1996). Accordingly, designing a valid reading test is of great concern of language educators and researchers (Bachman and Palmer, 1996; Alderson, 2000; Jin Yan, 2002).

The Vietnamese Standardized Test of English Proficiency (VSTEP.3-5) has been implemented for Vietnamese learners of English since March 2015. The test aims at assessing English proficiency from level 3 to level 5 according to the Common European Framework of Reference for languages for Vietnamese learners (CEFR-VN) or from level B1 to level C1 according to the Common European Framework of Reference for Languages (CEFR) for users in various majors and professions using four skills. There have not been many studies on this test since only two articles were published regarding the rater consistency in rating L2 learners' writing task by Nguyen Thi Quynh Yen (2016) and the washback effect of the test on the graduation standard of English-major students at University of Languages and International Studies (ULIS), Vietnam National University (VNU) by Nguyen Thuy Lan (2017). The test analysis has been so far under-researched.

* Tel: 84-963716969

Email: phuongthaonguyen310@gmail.com

1 This study was completed under the sponsorship of the University of Languages and International Studies (ULIS-VNU) in the project N.16.23

Like the other skills, the reading tests have been developed, designed and expected to be valid in its use. It is of importance that the test measures what it is supposed to measure (Henning, 2001: 91). In this sense, validity "refers to the interpretations or actions that are made on the basis of test scores" and "must be evaluated with respect to the purpose of the test and how the test is used" (Sireci, 2009). In the scope of this study, the author would like to evaluate the content validity of a specific VSTEP.3-5 reading test with a focus on the content of the test and the test scores. The results of this study, to an extent, are expected to respond to concerns about the quality of the test to the public.

2. Literature review

2.1. Models of validity

As it is claimed by researchers, validity is the most important quality of test interpretation or test use (Bachman, 1990). The inferences or decisions we make based on the test scores will guarantee the test's meaningfulness, appropriateness and usefulness (American Psychological Association, 1985). In examining such qualities related to the validity of a test, test scores play the key role but are not the only factor as it needs to come together with the teaching syllabus, the test specification and other factors. As a result, the concept of validity has been seen from different perspectives, which leads to the fact that there are different viewpoints to categorize this most crucial quality of a test. Due to the purpose and the scope of this paper, the researcher will present two main types of validity, and how content validity can be examined.

Content validity

As test users, we have a tendency to examine the test content, which can be seen from the copy of the test and/or test design guidelines. In other words, test specifications and example items are to be investigated. Likewise, when designing a test, test developers also pay their attention to the content or ability domain covered in the test from which test tasks/items are generated. Therefore, consideration of the test content plays an important role to both test users and test developers. "Demonstrating that a test is relevant to and covers a given area of content or ability is therefore a necessary part of validation" (Bachman, 1990:244). In this sense, content validity is concerned with whether or not the content of the test is "sufficiently representative and comprehensive for the test to be a valid measure of what it is supposed to measure" (Henning, 2001:91).

As regards the evidential basis of content validity, Bachman (1990) discussed the two following aspects: content relevance and content coverage. *Content relevance* requires "the specification of the behavioral domain in question and the attendant specification of the task or test domain." (Messick, 1980:1017). According to Bachman (1990), content relevance should be considered in the specification of the ability domain – or the constructs to be tested, and the test method facets – aspects of the whole testing procedure. This is directly linked with the test design process to see whether the items generated for the test can reflect the constructs to be measured and the nature of the responses that the test taker is expected to make. The second aspect of content validity is named *content coverage* or "the extent to which the tasks required in the test adequately represent the behavioral domain in question" (Bachman, 1990:245). Regarding test validation, this is the basis to evaluate how much the test items represent the domain(s); in other words, how much they match the specification.

The limitation of content validity is that its does not take into account the actual performance of test takers (Cronbach, 1971; Bachman, 1990). It is an essential part of the validation process, but it is sufficient all by itself as inferences about examinees' abilities cannot be made from it.

Construct validity

According to Bachman (1990:254), construct validity "concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs." This is related to the way test scores are interpreted and how this interpretation can reflect the abilities the test aims to measure in advance.

By the 1980s, this model was widely accepted as a general approach to validity by Messick (1980, 1988, and 1989). Messick adopted a broadly defined version of the construct model to make it a unifying framework for validity when he involved all evidence for validity (namely content and criterion evidence) into the construct validity. He considered the two models' supporting roles in showing the relevance of test tasks to the construct of interest, and validating secondary measures of a construct against its primary measures. According to Messick (1988, 1989), there are three major positive impacts of utilizing the construct model as the unified framework for validity. Firstly, the construct model focuses on a number of issues in the interpretations and uses of test scores, and not just on the correlation of test scores with specific criteria in specific settings for specific test takers. Secondly, its emphasis lies in how the assumptions in score interpretations prove their pervasive role. Finally, the construct model allows for the possibility of alternative interpretations and uses of test scores. As can be seen from this analysis, the construct validity is based on the interpretations of test scores in "a two-step process, from score to construct and from construct to use" (Kane, 2006:21)

2.2. Examining the content validity of the test

In the previous parts of the literature review, content validity and construct validity have been discussed on their own. In this section, the content validity is going to be examined with a link to the construct validity in some recent researchers' view to explain why the author chose to cover both the content and test performances in the analysis.

As synthesized by Messick (1980), together with criterion validity, content validity is seen as part of construct validity in the view of "unifying concept." However, the current standards suggest five sources of validity evidence in which rather than referring to "types", "categories", or "aspects" of proposes, a validation framework is proposed based on five "sources of validity evidence" (AERA et al., 1999: 11, cited in Sireci, 2009). The five sources include test content, response processes, internal structure, relations to other variables, and consequences of testing. Among them, evidence based on test content "refers to traditional forms of content validity evidence" (Sireci, 2009: 30).

Furthermore, Lissitz and Samuelsen (2007: 482) are "attempting to move away from a unitary theory focused on construct validity and to reorient educators to the importance of content validity and the general problem of test development." Chalhoub-Deville (2009:242) absolutely supported the focus of attention on content validity which should be examined through "the qualities of test content, the interpretation and uses of test scores, the consequences of proposed score interpretation and uses, and theory refinement." The investigation of content validity, according to Chalhoub-Deville (2009), follows the operationalization of content that Lissitz and Samuelsen presented

in their 2007 article. It includes test standards and tasks which are captured by domain description of the test in general, and test specification in particular. As a result, the content validity of the test can be primarily seen from the comparison between the test tasks/items and the test specification. This is what we do before the test event, called "a *priori* validity evidence" (Weir, 2005). After the test event, "*posteriori* validity evidence" is collected related to scoring validity, criterion-related validity and consequential validity (Weir, 2005). To ensure scoring validity, which is considered "the superordinate for all the aspects of reliability" (Weir, 2005:22), test administrators and developers need to see the "extent to which test results are stable over time, consistent in terms of the content sampling, and free from bias" (Weir, 2005:23). In this sense, scoring validity helps provide evidence to support the content validity.

In summary, the current paper followed a combination of methods in assessing the content validity of the reading test. It is a process spanning before and after the test event. For the pre-test stage, the test content was judged by comparing it with the test specification. Later the test scores were analyzed in the post-test stage for support of the content validity by examining if the content of the specific item needs reviewing based on the analysis of item difficulty and item fit to the test specification.

3. Research methodology

3.1. Research subjects

The researcher chose a VSTEP.3-5 reading test used in one of the examinations administered by the University of Languages and International Studies (ULIS), Vietnam National University, Hanoi (VNU). This is one among the four separate skill tests that examinees are required to fulfill in order to achieve the final result of VSTEP.3-5 test. Like other skills, the reading test focuses on evaluating English language learners' reading proficiency from level 3 (B1) to level 5 (C1). There are four reading passages with 10 multiple choice four-option question per passage for test takers to complete in the total time of 60 minutes. The passages range in terms of length and topics. As a case study which is seen to be the basis of future research, this paper only focused on one test.

The particular test assessed was selected at random from a sample pool of VSTEP.3-5 tests which have undergone the same procedure of designing and reviewing. This aims at providing objectivity to the study. Also, only tests that were taken by at least 100 candidates were included in the sample pool to increase the reliability of test score analysis.

3.2. Research participants

For the pre-test stage, three experienced lecturers who have been working in the field of language testing and assessment participated in the evaluation of the test content by working with both the test paper and test specification based on a framework of language task characteristics including setting, test rubric, input, expected response, the relationship between input and response, which is originally proposed by Bachman and Palmer (1996).

The research participants also included 598 test takers who took the VSTEP.3-5 reading test. This population is a combination of majored and non-majored English students at VNU and candidates who are working in a range of fields at various ages throughout the country. Therefore, the test scores are expected to reflect the performance of a variety of English language learners when taking the reading test.

3.3. Research questions

1. To what extent is the content of the reading test compatible with the test specification?

2. To what extent do the reading test results reflect its content validity?

3.4. Research methods and data analysis

The study made use of both quantitative and qualitative data collection. Firstly, an analysis of the test paper comparing it with the test specification was conducted. The framework followed the original one proposed by Bachman and Palmer (1996). This widely used framework in language testing has been applied in previous studies such as Bachman and Palmer (1996), Carr (2006), Manxia (2008) and Dong (2011). However, as analyzed from Manxia (2008), this framework was not designed for any particular types of test tasks or examinations. According to the nature of reading and characteristics of reading tests, "characteristics of the input" and "characteristics of the expected response" are advised to be evaluated. In this study, "input" refers to the four reading passages that test takers were asked questions about during their examination. It involves length, language of input, domain and text level. This is also an adaptation from Bachman and Palmer's model since it is closely related to the test specification – the blueprint or the guidelines of test design that test writers are supposed to follow. "Expected response" aims at the response types and specifically the options of each question. The analysis pointed out how similar and different the test paper under evaluation was written compared with the test specification. To be specific, regarding characteristics of the input, the study compared the length, language of input, domain and text level. In terms of expected response, it is response type and reading skills which are analyzed. The analysis was conducted by comparing these features of the test with the description in the test specification. The data was collected using the Compleat Lexical Tutor software version 6.2 which is a vocabulary profiler tool (http://www.lextutor.ca/), the software provided the statistical data of inputted text based on the research from the British National Corpus (BNC) representing a vocabulary profile of K1 to K20 frequency lists. Moreover, the readability index was checked from the website https://readable.io/ and cross checked with the result from Microsoft Word software. The website showed the level of the text at A, B and C; rather than one of the six levels of the CEFR.

After that, more qualitative data were collected through a group discussion between the researcher and three experts who did the analysis of the test paper. In the discussion, the experts shared their thoughts about the insights of the test related to the proposed and estimated item difficulty level, the characteristics of the stems and options as well as an overall evaluation of the compatibility between the investigated test paper and the reading test specification. These two methods helped collect the data to answer research question one which aims at the compatibility between the test items/questions and the test specification.

Secondly, the test scores were reported with descriptive statistics and item response theory (IRT) results as a means of incorporating examinee performance into the Bachman and Palmer model. IRT is basically related to "accurate test scoring and development of test items" (An & Yung, 2014). There are some parameters than can be calculated; however, this study focused on the item measure which means item difficulty and item fit to see how the examinees' performance in each item/question matches the estimated item/question levels in the test specifications. In this way, we can evaluate the quality of the items with a real pool of examinees.

4. Results and discussion

4.1. Research question 1: To what extent is the content of the reading test compatible with the test specification?

As presented in the methodology, Bachman and Palmer's framework was adopted in this study with a focus on the analysis of characteristics of the input and the response.

Characteristics of the input

In terms of the input, attention was paid to specific features that suited reading passages. Table 1 displays the detailed illustration of the analysis by comparing the requirements in the test specification and the manifestations of the investigated test paper.

Table 1. Characteristics of the input

	Characteristics of the input described in the test specification	**Characteristics of the input in the test paper**
Length	Passage 1, 2, 3: ~ 400 words/passage Passage 4: ~ 500 words/passage	Passage 1: 452 words Passage 2: 450 words Passage 3: 456 words Passage 4: 503 words
Language of input	*Vocabulary* Passage 1, 2: mostly high-frequency words, some low-frequency words Passage 3, 4: more low-frequency words are expected	Passage 1: K1+K2 words 94.31% Passage 2: K1+K2 words 87.23% Passage 3: K1+K2 words 77.13% Passage 4: K1+K2 words 77.41%
	Grammar Passage 1, 2, 3: a combination of simple, compound and complex sentences Passage 4: a majority of compound and complex sentences	Passage 1, 2, 3, 4: the majority is compound and complex sentences
Domain	The passage should belong to one of the four domains: personal, public, educational and occupational	Passage 1 & 2: educational domain Passage 3 & 4: public domain
Text level	Passage 1: B1 level Passage 2 & 3: B2 level Passage 4: C1 level	Passage 1: Level B (Average grade level 6.9 Reading ease: 76.6%) Passage 2: Level B (Average grade level 9.4 Reading ease: 58.4%) Passage 3: Level C (Average grade level 11 Reading ease: 50%) Passage 4: Level C (Average grade level 13.8 Reading ease: 34.5%)

The table shows that the test was generally an effective presentation of the test specification under the investigated characteristics of the input. Most of the description was satisfactorily met in the four reading passages. Regarding the length of the input and domain, all the passages were accepted in the range of word number as the total word counts can fluctuate within 10% of the total number and belonged to reasonable domains with suitable topics. In terms of lexical resource of the input, according to O'Keeffe et al. (2003), Dang and Webb (2016) cited in Szudarski (2017), the first two thousand words, i.e. K1 and K2 words are the high-frequency ones and the rest from K3, academic word list and off-list words. Based on these studies, it can be claimed that the proportion of high and low-frequency words in the four passages satisfied the test specification. Last but not least, the text level should be mentioned in this study as it is of the priority of the test design according to the test specification. As the goal of the test is to distinguish examinees' reading proficiency level at levels B1, B2 and C1, the requirement from the test specification also aims at these three levels as seen from the table. The four passages were checked with the website https://readable.io/ and Microsoft Word; however, it is admitted that there is not any official tool to assess the readability of the inputted text. Therefore, the result should be considered a reference to the study which partially reflects the requirement and needs more discussion with the test reviewers.

With regards to the discussion with the three reviewers, positive comments on the quality of the texts were noted. Reviewer 1 saw a good job in the capability to discriminate the level of the four passages, i.e. the difficulty level changed respectively from passage 1 to passage 4. Also the variety of specific topics allowed for examinees to demonstrate a breath of understanding. This feedback was also reported from reviewer 2 and 3. Reviewer 2, however, pointed out the problem with grammatical structures that the above table displays. The percentage of compound and complex sentences in all four texts outnumbered the simple ones, which might be challenging for readers at lower levels like B1 to process. For the text level, the experts emphasized the role of test developers in evaluating the difficulty of the input which should not solely depend on the readability tool. It is ultimately the test writer's expertise at analyzing the language of the passage that best assesses the reading level of a text.

Characteristics of the response

Following the analysis suggested by Manxia (2008), this paper focused on two features of the response, namely response type and reading skills. Typically, the reading skills should be mentioned in the input regarding the test item; however, to make the analysis coherent and compatible with the test specification, the researcher decided to keep both the test item and the item options in this part. The analysis results can be seen in Table 2.

Table 2. Characteristics of the response

	Characteristics of the response described in the test specification	Characteristics of the response in the test paper
Response type	Multiple choice questions with four options	Multiple choice questions with four options
Reading skills	Reading for main idea Reading for specific information/details Reading for reference Understanding vocabulary in context Understanding implicit/explicit author's opinion/attitude Reading for inference Understanding the organizational patterns of the passage Understanding the purpose of the passage	Reading for main idea Reading for specific information/details Reading for reference Reading for vocabulary in context Reading for author's opinion/attitude Reading for inference Understanding the organizational patterns of the passage Understanding the purpose of the passage

The table shows that the test met the requirement of the test specification in terms of response type and reading skills. All forty items were written in the form of multiple choice with four options and covered a number of sub-skills that the test specification suggested for different question levels. For an in-depth analysis into the test items, to evaluate the extent they matched the test specification, i.e. the content coverage, three reviewers were arranged to work individually and discuss in groups to assess the quality of test items. In the assessment, firstly, all reviewers agreed that there were a range of question types that aimed at different skills in the test. All these types appeared in the test specification. Secondly, the majority of the questions or items appropriately reflected the intended item difficulty. The test covers three CEFR levels (B1, B2, and C1); furthermore, the test specification adds three levels of complexity (low, mid, high) to each level, creating nine levels of questions from the test. Due to the confidentiality of the test, a detailed description cannot be presented here for either the test specification or the current test itself. In this research, the reviewers all claimed that nine levels of difficulty could be pointed out from the forty items. However, a problem came about in this aspect when fewer B1 low questions were found than planned. Otherwise, there were more B1 mid, B2 low and B2 mid questions in the investigated paper compared to the test specification. There was an agreement among the test reviewers that the number of high-level items was more than that in the test specification. This explains a finding that low-level test takers had difficulty with this test, i.e. the test was more difficult than the requirement of the test specification. The reviewers also commented on the tendency to have several questions that test a specific skill in one passage. For example, in passage 2, four out of ten questions focus on sentence meaning, whether explicitly or implicitly expressed; and another passage had one question for main idea and one question for main purpose. In fact, this is not mentioned in the test specification as a constraint for the test designers; however, the test specification recommends that the test writer should balance and vary the kind of skills tested in each passage particularly and in the whole test overall.

To sum up, it can be concluded in this study that the test paper followed the test

specification with all requirements regarding its content. The analysis of the input and response by presenting statistical data and reviewers' feedback made it possible to confirm the content validity via content relevance and content coverage of the test.

4.2. Research question 2: To what extent do the reading test results reflect its content validity?

The evidence to answer this question was obtained from the analysis of test scores by using the descriptive statistics and the IRT model.

Descriptive statistics

The descriptive statistics of the reading test are presented in Table 3 and Figure 1.

Table 3. Score distribution of the test (N = 598)

Items	N	Min	Max	Mean	Mode	SD	Skewness	Kurtosis
40	598	4	37	15.080/40	15	5.082	.288	-.153

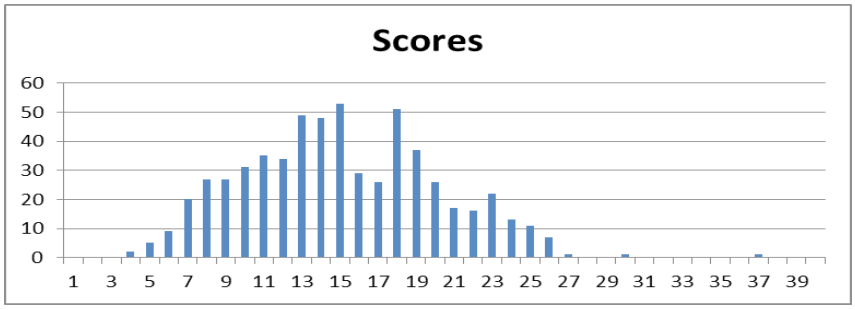

Figure 1. Score distribution of the test (N = 598)

It can be seen that the mean score is relatively low at 15.080/40. More importantly, the skewness is positive (.288), showing that the score distribution is slightly skewed to the right. This indicates that the reading test was rather difficult to the test takers. The initial analysis of descriptive statistics strengthened the comments that the three experts made about the level of the test providing an overall impression that it is more difficult than what is required in the specification.

IRT results

In order to get a detailed description of the test items and personal performance, the IRT results which focus on item difficulty and item fit to the test specification were collected. These are significant tools to assess whether the content specification is maintained in the real test.

As shown in Table 4, the mean measures (difficulty) for item and person are .00 and -.62 respectively, which means the test takers found the test difficult in general. Additionally, it is evident from Table 4 that the infit and outfit statistics for both item and person are within the desirable range which is from .8 to 1.2 for the mean square and from -2 to +2 for the *z*-standardized values (Wright & Linacre, 1994). Therefore, it is safe to say that overall, the data fit the model expectations for both person and item. That is, the test is productive for measuring the construct of reading comprehension, and the data have reasonable predictability in general.

Table 4. Measure, fit statistics, reliability, and separation of the test (N = 598)

	Measure		Infit		Outfit		Reliability	Separation
	Mean	SE	MNSQ	ZSTD	MNSQ	ZSTD		
Item	.00	.10	1.00	-.3	1.03	.0	.99	9.37

Furthermore, the reliability estimate for reading items and the item separation resulted from Rasch analysis are high at .99 and 9.37 respectively, showing very high internal consistency for the items in the reading test. Simply put, the test has a wide spread of item difficulty, and the number of test takers was large enough to confirm a reproducible item difficulty hierarchy. This point matches the description in the test specification that the item difficulty levels range from B1 low to C1 high; and also matches the qualitative analysis from the three test reviewers presented in research question 1.

Item and person measure

First, a correlation analysis was run to examine the correlations between the person measure and the test takers' raw scores, and between the item measure and the proportion correct p value. The results are presented in Table 5, which shows that the correlations are nearly perfect, very close to ±1. From such results, the reading raw scores can be used legitimately to determine the performers' level of reading proficiency.

Table 5. Correlations between person measure and raw scores, item measure and proportion correct (N = 598)

	Person measure	Item measure
Raw scores	.995***	
Proportion correct (p)		-.992***

*** $p < .001$

Secondly, the item measure (item difficulty) of the test was investigated through Rasch analysis. Table 6 provides the logit values of items which represent the difficulty of items (item measure) estimated by the Rasch model. In the Rasch model, the item with the higher logit value is more difficult, thus requiring a higher ability to solve. Figure 2 illustrates the spread of test takers' reading proficiency levels and the difficulty range of reading items over the same measure scale. As observed from the table and the figure, the item difficulties in the reading test ranged widely from -2.9 to 2.05 with the mean set at 0 by the model. Items 2, 13, and 28 are the most challenging while items 1, 11, and 7 are the easiest. It is easily seen from Figure 2 that the spread of item difficulty covered nearly the whole range of all persons' abilities. Only the persons at the top and bottom of the scale did not have the items of equivalent levels. That is, the easiest item seemed difficult for several examinees, and there were a few examinees whose reading proficiency surpassed the highest level tested. However, in general, the test could measure the proficiency of the vast majority of test takers. That the test cannot measure English reading proficiency at either extreme (low and high level) should not be considered detrimental to the test quality because the VSTEP.3-5 does not aim at identifying examinees' English proficiency at all six CEFR levels. Instead, the test targets are only three levels B1, B2, and C1. Therefore, if the examinees are at level A1, A2, or C2, their ability is not likely to be well measured by the VSTEP.3-5. It can be considered that the test items fulfilled their purpose to focus on three specific levels of the CEFR, rather than spread through all six levels.

Table 6. Item measure and item fit of the test (N = 598)

Item	Measure	Infit		Outfit	
		MNSQ	ZSTD	MNSQ	ZSTD
1	-1.78	0.90	-2.24	0.83	-2.87
2	2.05	1.06	0.51	1.57	2.99
3	0.31	0.86	-3.57	0.83	-3.36
4	-1.52	0.80	-5.47	0.73	-5.70
5	-0.45	0.94	-2.55	0.93	-2.37
6	-1.28	0.84	-5.07	0.80	-4.90
7	-2.09	0.89	-2.02	0.78	-2.91
8	-0.77	0.96	-1.78	0.95	-1.70
9	-1.32	0.89	-3.34	0.85	-3.57
10	0.17	1.03	0.77	1.04	0.80
11	-2.02	0.95	-0.95	0.83	-2.36
12	0.50	1.05	1.08	1.09	1.43
13	1.69	1.00	0.02	1.15	1.11
14	-0.33	0.95	-1.91	0.95	-1.57
15	-0.41	1.00	0.09	1.01	0.28
16	-0.30	0.95	-1.74	0.95	-1.44
17	-0.26	1.01	0.52	1.01	0.42
18	0.37	1.04	0.95	1.08	1.33
19	0.27	0.98	-0.57	0.99	-0.17
20	0.38	1.07	1.74	1.10	1.79
21	-0.53	1.04	1.76	1.05	1.75
22	-0.23	1.02	0.82	1.04	1.01
23	-1.03	0.97	-1.14	0.97	-0.89
24	0.36	1.09	2.24	1.15	2.65
25	0.27	1.07	1.74	1.09	1.63
26	-0.33	0.93	-2.88	0.93	-2.27
27	-0.09	1.06	2.08	1.10	2.44
28	1.65	1.11	1.11	1.40	2.72
29	0.07	1.01	0.27	1.04	0.84
30	-0.03	0.91	-2.86	0.90	-2.61
31	1.05	1.06	0.94	1.26	2.64
32	0.72	1.04	0.71	1.09	1.18
33	0.78	1.04	0.73	1.10	1.35
34	0.30	1.10	2.58	1.14	2.60
35	0.62	1.09	1.78	1.14	2.02
36	0.74	1.11	2.02	1.19	2.44
37	0.76	1.03	0.62	1.08	1.09
38	0.43	0.98	-0.43	0.96	-0.60
39	0.74	1.08	1.47	1.13	1.66
40	0.54	1.01	0.28	1.02	0.34

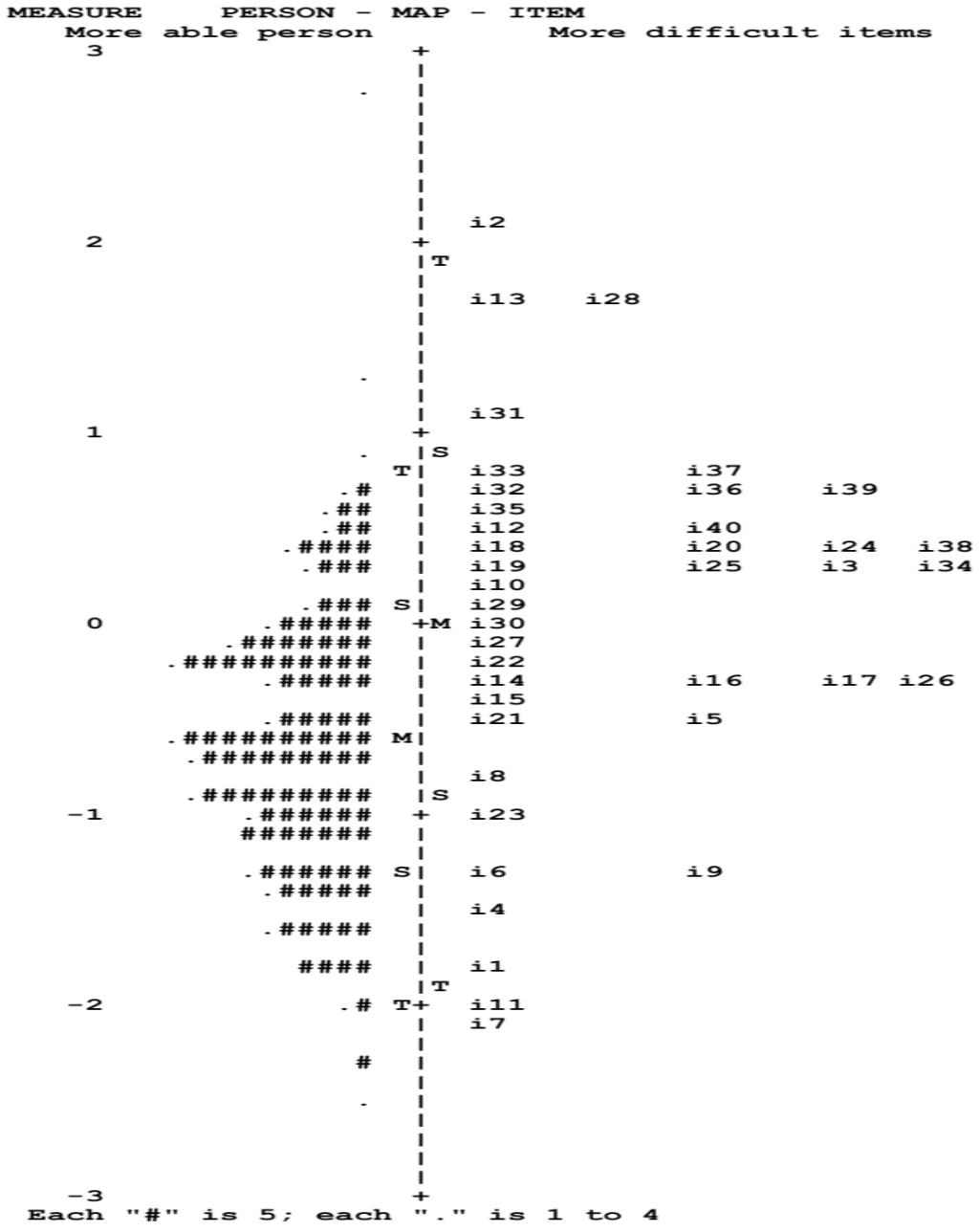

Figure 2. Person maps of items of the test (N = 598)

Furthermore, the Rasch analysis also reveals the actual difficulty of the items. It is illustrated in Figure 2 that several items do not follow the difficulty order they were intended for. For example, at the top of the scale, items 2 and 13, which were designed to be at lower level, are shown to be above items 31 and 32 which are of higher level. From the figure, more problematic items can be seen as item 3, 10, and 12 which are more difficult than expected; whilst items 23 and 26 are easier. This means some items do not

perform as expected with this group of test takers. As a result, content review is necessary for them. This point is worth more effort of item review before and after the test as it is directly related to the test content regarding item difficulty. Again, this is what the three test reviewers commented in their analysis when showing that it was hard to find low-level items in the test, while more items were found at mid or higher levels compared with the test specification. It can be claimed that the statistical analysis did support the test analysis of content validity in this study.

5. Conclusion

5.1. Summary of major findings

The qualitative and quantitative data analysis has shown that both the test content and test results reflect its content validity. In the first place, the paper followed the guidelines of the test specification when considering its input characteristics such as length, language, domain, text level and its response features of type and skills. This claim is made from the data comparison and the three test reviewers' feedback. What was developed in the test covered the main requirements of the test specification, and this is proved from the analysis of the test paper made by the reviewers. Some problems, nevertheless, were seen to remain with the study. Texts chosen for the test had a majority of compound and complex structures while the first two passages should contain more simple structures according to the test specification. With an online readability tool, the analysis also showed that the readability level of one passage was higher than it should have been. This is not a particularly big concern, but it is worth noting for future test review.

Secondly, a wide range of difficulty levels in the questions that spread from B1 low to C1 high was reported, following the CEFR levels applied for VSTEP.3-5. There exists an agreement between reviewers about the variety of item difficulty levels throughout the test, especially that all nine required levels appear in the test. However, the analysis from the three experts and the test scores reveal a gap between the proposed difficulty and actual difficulty of some items. In the test, some questions did not follow the difficulty order assigned for them, and the levels seemed to be higher or lower than planned. This leads the researcher to believe that the test is a bit more difficult than what is designed in the test specification.

As a result, it is necessary that the specific items pointed out from the analysis be edited. The item edition should begin by reviewing reading skills assessed by the question to reduce the concentration of such questions for any one text. Additionally, some options that were excessively challenging in terms of lexical and grammatical structures should be rewritten.

Generally speaking, the investigated test can be considered a success to guarantee the content validity of VSTEP.3-5 reading comprehension test.

5.2. Limitations of the study

It cannot be denied that the current research has some limitations which should be taken into consideration for future studies. As this is a small-scale study, the focus was one reading test with three reviewers involved. Therefore, to reach generalized conclusions, more tests should be investigated.

References

Vietnamese

Nguyễn Thúy Lan (2017). Một số tác động của bài thi đánh giá năng lực tiếng Anh theo chuẩn đầu ra đối với việc dạy tiếng Anh tại Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội. *Nghiên cứu Nước ngoài, 33*(6), 123-141.

English

Alderson, J.C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

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

Bachman, L. & Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Carr, N.T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23*(3), 269-289. Available through http://ltj.sagepub.com/. Accessed 01/03/2018 14:15.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R.W.Lissitz (Ed.), *The concept of validity* (pp. 241-259). Charlotte, NC: Information Age Publishing, Inc.

Cronbach, L.J. (1971). Test validation. In R.L.Thorndike (Ed.), *Educational Measurement* 2nd ed. (pp. 443-507). Washington, DC: American Council on Education.

Dong, B. (2011). A content validity study of TEM-8 Reading Comprehension (2008-2010). *Kristianstad University Sweden*. Available through www.diva-portal.se/smash/get/diva2:428958/FullText01.pdf *Accessed 20/02/2018 09:00.*

Henning, G. (2001). *A guide to language testing: Development, evaluation and research. Beijing*: Foreign Language Teaching and Research Press.

Kane, M.T. (2006). Validation. In R.L.Brennan (Ed.), *Educational Assessment* 4th ed. (pp. 17-64). New York: American Council on Education.

Lissitz, R.W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448.

Manxia, D. (2008). Content validity study on reading comprehension tests of NMET. *CELEA Journal, 31*(4), 29-39.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologists, 35*, 1012-1027.

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

O'Keeffe, A. & Farr, F. (2003). Using language corpora in language teacher education: pedagogic, linguistic and cultural insights. *TESOL Quarterly*, *37*(3), 389-418.

Nguyen Thi Quynh Yen (2016). Rater Consistency in Rating L2 Learners' Writing Task. *VNU Journal of Science: Foreign Studies*, *32*(2), 75-84.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W.Lissitz (Ed.), *The concept of validity* (pp. 19-39). Charlotte, NC: Information Age Publishing, Inc.

Szudarski, P. (2018). *Corpus Linguistics for Vocabulary*: *A Guide for Research*. *Routledge Corpus Linguistics Guides*. New York: Routledge.

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Wright, B.D. & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370-371.

NGHIÊN CỨU TÍNH GIÁ TRỊ NỘI DUNG CỦA BÀI THI ĐỌC THEO ĐỊNH DẠNG ĐỀ THI ĐÁNH GIÁ NĂNG LỰC SỬ DỤNG TIẾNG ANH BẬC 3-5 (VSTEP.3-5)

Nguyễn Thị Phương Thảo

Trung tâm Khảo thí, Trường Đại học Ngoại ngữ, ĐHQGHN,

Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Tóm tắt*:* Bài viết này trình bày kết quả của một nghiên cứu về tính giá trị nội dung của một bài thi Đọc theo định dạng đề thi đánh giá năng lực sử dụng tiếng Anh bậc 3-5 (VSTEP.3-5) thông qua phân tích số liệu định lượng và định tính. Mục đích của nghiên cứu là đánh giá tính phù hợp của nội dung đề thi với bản đặc tính kĩ thuật của đề thi và năng lực thực tế của thí sinh dự thi. Nghiên cứu mời ba giảng viên có chuyên môn về lĩnh vực khảo thí phân tích nội dung đề theo khung phân tích tác vụ đề thi của Bachman và Palmer (1996). Đồng thời, nghiên cứu phân tích điểm thi thực tế của 598 thí sinh thực hiện bài thi này. Nghiên cứu chỉ ra rằng tính giá trị nội dung của đề thi được khảo sát phù hợp với các công cụ phân tích. Tuy nhiên, đề thi vẫn cần được kiểm tra lại để hoàn thiện với một số vấn đề nghiên cứu đã chỉ ra.

*Từ khóa***:** kiểm tra đánh giá ngôn ngữ, tính giá trị nội dung, bài kiểm tra kĩ năng đọc hiểu, bài thi chuẩn