ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 1, No. 5, pp. 525-529, May 2011
© 2011 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/tpls.1.5.525-529

A Comparative Study of the Test Tasks and Target Use Tasks

Seyyed Ali Ostovar-Namaghi
Shahrood University of Technology, Iran
Email: saostovarnamaghi@yahoo.com

***Abstract*—This study aimed at exploring possible construct underrepresentation and construct irrelevant difficulty in Shiraz University Language Proficiency Test (SULPT). To this end, a checklist of target-use tasks was developed based on an analysis of the data obtained from an open-ended questionnaire. The checklist, comprising 24 areas of target use tasks, was administered immediately after the test. The students rated the degree to which the test tasks correspond to each of the items included in the checklist on a scale of 0-4. They also rated the comparative difficulty of test tasks and target use tasks. The frequency of the answers was analyzed using χ^2 test of significance. All items, except for item 4, showed very negligible p values ($p < 0.05$). Such small p-values indicate that there is a significant difference between the test tasks and target use tasks. As for construct irrelevant difficulty, nearly 80% of the students indicated that the reading tasks included in the test were more difficult. This indicates that the test tasks produce construct-irrelevant difficulty.**

***Index Terms*—construct underrepresentation, construct irrelevant difficulty, test task, target use task**

I. Introduction

Although the communicative approach has always laid great stress on authenticity, even a cursory reading of the relevant literature will bring to light a confused and contradictory picture. Taylor (1994) points out that while there are relatively clear definitions available of what is meant by authenticity in teaching materials and texts, there is much less agreement about what constitutes authenticity of context and of the task or activity. Similarly Heltai (1998) states that the term authentic is rather problematic since it has several meanings and is often used in rather a loose manner. The first and the original meaning of authenticity refers to texts. Widdowson (1978) recommends the term genuine to differentiate authentic texts from contrived or doctored texts. Moreover, he introduced the distinction between 'genuineness' and 'authenticity' of language by arguing:

To present someone with a set of extracts and require him to read them not in order to learn something interesting and relevant about the world but in order to learn something about the language being used is to misrepresent normal language use to some degree. The extracts are, by definition, genuine instances of language use but if the learner is required to deal with them in a way, which does not correspond to his normal communicative activities, and then they cannot be said to be authentic instances of use. Genuineness is a characteristic of the passage itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response. (Widdowson, 1978, p. 80)

Morrow (1997) also focuses on the authenticity of texts. He argues that in real life language is seldom simplified to suit a person's linguistic ability. It contains irregularities, interruptions, irrelevancies depending on the context of interchange. It should be noted, however, that Morrow does not limit himself to authenticity and argues that language tests should measure the seven features of communication. These seven features are; interaction-based, unpredictability, context, purpose, performance, authenticity, and behaviour-based.

Lewkowicz (2000) believes that "Sticking to authentic texts has proved unhelpful since the very act of extracting a text from its original source, even if it is left in its entirety, could be said to disauthenticate it" (p.46). Taking this dilemma into account, Spolsky (1985) focused on authenticity of tasks. He believed that in the post-modern period authenticity of task has come to have special importance. He states that authentic tasks simulate real-life tasks but the test taker need to cooperate and be willing to abide by the rules of the game if simulations are to be successful in testing situations. Otherwise the validity and fairness of the assessment procedure remain suspect. Throughout the 19880s, the authenticity debate remained on the use of authentic texts and tasks taken from real life situations.

Bachman in the early 1990s suggested that there was a need to make a distinction between situational and interactional authenticity. Situational or real life (RL) approach considers the extent to which test performance replicates some specific non-test language performance. According to Bachman (1990),"The primary concerns of this approach are: (1) the appearance or perception of the test and (2) the accuracy with which test performance predicts future non-test performance"(p. 301). Therefore, to the extent that test tasks resemble the real life tasks, the test is seen to be a direct test of proficiency and by definition valid.

The second approach is the interactional ability (IA) approach which aims to ensure authenticity by identifying and incorporating critical features of communicative language use, rather than attempting to capture holistic language use situations. The IA approach is based on a carefully developed framework, which specifies both the components of communicative language ability and the characteristic of the test method. The basic premise of this approach is that it is the interaction between the characteristics of the test taker and the characteristics of the test task that determines authenticity.

Bachman and Palmer (1996) divide the notion of authenticity into two components: authenticity and interactiveness, roughly corresponding to RL and IA approaches in Bachman (1990). Bachman's perspective towards authenticity has changed over time. While in 1990 Bachman rejected the RL approach as the sole basis of test validity, he now advocates a consideration of authenticity in terms of the degree to which the test tasks resemble real life tasks. They define authenticity as "the degree of correspondence between the characteristics of a given test tasks and the features of TLU task"(Bachman and Palmer, 1996, p. 23).

Many authors have highlighted the need to ensure that tasks engage and challenge the abilities, which we wish to test. Bachman and Palmer (1996) term this dimension interactiveness, defining it as the degree of involvement of test takers competencies and other individual characteristics in accomplishing a task. It thus refers to the interaction of the competencies with the task. It should be noted, however, that for Bachman interaction is interpersonal.

McNamara (1996) rejects this restricted notion of interaction by arguing that attention should be given to social dimensions of interaction in language tests, including the interaction between the test-takers, examiners and other interlocutors. He backs up his position by referring to constructivsm, functionalism, and Vygotssky's (1978) zone of proximal development (ZDP). The thrust of his argument is that interaction is something interpersonal. As such performance is affected by the rater and other interlocutors.

From the previous discussion it is evident that authenticity has been an ever-changing concept. Although Bachman has recently advocated the RL approach, it has been attacked on several counts. Stevenson (1985) points out that instances of language use are by definitions context-dependent and hence unique. Therefore, authenticity is not transferable. But the main criticism with the real life approach is that duplication or simulations of real life tasks do no solve the problem of coverage. Skehen (1984) cogently addresses this problem by arguing that merely making an interaction authentic does not guarantee that the sampling of the language involved will be sufficient, or the basis for wide-ranging and powerful prediction of language to other situations. Skehan's concept of sufficiency is in line with Messick's (1996) notion of ‘construct under-representation’. These two concepts act synergistically to add a new dimension to authenticity and hence fill in the gap in our present discussions about authenticity.

Messick (1996) presents authenticity and directness as validity standards. To him, authenticity ensures that nothing important be left out in assessing the focal construct. Similarly, directness ensures that nothing irrelevant be added that interferes with the construct. He argues:

In the threat to validity known as construct under-representation (which jeopardizes authenticity) the assessment is deficient: the test is too narrow and fails to include important dimensions or facets of the focal constructs. In the threat to validity known as construct-irrelevant variance (which jeopardizes directness), the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct (Messick, 1996, p.244).

These threats are more likely when language skills are included in the construct definitions. Bachman and Palmer (1996) warn that skills should not be included in construct definition by arguing that:

Characterizing skills in terms of channel (audio, visual) and (receptive, productive) is inadequate on two counts: first, these features fail to capture important differences among language use activities that are within the same skill [construct under-representation]… Second, this approach to distinguishing skills treats them as abstract aspects of language ability, ignoring the fact that language use is realized in specific situated language use tasks (p. 79).

Another reason for not including a specific skill in the construct definition is that as Widdowson (1978) has pointed out, many language use tasks involve more than one skill. What he calls the communicative ability of 'conversion' involves listening and speaking, while what he calls 'correspondence' involves reading and writing. Therefore rather than defining a construct in terms of skills, Bachman and Palmer (1996) suggest that the construct definition include only the relevant components of language ability and the skills elements be specified as characteristics of the tasks in which language ability is demonstrated.

Nonetheless, many of the popular tests are based on skills and components. Using the format and components of the popular tests has placed us in a position to defend the indefensible. Being skill-based, SULPT may not include many important aspects of the focal constructs, i.e., it under-represents the constructs. Moreover, since test tasks are different from target language use tasks, they are likely to include construct-irrelevant difficulty. Knowing that both threats are operative in all assessment, the researcher aims to validate the reading section of SULPT to see whether and to what extent it stands the tests of construct under-representation and construct irrelevant difficulty.

To meet the specific needs of the test takers and to find the relevant reading tasks for PhD students in Iran, this study tried to elicit real life reading tasks from the test takers themselves, i.e., from PhD students at Shiraz University. The elicited tasks presented us with important aspects of reading comprehension. So they can be used as benchmarks against which authenticity and validity can be assessed.

II. Purpose of the Study

Taking construct under-representation and construct-irrelevant variance into account, this study mainly aims at answering the following questions:

- Is there any significant difference between test tasks and target use tasks in SULPT?
- Do the reading tasks included in SULPT produce construct irrelevant difficulty?

III. METHOD

A. *Participants*

Because the population of SULPT is mainly PhD students, it was important to explore how they perceive the test. To this end, a questionnaire was distributed among all test takers, both male and female, from different departments of Shiraz University. A total of 150 students returned the questionnaires.

B. *Instruments*

This study utilized three instruments for data collection. The first instrument is a questionnaire comprising one open-ended question. This questionnaire was used to elicit the different reading tasks PhD students are engaged in during their academic studies. The second instrument is a checklist developed on the basis of students' responses to the questionnaire. This checklist is a summary of the important aspects of reading comprehension for PhD students in Shiraz University. It is composed of 27 items; the third instrument is the reading section of SULPT.

C. *Data Collection Procedure*

To identify the target language use reading tasks, an open-ended questionnaire was distributed among 57 PhD students who reside in the dormitory. To make sure that they have no problem in expressing their ideas, the questionnaire was written in Persian. Then, their responses were summarized and developed into a checklist which is composed of 24 items, covering the different reading tasks students are involved in during their academic studies. The checklist is also in students' native language. Finally, the checklist was administered to all the test takers after the SULPT. 150 students returned their checklist after the exam. Only 108 papers were analyzed because the rest of the papers were either anonymous or in blank.

IV. RESULTS

The students rated the degree to which test tasks correspond to each of the items included in the checklist on a scale of 0-4. Table 1 shows the mean and the standard deviation of the 24 items of the checklist. Taking their means into account, only items 1, 2, 3, and 22 are well represented in the test. Items 15, 17, 18, 20, and 24, which are very important in an EFL context, are completely under-represented. According to Messick (1996) construct under-representation, which is a threat to validity, jeopardizes authenticity.

TABLE 1:
DESCRIPTIVE STATISTICS OF THE ITEMS

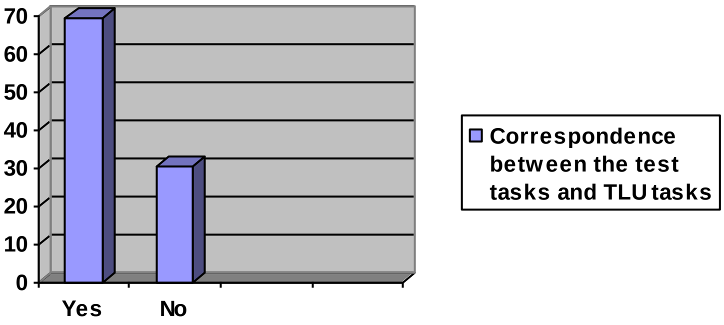
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Mean	3.15	3.20	2.24	1.88	1.74	1.47	1.46	3.05	1.35	1.51	1.46	1.07
SD	1.23	1.15	1.45	1.40	1.33	1.16	1.21	1.23	1.25	1.34	1.46	1.19
Item	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Mean	1.07	1.01	.95	.97	1.04	.98	1.03	.99	1.00	3.07	1.16	.80
SD	1.08	.99	1.01	1.07	1.09	1.12	1.34	1.08	1.11	1.49	1.50	1.07

The frequency of the answers was analyzed using χ^2 test of significance. All items, except for item 4, showed very negligible p values ($p < 0.05$). Such small p-values indicate that there is a significant difference between the test tasks and target use tasks. Table 2 shows that in all items, except item 4, there is a significant difference between test tasks and target use tasks.

TABLE 2: χ^2

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
χ^2	116.7	142.	21.23	3.94	11.33	22.58	21.76	94.84	21.32	13.29	45.47	48.95
Df	4	4	4	4	4	4	4	4	4	4	4	4
P	.000	.000	.000	.414	.023	.000	.000	.000	.000	.010	.000	.000
Item	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
χ^2	44.39	22.70	26.23	59.18	46.72	62.80	89.05	55.23	53.09	128.7	81.42	89.14
Df	4	4	4	4	4	4	4	4	4	4	4	4
P	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Moreover, student's general impression about the degree of correspondence between the test tasks and real life reading tasks indicates little similarity between the two.

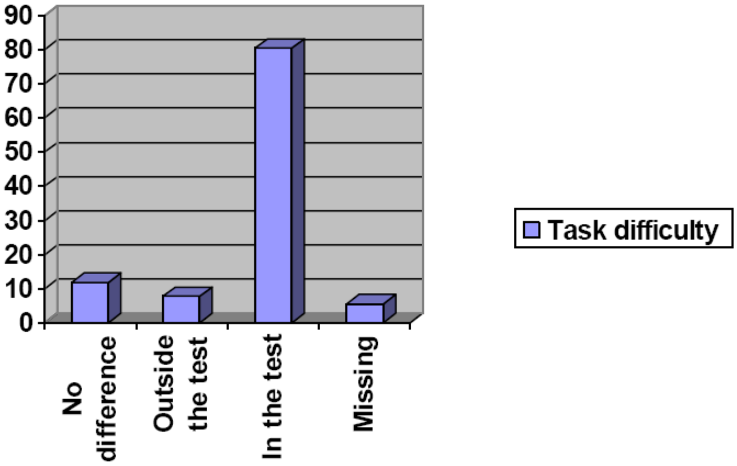


To explore construct irrelevant difficulty, test-takers were asked to mark the comparative difficulty of test tasks and target use tasks. Eighty two test takers pointed out that test tasks were more difficult than the target use tasks. Table 3 gives a summary of the results.

TABLE 3:
FREQUENCIES

	Frequency	Percent	Valid Percent	Cumulative percent
No difference	12	11.1	11.8	11.8
Outside	8.	7.4	7.8	19.6
In the test	82	75.9	80.4	100.0
Missing	5.00	6	5.6	
Total	108	100.0		

The bar graph clearly shows that nearly 80% of the students commented that the reading tasks included in the test were more difficult. This indicates that the test tasks produce construct-irrelevant difficulty.



V. LIMITATIONS OF THE STUDY

First, respondents were not much cooperative since in both phases of data collection the return rate was low. Moreover, in the open-ended questionnaire students were a bit reluctant to respond. However, such a problem is naturally expected due to the nature of the study. It should also be acknowledged that after summarizing the students' open-ended responses, the researcher came up with some items, which seemed either trivial or un-operationalizable. Therefore, they were not included in the checklist. There are still some items in the checklist, which may seem trivial or un-operationalizable but this is a problem with any types of needs analysis, especially in testing EAP. Finally, since the SULPT was not accessible for further scrutiny the analysis was limited to the students' responses. Had the test been accessible, the analysis could have been extended in other dimensions.

VI. DISCUSSION

The results clearly show that the SULPT suffers from construct under-representation and construct-irrelevant difficulty. First, 80.4 percent of the students indicate that reading tasks included in the test are more difficult than real life reading tasks. As such, the test tasks have produced evaluative anxiety that is not operative in the criterion performance. This is due to lack of correspondence between the test tasks and real life tasks. Second, the test does not capture important aspect of reading proficiency. Therefore, the results are not generalizable to those aspects which are excluded. That is, the test scores lack predictive utility.

Moreover, construct under-representation and construct-irrelevant difficulty leads to bad educational practices. Because important aspects of reading proficiency are under-represented in the SULPT, students overemphasize those aspects that are well presented and downplay those parts that are not. Living in the dormitory with PhD students, the researcher has had the chance to observe such unhelpful practices on the part of students. Similarly, since the test employs test tasks which are different from real life reading tasks, students pay undue attention to overcoming the irrelevant difficulty as opposed to fostering reading proficiency. If these threats are minimized and if test tasks are deeply rooted in students’ needs, the student's immediate goal- i.e., to achieve a given test score- materializes his long-term goal-i.e., to enhance his language proficiency. As Messick (1996) states, "the best defense is to minimize such irrelevant difficulty in the first place as well as construct under-representation (p. 253). Having done this, one does not feel guilty to see that even those who can do academic reading tasks quite efficiently are under the heavy and stressful pressure of the SULPT. Knowing the importance of authenticity, and being aware of the fact that authenticity affects performance, test developers are challenged to eliminate the problems with the current form of the SULPT, i.e., construct under-representation and construct-irrelevant variance.

REFERENCES

[1] Bachman, L. F. & Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University Press.

[2] Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

[3] Heltai, P. (1995). Communicative language tests, authenticity and the mother tongue. *NovELTy*, 8(2), 65-74.

[4] Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing*, 17(1), 43-66.

[5] McNamara, T. F. (1996). Measuring second language performance. London: Longman.

[6] Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.

[7] Morrow, K. (1979). Communicative language testing: Revolution or evolution. In Jordon R. editor, *Case Studies in ELT*, (pp. 102-107), London: Collins ELT.

[8] Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing*, 1(2), 202, 220.

[9] Spolsky, B. (1985). The limits of authenticity in language testing. Paper Presented at the Seventh World Congress of Applied Linguistics. Brussels, Belgium.

[10] Stevenson, D. K. (1985). Authenticity, validity, and a tea party. *Language Testing* 2, 41-47.

[11] Taylor, D. (1994). Inauthentic authenticity or authentic in-authenticity. *TESL-EJ*, 1(2), 80-91.

[12] Vygotsky, L.S. (1978). Mind in Society: The development of the higher psychological processes. Cambridge, MA: Harvard University Press.

[13] Widdowson, H. G. (1978). Teaching language as communication. Oxford: Oxford University Press.

Seyyed Ali Ostovar-Namaghi was born in 1969. Presently, he runs EAP courses at Shahrood University of Technology, Iran. His chief area of research interest is language teacher education. He has published in a number of leading peer-reviewed journals including the Reading Matrix, Teacher Education Quarterly, the Qualitative Report, and the Asian EFL Journal.