DOCUMENT RESUME

ED 360 838 FL 021 415

AUTHOR Ghonsooly, Behzad
TITLE Development and Validation of a Translation Test.
REPORT NO ISSN-0959-2253
PUB DATE 93
NOTE 11p.; For serial publication in which this paper appears, see FL 021 410.
PUB TYPE Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT Edinburgh Working Papers in Applied Linguistics; v4 p54-62 1993

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Comparative Analysis; Criterion Referenced Tests; English (Second Language); Foreign Countries; Higher Education; Item Analysis; Language Proficiency; *Language Tests; Scores; *Test Construction; Testing; Test Items; *Test Reliability; *Test Validity; *Translation
IDENTIFIERS Iran

ABSTRACT

Translation testing methodology has been criticized for its subjective character. No real strides have so far been made in developing an objective translation test. In this paper, certain detailed procedures including various phases of pretesting have been performed to achieve objectivity and scorability in translation testing methodology. In validating the newly-developed objective translation test, the following research questions are asked: (1) What is the reliability of scores of the translation test and how does it compare with the criterion measure; (2) What is the concurrent validity of the test and of the criterion measure? and (3) Are there any factors such as underlying constructs that the translation test and each subtest of the criterion measure may assess? The following general hypothesis is proposed: in measuring the English proficiency of Iranian EST university learners, a translation test is as valid and reliable as a standardized objective test. Results showed significant reliability for the new test. Contains 10 references. (Author/JL)

DELVELOPMENT AND VALIDATION OF A TRANSLATION TEST

BEHZAD GHONSOOLY (DAL)

DEVELOPMENT AND VALIDATION OF A TRANSLATION TEST

Behzad Ghonsooly (DAL)

Abstract

Translation testing methodology has been criticized for its subjective character. No real strides have so far been made in developing an objective translation test. In this paper certain detailed procedures including various phases of pretesting have been performed to achieve objectivity and scorability in translation testing methodology. In validating the newly-developed objective translation test, the following research questions are asked: a) What is the reliability of scores of the translation test and how does it compare with the criterion measure?, b) What is the concurrent validity of the test and of the criterion measure?, c) Are there any factors such as underlying constructs that the translation test and each subtest of the criterion measure may assess? The following general hypothesis is proposed: in measuring the English proficiency of Iranian EST university learners, a translation test is as valid and reliable as a standardized objective test. Results showed significant reliability for the new test.

1. Introduction

As early as the beginning of the twentieth century,the grammar-translation method was disfavoured on the grounds that it did not take into account speaking, writing and listening as important skills of second/foreign language teaching and learning. It was, therefore, excluded from the teaching paradigm. With the exclusion of the traditional method, translation as a testing device was excluded too. Lado(1964) argued that translation tests were highly subjective, referring to the interference of the teacher's taste in scoring a translation test, which resulted in its unreliability. It was also maintained that translation tests lacked the property of scorability (Lado 1964; Harris 1969). The scorability of a language test is defined in terms of how well and easily it is scored. This idea of scorability, which has served as one of the distinguishing features between essay or subjective type questions and the so-called objective tests, draws upon the notion of convenience and speed in scoring a test. Thus, a well-designed test which collects all the responses on a separate sheet and can be scored by machine is much more convenient and less time-consuming and thus more scorable than one which has the responses scattered in the pages of the test. In fact, one might just imagine how difficult an undertaking it may appear for a teacher who is to correct an average number of, for example, 40 students' responses on a rendered text with a length of one or in some cases more than one paragraph.

EDINBURGH WORKING PAPERS IN APPLIED LINGUISTICS, number 4 (1993) ISSN 0959-2253

Taking this into account, it has been argued that scoring essay-type questions including translation tests is not as easy and convenient as,for instance, a multiple-choice question; therefore, they have been judged to be too burdensome and time-consuming.

However, attempts have recently been made to revive translation as a useful device for the purpose of language teaching (Titford 1983; Tudor 1987). As a result of this movement to re-assess the potential contribution which translation can make to ELT after Lado's rather sweeping dismissal of it, new theories of translation have evolved to pave the way for the development of translation teaching activities (see Newmark 1981; Nida 1982). Nevertheless, while translation methodology has been influenced by improvements in translation theory, its testing counterpart has remained untouched. No real advance has so far been made towards constructing an objective translation test to remedy for the above-mentioned deficiencies. This paper is oriented towards the essential procedures for the development of an objective translation test which may fulfil the scorability criterion of the newly developed test and guarantee its objectivity.

2. Design of the study

2.1 Hypothesis and research question

To determine the statistical characteristics of the new translation test, the following hypothesis was adopted: in measuring the general English proficiency of Iranian English for Science and Technology (EST) learners, a translation test would be as valid and reliable as a standardized objective proficiency test. To provide data for testing the hypothesis the following research questions were addressed: a) What is the reliability of the translation test and how does the test compare with the Michigan EFL test? b) What is the concurrent validity of the new translation test and of the criterion measures? c) Are there any common factors such as underlying constructs that the translation test and each subtest of the criterion measure may be assessing?

2.2 Subjects

The total sample of subjects who were exposed to various phases of pre- and post-testing were 315 male and female university students from the Department of Electronics of Tehran University (TU) and Science and Technology University (STU) who had passed ESP courses in the current Iranian educational system. They were supposed to have acquired general English proficiency.

2.3 Instrumentation

Two classes of multiple-choice item tests were administered in this study: the new translation test, which consisted of twenty multiple-choice items and the Michigan test (used as the criterion measure) which comprised forty grammar M/C questions and forty vocabulary M/C questions together with two reading comprehension passages, each of which consisted of five M/C questions.

2.4 Methods of data collection

The decision as to what translation elements should be selected for the construction of the translation test was one of the difficulties in the investigation. Since the content of the translation test was hypothesized to be independent of the content of the materials used in a particular course of instruction, it was not felt necessary to impose any limitation on the content of the test except that the content had to be compatible with the examinees' field of study, namely electronics. Consequently, scientific and technical English texts were chosen as content elements of the translation test. Since each English scientific text (EST) unit of discourse is a coherent paragraph comprising a number of sentences and is too long to be included in the translation test, it was decided to narrow down the task of selection and search for smaller units of discourse, typically sentences. But due to the typological variety of sentences in English, the decision as to which sentence type should be selected posed another problem. It was decided to deal with those rhetorical functions which, as Trimble (1985) argues, are fundamental elements in the organization of an EST paragraph.

2.4.1 Selecting the rhetorical functions

Determining rhetorical functions with regard to the kind and amount of information each provides the reader with, Trimble (1985) distinguishes five major functions and fifteen related sub-functions. Making full use of the rhetorical functions and their related sub-functions in the translation test seemed to be impractical if not impossible. Therefore, setting some criteria for the selection of functions became necessary. Functions and sub-functions were used in the construction of the translation test only if they met these criteria:

1. is always used in written EST discourse;
2. has high frequency of occurrence and usage in academic settings;
3. does not overlap with other functions or sub-functions.

On the basis of the above criteria, the following rhetorical functions and sub-functions were selected.

Rhetorical Function	1	Description
sub-function	1.1	physical
sub-function	1.2	function
sub-function	1.3	process description
Rhetorical Function	2	Definition
sub-function	2.1	formal
sub-function	2.2	semi-formal
Rhetorical Function	3	Classification
sub-function	3.1	complete

Rhetorical Function	4	Instruction
sub-function	4.1	direct
sub-function	4.2	indirect
sub-function	4.3	instructional information
Rhetorical Function	5	Visual-verbal relationship

All the examples of the above-selected rhetorical functions used were taken from EST paragraphs. A preliminary version of the test based on the selected rhetorical functions within EST paragraphs was prepared for different phases of pretesting.

2.4.2 Pretesting

One of the fundamental purposes of pretesting is to draw out a variety of responses which can be used as distractors for the final test items. For this reason care was taken over the different phases of pretesting. These are briefly explained here.

2.4.2.1 Phase 1. Pretest with sample population of students

In this phase, one hundred students at TU were pretested. They were both male and female and were randomly selected from 825 Engineering students who had been registered for English proficiency tests such as TOEFL and the Michigan test. These tests are occasionally administered at TU for those students who are eager to get an objective view of their English proficiency. The purpose of this phase was to elicit different alternatives. Hence, a preliminary version of the test, consisting of forty items in an open-ended form, was given to the subjects. They were required to read each EST paragraph and translate the underlined rhetorical function of each paragraph.

2.4.2.2 Phase 2. Pretest with translation expert

The same forty items in an open-ended form were given to two translation experts who were required to write the most desirable translation for each underlined rhetorical function. The purpose of this phase was to obtain the most appropriate response for each item by comparing students' responses for the construction of the test items and to ensure its objectivity.

2.4.2.3 Selecting the alternatives

As to the correct response, only those responses agreed upon by the translation experts were inserted in the tests as the most desirable choices. Other distractors were selected from among students' responses which did not conform to those of the translation experts. But the decision as to what distractors should be selected for each item appeared to be a problem. To solve the unwanted obstacle and to be objective, a tentative criterion was proposed. The criterion was set such that the distractors should have a high frequency of occurrence and be often used by the students. The most common mistakes elicited from students' responses were mainly those of comprehension of the functions, word for word translation and deviant translation including errors of style, grammar and lexicon. Each item was, therefore, given the following arrangement of choices: 1. the correct response, 2. reading comprehension distractor, 3. word for word translation, 4. deviant response distractor.

2.4.2.4 Phase 3. Pretest with sample population of students

After developing the test in M/C form, in order to ensure the difficulty level of the test items, the items were administered to another population of 55 students of Electronics at STU. An example of a sample item together with transliterations of each alternative and their closest area of meaning is given here.

> The first man to produce a practical steam engine was Thomas Savery, an English engineer (1650-1715), who obtained a patent in 1698 (for a machine designed to drain water from mines). The machine contained no moving parts except hand-operated steam valves and automatic check valves, and in principle it worked as follows: Steam was generated in a spherical boiler and then admitted to a separate vessel where it expelled much of the air. The steam valve was then closed and cold water allowed to flow over the vessel, causing the steam to condense and thus creating a partial vacuum.
>
> 1. *Bokhar mishod tolid dar yek makhzane bokhar va rah yafi be yek luleye joda jaee ke an kharej kard bishtare hava.*[Steam is generated in a steam tank and then entered into a separate vessel where it expelled much of the air.] **Word for Word**
>
> 2. *Bokhar tolid mishod dar yek jush konandeye koravi ke be yek zarfe joda konande vasl shode bud va meghdare ziyadi hava as an kharej mishod.*[Steam is generated in a spherical boiling device which was attached to a separate vessel and a considerable amount of air was coming out.] **Reading Comprehension**
>
> 3. *Bokhar dar digi koravi tahiyye mishod va angah be zarfe digari hedayat mishod ke meghdare motanabehi hava ra ba feshar aghab mirand.*[Steam was generated in a spherical boiler and then admitted to a separate vessel where it expelled much of the air.] **Correct**
>
> 4. *Bokhar dar digi koravi ke be zarfe digari vasl mishod tahiyye shod ke meghdare motanabehi hava ra ba zoor birun kard.*[Steam in a spherical boiler attached to another vessel was generated that pulled out a considerable amount of air by force.] **Deviant**

2.4.2.4.1 Item analysis

To discard and/or revise items that were either too difficult or too easy, the researcher used the classic item analysis technique with the typical range of 0.33 to 0.67. Of the original 50 test items only 20 items remained to fit the standard item analysis range.

2.4.2.5 Post-test with sample population of students

After the necessary revision and clarification of the items, the final version of the translation test was prepared to be administered together with the Michigan test to another group of Electronics students. The testees were 60 male and female students from STU who were randomly selected from among 150 Engineering students.

3. Results

Based on the research questions stated earlier in this paper, statistical analyses were performed. The results for reliability, validity and factor analysis are given below.

3.1 Reliability

Reliability is defined as the extent to which a test produces consistent results under similar conditions with similar subjects. There are various statistical methods for measuring the reliability coefficient of a test (see Hatch and Farhady 1982). One of the most commonly-used ways of determining the reliability coefficient is the measure of internal consistency. In this study, in order to determine the reliability of the translation test and the subtests of the criterion measure, the measure of internal consistency (Kuder-Richardson formula 21) was used. As can be seen in the table below, the reliability of the translation test is lower than that of the subtests of the criterion measure. One of the most important factors which influence the reliability of a test is the number of test items: the more items used in a test, the higher the reliability of that test will be. Taking this into consideration, the main reason for the somewhat lower reliability coefficient of 0.74 may be the insufficient number of test items (the final version of the translation test consisted of 20 items which in comparison to the total 100 test items of the criterion measure is rather few). This being so, the translation test would probably have had a higher reliability coefficient if more items had been used. However, even the reliability coefficient actually achieved is satisfactory and encouraging.

Table 1. Reliability coefficients of the study measures

Subtests	
Grammar	0.90
Vocabulary	0.92
Reading Comprehension	0.93
Translation	0.74

3.2 Validity

Validity is defined as the extent to which a test measures what it is claimed to measure. To determine the validity of the translation test, correlational analysis was carried out. The concurrent validity of the translation test, as can be seen in Table 2., was low and not significant. In attempting to account for this, it should be pointed out that the coefficient of validity is influenced by many factors, including the size of sample. The greater the number of subjects taking a test, the higher the correlation coefficient of

test results will be. This being so it is likely that one of the main reasons for the apparent low correlation of the translation test with the subtests of the criterion measure is the restricted sample of students who took the test (N=60). The correlation coefficient of the two tests might have been increased if a larger sample of test-takers had taken the test. It is also worth mentioning that the translation test and the criterion measure are fundamentally different from each other in terms of the purposes for which they are designed. Whereas the EFL criterion Michigan Test is primarily designed to assess the general language proficiency of the testees irrespective of their field of study, the newly developed translation test is mainly constructed for a specific group of students, namely students of Engineering and more specifically students of Electronics.

While both the criterion measure and the translation test are measures of language proficiency, the latter is more specific in that it claims to assess the language proficiency of the EST university learners. Therefore, it could be argued that there is something specific to the translation test which is not shared by the subtests of the criterion measure and that is the specific variance of the translation test.

Table 2. Correlation coefficients between the translation test and other subtests of the criterion measure

Variable	1	2	3	4
Grammar	*			
Vocabulary	0.27	*		
Reading Comprehension	0.24	0.30	*	
Translation	0.44	0.29	0.20	*

3.3 Factor analysis

Factor analysis, as Hatch and Farhady (op. cit.) point out, is based on the assumption that in any test there are probably one or more underlying traits being assessed. Through factor analysis the information on factors underlying a test is obtained by examining the common variance among items. Using the varimax rotation procedure in the SPSS computer package, the following data were obtained.

Table 3. Varimax factor matrix

Variable	Factor 1	Factor 2
Translation	0.54294	0.49639
Grammar	0.64303	0.48268
Vocabulary	0.83213	-0.16086
Reading Comprehension	-0.04363	0.86164

The data show us that there are loadings on factor 1 with vocabulary, grammar and translation. Factor 2 is heavily loaded with reading comprehension and moderately loaded with translation and grammar. Factor 2 and factor 1 contribute negatively as underlying factors for the vocabulary and reading comprehension respectively. The most crucial step in the interpretation of the above matrix is that of labelling these factors. It can be observed that factor 1 is highly loaded with grammar and vocabulary while reading comprehension contributes negatively to factor 1. Due to the function of the grammar and vocabulary tests which are considered to be discrete items, factor 1 could be labelled the discrete factor or comprehension of smaller chunks of language. On the other hand, factor 2 contributes negatively as an underlying factor for the vocabulary and is heavily loaded with reading comprehension and to some degree with grammar and translation. Given the integrative purposes for which reading comprehension passages are devised, and the negative load of vocabulary as a discrete item on factor 2, the second factor may be labelled integrative factor or comprehension of larger chunks of language. Factor 2 is also loaded with grammar, a discrete item type. This is probably due to the fact that grammatical knowledge is required for understanding a piece of text, namely, reading comprehension passages.

Taking the translation variable into account, it appears that factor 1 and factor 2 both contribute, if not highly, at least moderately to the translation. Thus, on this interpretation of the factor matrix the translation test may be labelled both as a discrete item and an integrative one.

4. Conclusion

The potential contribution of neglected translation methodology to ELT has recently been re-assessed. While translation methodology has been influenced by improvements in translation theory, its testing counterpart has been less enriched. The main purpose of this project was to develop procedures for the constru... of an objective translation test. The procedures were designed to eliminate the possibility of subjectivity in the test and to achieve one of the essential properties of an objective test, called scorability. Compared with some batteries of language testing methods (mainly discrete tests (DP) and integrative tests (IN)) the translation test developed in this study has some advantages. Firstly, the translation test does not have the deficiency of the DP test, which has been criticized for not being able to take into account extra-linguistic factors (see Oller 1976); rather it is constructed at the level of a meaningful coherent unit of discourse. This means that every example of a rhetorical function used in this study has the property of being used in a natural context. Therefore, the translation test developed in this study does not violate the assumption of 'incoherent segments', the outstanding negative property of DP tests. Secondly, the translation test does not have the problem of independence of items which has raised doubts about the reliability of the cloze test (see Farhady 1980). Thirdly, through factor analysis, it has been shown that the translation test devised in this study can function not only as a discrete point test but also as an integrative test. Accordingly, the translation test can be supposed to assess both skills relating to the comprehension of smaller chunks of language (i.e. grammar and vocabulary) and those which relate to the comprehension of larger chunks of language (i.e. reading comprehension).

Further investigations are needed to shed more light on translation testing methodology. However, in our attempt to objectify translation tests we should be careful not to underestimate the potential value of the so-called subjective tests. We

must always remember that the real merit of a translation test lies in its authentic practice of rendering a text. By carefully designing an open-ended translation test and training translation raters as well as specifying various weighting or scores for different types of translation errors, we ma, achieve objectivity in translation testing methodology.

Acknowledgement

I wish to acknowledge my sincere thanks to Alan Davies for editing the first draft of this paper and particularly Cathy Benson for her insightful comments on this project. I alone am responsible for any mistakes in this paper.

References

Farhady H. 1980. Justification, development and validation of functional language testing. Unpublished PhD. dissertation. University of California.

Harris D.P. 1965. Testing English as a Second Language. New York:McGraw-Hill Book Company.

Hatch E. and H. Farhady. 1982. Research Design and Statistics for Applied Linguistics. Rowley, Mass: Newbury House Publishers Inc.

Lado R. 1961. Language Testing. New York: McGraw-Hill Book Company.

Newmark P. 1981. Approaches to Translation. Oxford: Pergamon.

Nida E.A. and C. Taber. 1982. The Theory and Practice of Translation. Leiden: Brill.

Oller W.J. 1976. Language Tests at School. London: Longman.

Titford C. 1983. Translation for advanced learners. ELT Journal. 37/1: 184-90.

Trimble L. 1985. English for Science and Technology : A Discourse Approach. Cambridge: CUP.

Tudor I. 1987. Using translation in ESP. ELT Journal. 41/4.