Journal of Applied Psychology *1973,* Vol. 57, No. 3, 328-334

# EVALUATING LANGUAGE TRANSLATIONS:

# EXPERIMENTS ON THREE ASSESSMENT METHODS[1]

H. WALLACE SINAIKO[2] AND RICHARD W. BRISLIN[3]

*Institute for Defense Analyses, Arlington, Virginia*

Experiments were run to assess three ways of evaluating the quality of language translations: back translation, knowledge testing, and performance testing. Twelve professional English-to-Vietnamese translators processed approximately 10,000 words of technical material (i.e., a helicopter maintenance manual). Subjects took knowledge tests or performed a difficult maintenance task using translated materials. Vietnamese Air Force technicians and U.S. Army technicians served as primary subjects and controls, respectively. The analysis of back translations showed the frequency and types of translation errors that occurred. Knowledge test scores satisfactorily discriminated different quality levels of translations. The performance tests demonstrated *(a)* the impact of translation quality on performance, *(b)* the value of working in one's native language (vs. having to learn English), and *(c)* the importance of providing high-fidelity translations where a complex task is to be done.

Technical documents—maintenance manuals, technical orders, and instructional material—are as critical in the use of complex military equipment as the hardware itself. Training men how to use and service equipment *is* inevitably tied to the quality of the technical documents they are given. And in the case of material intended for foreign nationals—in this research, the Armed .Forces of the Republic of Vietnam—there is an added class of problems: Most of the intended users do not read English, and documents must be translated. In addition, the Vietnamese language contains very few technical terms. Language translation methods are as old as the printed word; but surprisingly, there *is* almost no literature on the technology of translation and on the accuracy that can be expected from it. One is forced to rely on the subjective views of translators or bilingual readers about the quality of a translated document.

[1] The authors would like to thank Vu Tam Ich, Nguyen Nhan, and the officers of Fort Eustis, Virginia, for making this work possible. Further information on all aspects of this investigation (e.g., background, more examples of technical English translated, performance task) can be found in Sinaiko and Brislin (1970).

[2] Requests for reprints should be sent to H. Wallace Sinaiko, who is now at the Smithsonian Institution, Arts and Industries Building, Room 3101; Washington, D.C. 20560.

[3] *N*ow at the Culture Learning Institute, East-West Center, University of Hawaii, Honolulu, Hawaii.

Several experiments were conducted to provide *(a)* information about different methodologies that could be used to assess the quality of translated technical English and *(b)* data on factors that affect the quality of text translated from English to Vietnamese. The three assessment techniques examined were back translation, knowledge testing, and performance testing.

## TECHNIQUES

### *Back Translation*

One method for evaluating translation quality is back translation—specifically, comparing the original English and the back-translated English. In the back-translation technique, the investigator asks one bilingual to translate from the original, to . the target-language, and then he asks another bilingual to translate back from the target to the original. The advantage of the technique is that, as opposed to other methods that have been suggested (e.g,, Carroll, 1966; Miller & Beebe-Center, 1956), the translation evaluator does not have to understand or speak the target language. A weakness is the fact that any mistakes in the back translation may be due either to the translator or to the back translator. Thus, even though we evaluate back translation to obtain insights about translation, a perfect translation can be misinterpreted by an incompetent back trans-

lator, or a good back translator can "correct" a poor translation. This is why back translation should always be complemented by other techniques, such as knowledge testing.

### *Knowledge Testing*

Knowledge testing refers to a method of evaluating translation quality in which subjects read a translated passage and then answer a set of questions about the content of the passage. If subjects can answer all the questions, the translation is assumed to be a good one. While the knowledge-testing technique resembles the standard reading comprehension method, it differs in one important respect: Measures of reading comprehension contain items of graded difficulty and are sensitive to individual differences. Knowledge testing is designed to elicit perfect scores if the translation is good and should be independent of individual differences. The technique was suggested by Miller and Beebe-Center (1956) and by Macnamara (1967) and was first used by Brislin (1970).

This approach asks, "How well can people read and understand Vietnamese that has been translated from English?" The knowledge-testing technique requires the researcher to write a series of questions in English about a passage and then to have them translated. He must also secure subjects who will read the passage and answer the set of questions. Tests must be scored by readers of Vietnamese, too, if they employ till-in type items. A multiple-choice format obviates the need for a native .reader.

### *Performance Testing*

This technique has subjects perform a task requiring them to use either English or trans lated instructions. To the extent that subjects can complete the task, the translation is regarded as equivalent to the original English text. As in the evaluation techniques previously described, the experimenter does not have to know the target language, since he only has to assess the product of the translated perform ance instructions.

Performance tests can be scored objectively. In the present experiment, a very demanding 12-step adjustment task on a portion of a helicopter engine made up the performance test. Three-man crews worked together, and the nature of the task required them to follow written instructions with care. Each of the 12 steps was assessed by a technically qualified observer as "error free," "minor error," or "major error."

Performance testing is the most. stringent translation evaluation technique, since it demonstrates the quality of a translation by observable behavior of subjects. However, the technique is the most expensive and time consuming of the three we have used because the experimenter has to *(a)* define a suitable task, (6) have it translated, *(c)* provide materials, for example, a helicopter, (*d*) secure suitably trained subjects, *(e)* have the subjects perform the task, and (*f*) obtain the services of observers who are technically competent to grade the task.

## METHOD

### *Bilingual Consultant*

A highly skilled consultant was hired who possessed the following qualifications: Vietnamese native, university teacher in Vietnam, 20 years in the United States, doctoral degree in educational psychology with additional training in linguistics, experience with translating technical materials, and had taught other Vietnamese how to translate.

### *Translators*

A group of t2 bilinguals was hired to provide translation services. At the time of these experiments, 7 of the 12 bilinguals were professional translators. All 12 had worked either part lime or full time as translators for an average of 11 years and had translated some technical materials in the past. None. however, had ever translated technical materials as a full-time job.

### *Materials to be Translated*

The 12 bilinguals translated three samples of technical material. The first was a section of the technical manual of the UH-1H helicopter (TM 55-1520-210-20). The second was a set of job performance aids for the C-141A aircraft. More specifically, we used PIMO (Presentation of Information for Maintenance and Operation). These materials have been designed so as to be more understandable than conventional technical manuals. The new format incorporates the following characteristics: organisation of tasks based on experimental analysis, a fixed syntax, a standardized verb list, and pictures corresponding closely to printed instructions (Goff, Schlesinger, & Parlog, 1969). The third type of material was the U. S. Air Force's

technical order for the C-141A aircraft (T.O. 1C-141A-2-12). This was chosen so that conventional and job jperformance aid materials for the same task could be compared.

An example of this material, from Chapter 7 of the UH-1H helicopter manual, *is as* follows:

> 7.2. This chapter provides all the instructions and information necessary for maintenance authorized to be performed by organizational maintenance activities on the power train system. The power train *is* a system of shafts and gear boxes through which the engine drives main rotor, tail rotor, and accessories such as DC generator and hydraulic pump. The system consists of a main drive shaft, a main transmission which includes input and output drives and the main rotor mast, and a series of drive shafts with two gear boxes through which the tail rotor is driven.

Other examples of technical materials translated by the bilinguals can be found in other sections of this article.

### *Translation Tasks*

All 12 bilinguals translated and back translated the three types of technical materials described above for eight hours on 2 different days. For instance, one bilingual would translate on the first day, and another • would back translate the first bilingual's work on the second day. All 12 bilinguals worked in quiet rooms and had access to an English dictionary *(Webster's Seventh New Collegiate Dictionary).* The instructions to the subjects were similar to those used by Brislin (1970).

### *Quality Measured by Back Translation*

The efforts of the 1.2 bilinguals produced 9,558 words of back-translated English, distributed as follows: 2,400 words of the UH-1H technical order, 3,486 words of the C-141A PIMO aids, and 3,672 words of the C-141A technical order.

Every word of the back-translated English was compared to the original, as in the following example: *original - English*—Man A performs activity (a test) in flight station; *back-translated English*—Mechanic A carries out the testing while in flight. In this example, the only combination of words that caused an error in the meaning of the back translation as compared to the original English is the substitution of "while in flight" for "flight station." All other words are judged to be equivalent.

The criterion for an error was simply this: Any place in the back translation that is not judged to convey the same meaning as the original English is called a meaning error. Meaning errors could be of six types:

1. An addition—an additional word or phrase appears in the back translation.
2. Minor omission—one or two words from the original arc omitted from the back translation.
3. Major omission—same as 2, but involving three or more words.
4. Garbling—three or more words in the back translation are not understandable.
5. Minor substitution- one or two words from the original do not have an equivalent in the back transla tion, but a phrase replaces the original words (e.g., "flight station" is back translated as "in flight").
6. Major substitution--same as 5, but involving three or more words. Finally, the back translation could be equivalent to the original and marked "O.K."

Our error analysis does not say anything about the operational seriousness of an error. We do not know, for example, whether a substitution error or addition of words would result in poor maintenance to the extent that a helicopter would operate in an unsafe condition.

### *Specific Method of Comparison*

Each of the three types of technical materials (described in Table 1) was arbitrarily divided into phrases averaging from eight to nine words. All phrases either were a complete sentence or contained a complete thought.

Dividing into phrases made it easy to look at a meaningful unit in the original and to find the equivalence or nonequivalence of that unit in the back translation. A given phrase could have more than one error. Each phrase, then, was tallied into one or more of the six error categories, or the "O.K." category. In addition, the exact wording that caused each error was noted.

Since the back translations of all three types of technical materials were examined, comparisons among their error scores can be made. This is possible since either all 12 bilinguals translated and back translated the material (as in the UH-1H technical order) or the 12 bilinguals were randomly assigned to translate or back translate the material (as in the C-141A PIMO aids and C-141A technical order). Thus, the quality of the people involved in work on the three types of material should be equivalent, and any differences should be due to the nature of each type of material. The main back-translation measure was simply a count of the number of meaning errors per passage. A second measure was derived by subdividing the total number of errors into the six categories.

### *Quality Measured by Knowledge Testing*

Two knowledge-testing experiments were run, each using different subjects and materials. In the first experiment, three translations of the same material from the Army's technical manual for the UH-1H helicopter were chosen that were judged to be of different quality. The quality ranking was based on the number of errors in the back translation; that is, Translation *A* had fewer back-translation errors than Translation B and Translation B had fewer errors than Translation C. In addition, a Vietnamese linguist read the original English and the three translations and then rank ordered the translations from best to worst. His rank ordering was the same as that based on the number of back-translation errors.

The knowledge test consisted of 10 fill-in type questions translated into Vietnamese. The same 10 questions were to be answered after the subject read one of the three translations. Since the questions were the same, any differences in the number answered would be due to the quality of the translations.

Subjects were 68 Vietnamese Air Force enlisted men being trained in helicopter maintenance at Fort Eustis, Virginia. These 68 subjects were randomly assigned to read either translation A ($n$ = 22 men), B ($n$= 23 men), or C ($n$ = 23 men). Subjects worked in an "open book" mode so that memory was not a factor on this test. An example of a question written about the previously quoted technical passage would be, "Who performs the maintenance on the power train system?" The correct answer is "organizational maintenance."

The second experiment was designed to compare translations of PIMO aids with those for the conventional U. S. Air Force technical order for the C-141A aircraft. A single bilingual translated both the PIMO aids and the technical order. He alternated between sections of one document and the other, so that he would not translate one document better simply because he had practiced on the other.

The questions to be asked about the passages were translated into Vietnamese by the same bilingual. Six of the questions were the same for the technical order and PIMO material, since the same topic was covered in the passages under study. These six questions allowed a range of 0-21 points. The other questions, also representing 21 points, were different for the PIMO and technical order, that is, they were unique to each passage. The "different" questions were added to increase the range of scores. An individual could thus achieve a score of 0-42. The major comparison between the PIMO and technical order would be in the "same" questions, since the same bilingual translated all test materials. Any difference in scores would be in the nature of the PIMO aids or the technical order.

Subjects were 36 Vietnamese Air Force enlisted men being trained in helicopter maintenance at Fort Eustis, Virginia. They read either PIMO or technical order material, and thus there were 18 subjects in a group. These subjects also worked in an "open book" mode. All tests, in both experiments, were scored by a Vietnamese linguist.

### *Quality Measured by Performance Testing*

Although it is a much more expensive and time-consuming approach to evaluating translations, the technique of observing men work with translated material comes closer to an ultimate criterion of the value of translations than any other method: Men do a task that is dependent on written material, and their performance is objectively scored. Good performance means that the writing was accurate and vice versa. In our experiments, teams of technicians carried out a very demanding adjustment task on a portion of the UH-1H helicopter main power plant.[4] Observers, U. S. Army sergeants who were both experts in helicopter maintenance and instructors on the system to be adjusted, assessed each of 12 steps in the task as "error free," "minor performance error," or "major error." Minor errors were those steps that the crews did wrong but then corrected, major errors were noted if crews could not proceed or if their performance was so poor that it required intervention by the observers.

There were four experimental language conditions: (*a*) the standard or original English technical manual, *(b)* a very high-quality translation, and (c*)* and *(d),* two lesser grades of translation. The high-quality translation was produced as follows: Two of our best translators each worked independently, then they reviewed each other's work and wrote a "consensus" translation. Finally, our linguist consultant reviewed and modified their combined effort. The translators had available two bilingual glossaries of technical terms. (We refer to this translation as "supervised.")

The first of the lesser quality translations was done by a free-lance, highly qualified translator to whom we gave copies of the same technical glossaries mentioned above. This man worked without review. (We call this the free-lance translation.) The second of the lesser quality translations was obtained by contracting with a Washington, D.C. translation service company for a fixed fee to have the approximately 1,000 words of English translated. We had no control of the method used by the translator nor did he have access to any of our glossaries or other aids. His work also was not reviewed. (We call this the commercial translation.) It is important to note that both the free-lance and commercial translators were highly qualified translators.

Crews used as subjects were assembled from two groups of men at the U. S. Army's Transportation School, Fort Eustis, Virginia: *(a)* Vietnamese airmen who had just completed the Army's aircraft maintenance and helicopter repairman course and *(b)* U. S. Army enlisted technicians who were also newly graduated from the same courses. Vietnamese airmen were assigned randomly to one of the four language conditions. In each language condition shown, there were six three-man crews, each of which worked independently. The American Army technicians who used English were tested for comparison purposes.

Only indirect comparisons between the three translation assessment methods can be made since practical considerations made it impossible to test the same materials with the three methods. Brislin (1970) was able to furnish comparative information in an earlier study.

## RESULTS AND DISCUSSION

### *Back Translation*

Reliability of the back-translation examination technique was adequate. Two raters independently examined the 12 back translations for the Army technical manual material, and their ratings of number of errors per passage and types of errors were in close

[4] Section 5-391 "Adjustment—Power Turbine Governor RPM Controls," U. S. Army Technical Manual, TM-55-1520-210-20.

TABLE 1

TRANSLATIONS EVALUATED BY KNOWLEDGE TESTING: TWO SESSIONS

| Translation | No.subjects | Mean score | SD |
|---|---|---|---|
| Session 1: Comparison of three translations of UH-1H technical manual | | | |
| A | *22* | 6.1 | *2.2* |
| B | *23* | 4.3 | 1.8 |
| C | 23 | *2.6* | 1.3 |
| Session 2: Comparison of PIMO and technical order translations for C-141A | | | |
| PIMO | 18 | | |
| Total score | | 34.8 | 3.3 |
| Same questions | | 16.2 | 3.7 |
| Different questions | | 18.6 | 1.2 |
| Technical order | 18 | | |
| Total score | | 33.2 | 6.7 |
| Same questions | | 16.1 | 2.9 |
| Different questions | | 17.1 | 4.7 |

*Note*. PIMO = Presentation of information for maintenance and operation.

agreement: $r$ = .88 and $r$ = .94, respectively. A comparison of the three types of technical material, that is, Army technical manual, Air Force technical order, and job performance aids (PIMO), showed very few differences in types of errors that occurred among translations. The only statistically significant difference was in the proportion of "minor substitution" errors for the Army material (13%) versus both technical order and PIMO material (32% and 30%, respectively). More striking was the fact that for seven categories of error there was very close agreement for translations of all three kinds of material. (A more detailed statement of this error analysis appears in Sinaiko and Brislin, 1970.) The major yield from the back-translation analyses was insight into how the translators went about performing a very difficult task. For example, translators in our experiment did one of four things when they came across unfamiliar words in English or words for which there were no Vietnamese equivalents:

1. They left the English word intact in the translation.

2. They transliterated the word using Vietnamese characters.

3. They coined terms to describe in a functional way the English word or concept. For instance, the translators looked at the word "tachometer" (for which there is no Vietnamese equivalent) and then decided that this meant "rotation measuring device," which they could express. This transformation of difficult technical English to simpler English and then to Vietnamese is called "the explain-around technique" by the present investigators. Wickert (1957) noted that he experienced the same technique when he asked Vietnamese to translate abstract concepts.

*Knowledge Testing*

Table I gives the results of both knowledge-testing sessions. For Session 1, where a perfect score is 10, it can be seen that subjects were able to answer more questions about Translation A than B and more about B than C. This rank ordering is the same as that found by errors in the back translation and by the judgments of a Vietnamese linguist. Differences among all combinations of the three means (A versus B, A versus C, B versus C) are statistically significant ($p < .01$). These results show that the knowledge test is sensitive enough to demonstrate differences in translation quality.

For Session 2, the data toward the bottom of Table 1 show that the translation of the PIMO aids and the technical order for the C-141A allow the same number of both the same (perfect score is 21) and different (perfect score is 21) questions to be answered. Thus, the total number of questions (perfect score is 42) are also the same for the technical order and PIMO aids. The very small differences are not statistically significant ($p > .10$).

*Performance Testing*

Table 2 presents the performance results of Vietnamese mechanics working with an English or translated text as well as the results of the control group of U. S. Army mechanics who worked only with an English text. Several striking things about translated technical material are illustrated. First, it is clear that working in one's own language, even if that

material is a translation of a difficult technical manual, is significantly better than having to use a second language. The difference is significant by chi-square at less than the .01 level ($\chi^2 = 16.6$, ($df = 1$). However, an important qualification is that the translation must be of high quality. Second, the performance task is sensitive to the quality of translation: Commercial quality[5] produced much higher rates of serious errors than the English text. That is, the Vietnamese airmen worked more effectively with English than they did using a poor translation ($\chi^2 = 13.5$, $df = 2$, $p < .01$). Third, the quality of translated technical documents as measured by performance is significantly influenced by the procedures of the translators. Thus, using a group of men who were approximately equal in their bilingual abilities as translators, we were able to produce very different levels of material. The mode of compensation, that is, placing a premium on speed, was one procedural variable. The availability of bilingual glossaries of technical terms was another.

Incorporation of team translation and a review procedure seemed to make a difference. Finally, the careful translation procedures outlined here can lead to documents that allow Vietnamese mechanics to perform as well as U. S. Army mechanics. (Note, however, that the best Vietnamese groups committed some "major errors," i.e., about 5%, while the Americans did not.)

*Subjective Opinions and Translation Quality*

An interesting fact emerged from discussions with some of the Vietnamese airmen who used the best translated material. Most of the men we talked with after they had worked on the performance task expressed a dislike for the translations. The principal objection seemed to be that there were unfamiliar Vietnamese terms used for some of the technical English words. To paraphrase the words of some subjects, "...we did not understand all the Vietnamese words. We would prefer to use the English manual on which we had been trained." It is particularly noteworthy that, in spite of their expressed dislike of even the best quality translation, the measured performance of the airmen was nearly equal to that of the American technicians. At the same time, we asked two bilingual readers (one of whom was an expert in helicopter maintenance) to review and comment on one commercial translation. Each of these men thought that, the latter document was "pretty good." However, in practice it resulted in the worst performance of any of the language conditions. The point we wish to underscore is the discrepancy between subjective assessment and performance testing as ways of evaluating translations. The verbal reactions of our subjects and of the linguists were reversed when we actually measured performance.

[5] Data for one crew, commercial translation, were lost because that crew was unable to follow the translation. This supports our contention that this specific translation was poor.

TABLE *2*

PERFORMANCE TEST RESULTS: ACCURACY

| Experimental condition | % error free | % major errors committed |
|---|---|---|
| Vietnamese: Supervised translation | 73.1 | 5.6 |
| Vietnamese: Free-lance translation | 40.3 | 4.2 |
| Vietnamese: Commercial translation | 11.0 | 37.0 |
| English (VNAF-subjects) | 40.7 | 20.6 |
| English (U. S. Army subjects) | 73.2 | 0.0 |

*Recapitulation: Three Methods Compared*

The experiments reported in this study are based on three approaches to assess the quality of translation: *(a)* back translation, *(b)* knowledge testing, and *(c)* performance testing. None of the three methods described requires that the experimenter have proficiency in the target language, although each approach requires the services of linguist translators. Relatively greater demands are placed on translator services in the first two methods than the last; particularly in the use of knowledge testing, translators must be used for the basic English text, the questions to be answered, and as test scorers. Back translation puts an analytic burden on the experimenter that is not present for the other techniques. However, there are no test items to be developed for back translation, while such items are

TABLE 3

COMPARISON OF THREE TRANSLATION EVALUATION METHODS

| Characteristic | Back trans-Lation | Knowledge testing | Performance testing |
|---|---|---|---|
| Experimenter proficiency in target language | No | No | No |
| Translators needed | | | |
| Original text | Yes | Yes | Yes |
| Test items | No | Yes | No |
| Scoring tests | No | Yes* | No |
| Back translating | Yes | No | No |
| Test construction | No | Yes | Yes (but may use available task) |
| Technical experts as observers | No | No | Yes |
| Special equipment needed | No | No | Yes |
| Relative cost; face validity | Lowest | Middle | Highest |
| Confidence in results | Lowest | Middle | Highest |
| Test subjects | No | Yes — any reader of the language | Yes — must be trained in the task |

[a] If fill-in type items are used.

at the heart of knowledge tests. Performance testing may require that a task be designed, although, as in the present experiments, an available task was used. In addition to translators, test subjects are required for knowledge and performance testing; this is not so with back translation. *Only* in the case of performance tests are technical experts needed to evaluate what subjects do. Similarly, special equipment or material is needed for performance tests but not for the other two approaches (see Table 3*)*. The relative costs of the three methods are probably in this order (low to high): back translation, knowledge tests, and performance tests. Finally, confidence in results or face validity of the methods is likely in the same order.

## REFERENCES.

BRISLIN, R. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology,* 1970, 1, 185-216.

CARROLL, J. Quelques measures subjectives en psychologie: Fréquence des mots, significativité et qualité de traduction. *Bulletin de Psychologie,* 1966, 19, 580-592.

GOFF, J., SCHLESINGER, R., & PARLOG, J. *PIMO test summary.* (Tech. Rep. 69-155, Vol. II) Andrews Air Force Base, Md,: Space and Missile Systems Organization, Air Force Systems Command, May 1969.

MACNAMARA, J. The bilingual's linguistic performance —a psychological overview. *The Journal of Social Issues,l967, 23,* 58-77.

MILLER, G. A., & Beebe-Center, J. Some psychological methods for evaluating the quality of translation. *Mechanical Translation,* 1956, 3, 73-80.

SINAIKO, H. W., & BRISLIN, R. W. Experiments in language translation: Technical English-to-Vietnamese. (Research Paper P-634) Arlington, Va.: Institute for Defense Analyses, 1970.

SINAIKO, H. W., Guthrie, G. M., & Abbott, P. S. *Operating and maintaining complex military equipment: A study of training problems in the Republic of Vietnam.* (Research Rep. P-501) Arlington, Va.: Institute for Defense Analyses, 1969.

WICKERT, F. An adventure in psychological testing abroad. *American Psychologist,* 1957, 5, 86-88.

(Received December 14, 1971)