Item Response Theory and Language Translation

An Aid to Comparative Studies of Management in the United States and Japan

WATANABE Naotaka

IN LINE WITH THE RECENT TREND toward the globalization of business activities, a lot of cross-cultural management research has been conducted in recent decades. In particular, a huge amount of comparative research on management in the United States and Japan has accumulated since Japanese economic success was recognized as having quite a significant influence on the world economy. The focus of these studies has been concentrated on various facets of such business activities as personnel management, production management, financial management, and marketing management.

Comparative studies of personnel management and organizational behavior in the two countries have been conducted in great numbers. According to the extensive review of the literature done by Peng, Peterson, and Shyi (1991), from 1981 to 1987 6.0 percent of a total of 8,403 management-related articles were cross-cultural organizational studies, while in comparative studies the most frequent was a comparison of Japan with the United States.

Despite the extensive academic interest in cross-cultural studies, the studies themselves contain a lot of problems that must be resolved. In all cross-cultural research we inevitably encounter problems of culture, language, customs, and other national differences. Many researchers have used quantitative methods involving a questionnaire to collect data from individuals who are working under "similar" organizational conditions and then have analyzed the data using multivariate analyses, such as multiple regression and other correlational methods. These methods strictly require "culturally equivalent data collecting conditions" as the prerequisite for statistical comparisons between the two countries.

WATANABE Naotaka is a member of the School of Business Administration at Nanzan University and a research associate in the Graduate School of International Development of Nagoya University. The author wishes to acknowledge the assistance provided by both universitites for carrying out this study, and to express gratitude to YAMADA Fumito, MIZOTA Yoshihiro, and KOBAYASHI Keiko for their help in analyzing the data.

In this context, language translation is one of the key factors determining whether the comparative research is meaningful or not, since almost all comparative surveys depend heavily on exact equivalence between the original language and the translated language. The purpose of this paper is to discuss the feasibility of item response theory (IRT) to attain language equivalence in cross-cultural management surveys. Discussed in particular is the theory and practice of analyzing cross-cultural quantitative data obtained by questionnaire surveys making use of IRT.

Problems of language translation

In order to achieve high-quality measurements in cross-cultural and comparative organizational research, we should resolve the serious problem of how we translate an original language to a target language. Language is generally considered to be a defining characteristic of culture. In cross-cultural organizational research, studying cultural similarity and difference is always a main goal of the research. In a questionnaire survey, a tool frequently used in cross-cultural research, cultural similarity and difference is operationally defined by response patterns to questions written in different languages. This means the translation of a questionnaire from one language to another always involves the risk of producing the problem of measurement inequivalence.

The problems of language equivalence in cross-cultural surveys have been addressed in quite a few discussions, and several linguistic and psychometric techniques have been developed. Casagrande (1954) argues that language translation can be categorized according to the linguistic purposes of the translation. He states that there are four types of translation, namely, pragmatic, aesthetic-poetic, ethnographic, and linguistic.

Brislin (1976) summarizes the major differences among these translations. Pragmatic translation emphasizes the accuracy of the content levels of the message. Here, the context of the message, the style, and grammatical forms are relatively ignored. Translation into Japanese of English commands for computer software would be an example of this kind of translation. The translation would be evaluated as successful if Japanese software users can use the commands without serious problems.

Aesthetic-poetic translation has a different goal. In this type of translation, the extent to which emotional states contained in the message (such as mood, affect, feeling, and atmosphere) are accurately translated is crucial. Translations of psychological tests to measure emotional aspects of the respondents, such as depression, anxiety, irritation, and so on, would be examples of the aesthetic-poetic translation.

Ethnographic translation refers to translations of a message in which the meaning and cultural content of the original language are maintained in the target language. To attain this goal, the translators have to be familiar with both the original and the target languages and cultures. For example, the meaning of commitment to the organization is totally different for Americans and Japanese. In the cultural context in the United States, the concept of commitment is based upon a kind of social contract between an individual and the company. On the other hand, in the Japanese cultural context, commitment contains a strong sense of loyalty toward the company, in which the individual has less autonomy and power than the organization has. Therefore, ethnographic translation is needed when we translate culturally biased concepts into the another language.

Linguistic translation refers to a translation whose end is to provide an equivalent structure and equivalent grammatical forms in the two languages. The goal of language translation by artificial intelligence in early days was to attain this type of translation.

Language translation becomes meaningless unless there is a careful examination of the purpose of the document. In cross-cultural research particularly, before doing a translation of the scales and/or items included in the questionnaire, we need to find out what the purpose of the research is, what kind of information is desired through the survey, and which aspects of both cultures we want to compare. Then we should select the type of translation that is suitable for attaining the goal desired.

Techniques for Assessing Language Translation

There are two groups of techniques for assessing the quality of a translation. One group is that of qualitative methods, the other is that of quantitative methods. I shall explain each.

Qualitative methods often include in the process of translation a comparison of the original items with the translated items. The back-translation technique and the decentering technique, often recommended by most cross-cultural researchers, are classified as qualitative methods. In back-translation, an item is translated into a target language and then translated back into the original language by someone who was not engaged in the first step of the translation. The degree of congruence between the original and back-translated versions is regarded as an indicator of the quality of the translation. In order to attain a translation of even higher quality, Brislin (1970) proposes an iterative procedure of back-translation. The decentering technique is a method that aims at ethnographic translation. In this technique, the original meaning and the cultural content included in the original item are preserved as far as possible. Phrases and expressions that have a common meaning in both cultures would be adopted, even if the literal meanings were different.

Needless to say, bilingual persons and/or highly skilled translators would play crucial roles in qualitative methods of translation. It is sometimes very difficult, however, to find people who are familiar with the language and culture *and* an academic field as specific as personnel management. Furthermore, language translation by qualitative methods does not ensure precise equivalence. Hence, some researchers regard the qualitative methods, such as back-translation, as merely a minimum requirement of language translation (Hulin,1987; Ellis,1989).

The quantitative methods include several methods that psychometrically assess the equivalence of translation. Unlike the qualitative methods, quantitative methods are usually used after data analysis is done. They try to check on whether or not items have similar patterns of response in each culture group.

The simplest approach to assessing equivalence is to compare the proportion of endorsing (positive responses) to the items written in both languages, or to compare the rank order of the proportion of endorsement across two subpopulations. This approach is so simple that we should not need to use sophisticated inferential statistics to examine the equivalence. But this approach has in it the serious problem that the statistics (the proportion of endorsement) are influenced by the distributions of strength of attitude across groups.

A second approach is to use the score of item-total correlations to indicate the extent of equivalence of translation. In this method, an item-total correlation computed in each subgroup is compared by statistical tests of differences. The logic of this method is that item-total correlations reflect the original item's and translated item's discriminatory power, in the sense that respondents with high test scores are more likely to endorse an item than respondents with lower test scores. But this approach is also largely influenced by the distributions of strength of attitude between two groups.

A third approach is to conduct factor analyses of the item responses in each subpopulation. Patterns of factor loadings are compared among groups. If the factor structure does not show similar loadings on items across two subpopulations, the conclusion would be that translation equivalence is not maintained. The problem with this approach is that it needs very sophisticated and professional skills to make the factor structures congruent among the groups.

A fourth approach is the use of chi-square statistics. In this method, the range of total scores on the test obtained from two subpopulations is divided into several discrete intervals. Chi-square figures are calculated to get expected frequencies of both endorsed and non-endorsed responses by the two groups in each score interval. If the translation is equivalent, the probabilities of an endorsed response for subjects who fall within a particular test-score interval should be the same between two subpopulations. This approach is based on the assumption that the test is measuring a unidimensional latent trait. Therefore, it is absolutely neceessary to examine the unidimensionality of the original and translated versions of the scale before this method is used.

A fifth approach is to use the statistics of item response theory (IRT), a new theory of measurement, for detecting equivalence of translation. IRT is a model-based approach to psychological measurement. The original purpose of IRT was to infer the statistical relationship between individuals' response patterns toward test items and an assumed latent trait. When it is applied to language translation problems, it is assumed that two groups of individuals who are equal in the latent trait measured by tests that maintain equivalence of translation have the same observed score. As will be shown later, the method of assessing translation equivalence by IRT uses more precise statistical tools than the other quantitative methods do.

An Overview of Item Response Theory

IRT supposes that an individual's response on a scale can be accounted for by defining characteristics called the latent trait (Lord & Novick, 1968). In IRT, the relationship between the unobserved latent trait and observed score is described in terms of a mathematical model.

IRT does not assume some typical distribution of the latent trait, such as normal distribution. Most IRT models also assume that the space of the latent trait is unidimensional. It means that it is assumed there exists only one trait that completely accounts for the performance of the test. Another assumption of IRT models is that of local independence, which is derived from the assumption of the unidimensionality of the latent trait. Local independence includes the notion that responses to different items by a given respondent over replications are independent. It means that responses to an item are fully explained by the traits being measured.

Mathematical functions of IRT models relate the probability of response occurring to an item, to the trait measured by the item. These mathematical functions are referred to as item characteristic curve(s), ICC(s). An ICC represents the regression of item scores on the latent trait, so that, given an ICC for an item, the probability of a particular response for any given latent trait level can be assessed. The property of ICCs can be obtained through the statistical estimation of the parameters that define the IRT models.

Many different IRT models have been proposed for a variety of different applications. The most common and widely used models are the binary response models, which include the one-parameter logistic, or Rasch model, the two-parameter logistic model, and the three-parameter logistic model. Figure 1 shows an ICC of the three-parameter logistic model. The horizontal axis indicates the strength of the latent trait and the vertical axis indicates the probability of a correct or endorsing response.

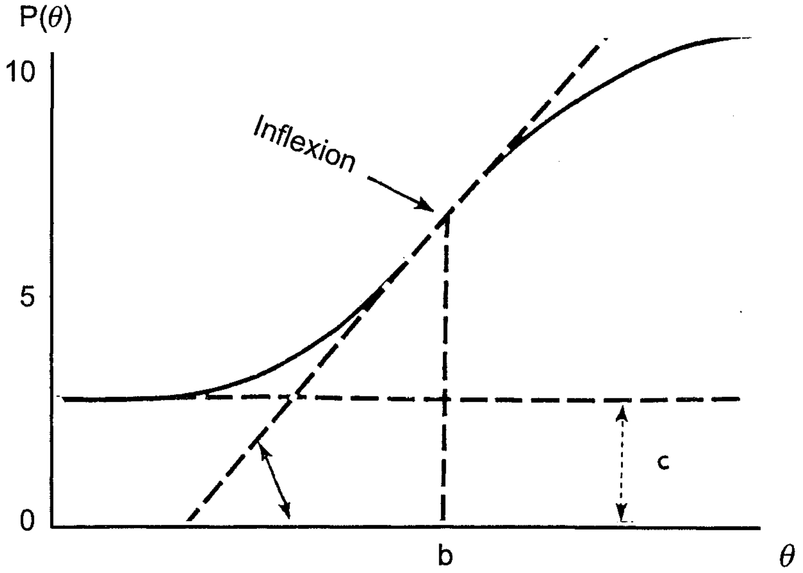

Fig. 1 The Relationships between ICC and the Parameters.

The equation of this model is as follows:

$$P_i(\theta) = c_i + (1-c_i)\ \frac{1}{[1 + exp\{-Da_i(\theta-b_i)\}]}$$

where: $P_i(\theta)$ is the probability of a positive response to the ith item among respondents with a score of θ on the latent trait assessed by the items on the scale; b_i, referring to item popularity parameter in attitude measurement and item difficulty parameter in ability measurement, controls the location of the ICC along the latent trait θ continuum; a_i refers to discrimination parameter and it controls the steepness of the ICC; c_i refers to guessing parameter, which indicates the lower asymptote of the ICC, and is used to model items where respondents with low θs sometimes respond positively; D is a scaling constant usually set equal to 1.702; and θ refers to the latent trait assessed by the group of items that constitute the scale.

Thus, the three-parameter logistic model provides *a, b,* and *c* parameters. On the other hand, the two-parameter model provides *a* and *b* parameters, while the one-parameter model only deals with the *b* parameter.

The attractiveness of IRT is derived primarily from its parameter invariance properties concerning item and person statistics. This property refers to the fact that item statistics estimated from the application of IRT models are independent of the sample of respondents. Likewise, the person statistics are independent of the items included in the test. This is true in theory. In practice, however, the estimated values of the parameters would differ from sample to sample if the samples differ in mean value or variability on the latent trait. The degree to which the property of parameter invariance is true is the degree to which the sample of respondents or items determines the parameter scale.

There are several methods for estimating IRT model parameters (Baker, 1992). Maximum likelihood estimation is one of the most widely used method. This method requires a relatively large number of respondents and items for accurate estimation. The marginal maximum likelihood estimation (Bock & Aitkin, 1981) approach can provide accurate item parameter estimates even with a relatively small number of respondents and items. Various kinds of Bayesian estimation are also used for parameter estimation. The accuracy of Bayesian estimates is maintained when some item parameters are known.

Several computer programs to assist the parameter estimation process are available. The LOGIST program is probably the most common one. This program simultaneously estimates the item and person parameters of the one-, two-, and three-parameter logistic models by using joint maximum likelihood estimation. The BILOG program computes marginal maximum likelihood estimates of the item parameters of the one-, two-, and three-parameter logistic models. After the item parameters have been estimated, the value of latent trait θ of each respondent will be obtained at the request of the user.

Detecting Translation Equivalence by IRT

As mentioned earlier, parameter invariance is one of the main attractive features of IRT models. It means that the values for IRT item and person parameters do not depend on the sample examinees nor on the sample of items. By utilizing this property of IRT, we can examine the equivalence between an original item and the translated item. The logic underlying this method is that if ICCs for the item estimated in samples from different subpopulations, which have different mother tongues, are not congruent within the limits of sampling fluctuations, the item can be regarded as inequivalent.

This notion of detecting the equivalence between an original and translated item is based on the logic and notion related to the detection of item bias, which is recently called differential item functioning (DIF). Item bias can be defined in terms of IRT. Since the probability of endorsing or giving a positive response is given by an ICC, an item can be regarded as unbiased if the ICCs across different subgroups are identical. This means ICCs must be identical, apart from sampling error, across different populations of interest.

Several methods of detecting the item bias have been developed. According to Ironson's (1983) review of procedures based on IRT for assessing the item bias, the major procedures fall into the following three categories: (1) comparison of ICCs; (2) comparison of vectors of item parameters; and (3) comparison of the fit of the item response models to the data.

The procedure for comparing ICCs is, first the item and latent trait parameters are estimated separately for the two groups, then the latent trait scale is divided into several small intervals, and, finally, the differences between areas defined by the height and the small intervals of latent trait are found (Runder, 1977).

The method for comparing vectors of item parameters conducts a simultaneous comparison of the item parameters of two groups by using standard multivariate techniques. Lord (1980) proposed a procedure for detecting item bias based on a large sample version of Hotelling's T statistics.

The procedure for comparing the fit of the item response models to the data is, first, the items and latent trait parameters are estimated for the aggregated data from two groups, then the probability of positive response for each person, the average probability of positive response for each group, and the proportion of positive response in a group (the classical item difficulty index) are computed, and, finally, the value of the average probability of each group is compared with the proportion of positive response in a group. In addition, the standard residual is computed for each person (Linn & Harnisch, 1981).

By using these procedures, item bias can be detected more meaningfully and accurately. Hambleton & Swaminathan (1985) recommend the first "area" method and the second parameter comparison method. Although these two methods are logically equivalent in that ICCs are compared, we might get incongruent results regarding the bias of an item due to the difference between the operational definitions. Preferably, both procedures should bé used simultaneously for accumulating evidence in terms of item bias.

Detecting Equivalence of Translation on English-Japanese Versions of the JDI

The author conducted an analysis of a translation equivalence between English and Japanese versions of the Job Descriptive Index (JDI). The JDI has long been used as a measure of job satisfaction in the United States. Since Smith, Kendall, & Hulin (1969) have developed the JDI, it has provided researchers and practitioners with a set of scales measuring job and work-role affects in a wide variety of settings.

The JDI covers five facets of job satisfaction: Work, Pay, Promotion, Supervision, and Coworkers. Adjectives or brief phrases are presented; the respondent indicates whether or not each describes his or her job by choosing "Yes", "?", or "No" alternatives.

Because of the simplicity of item expression and required response, the JDI readily lends itself to translation into different languages. As a matter of fact, the JDI is already translated into Spanish (Hulin, Drasgow, & Komokar, 1982) and Hebrew (Hulin & Mayer, 1986) and the equivalence is detected by using IRT.

PROCEDURES

For this study, the "Work" subscale of the JDI, including 18 adjectives and phrases, was chosen. Firstly, all 18 items were translated into Japanese and then translated back into English. The original expression and the expression translated back was compared. As a result, it became clear among the assessors that "hot" and "on your feet" in the original version had no adequate and universal word in Japanese. Therefore, these two items were eliminated from the JDI.

The 16 items in the JDI of both the English and the Japanese versions were administered to groups of American and Japanese workers. They were all blue-collar employees working in a Japanese-owned automobile manufacturing plant in the United States. The Japanese employees were basically trainers and the Americans trainees working in the same workplace. The numbers of Japanese and American respondents were 92 and 72, respectively. The respondents were required to answer whether or not their "Work" had the characteristics shown by each expressed item by circling "Yes", "?", or "No".

The scoring method was to give one point for a "Yes" response to a positive item and for a "No" response to a negative item, and to give zero points for the other response patterns, including the "?" response.

The data analyses were done as follows:

1. Factor analyses were conducted for both versions of the JDI to examine the unidimensionality of the scale.
2. A two-parameter logistic model was adopted and the values of item parameters and latent trait were estimated by maximum likelihood estimation on the LOGIST 5 program.
3. For detecting the equivalence of the translation, a comparison of ICCs between English and Japanese items was conducted by the chi-square test proposed by Lord (1980).

RESULTS

The results of factor analyses showed that the English and Japanese versions of the JDI does not necessarily maintain unidimensionality by the Kaiser-Guttman criteria, an eigen value greater or equal to 1.0. But, the results of a scree test showed that the first factors have a large amount of contribution. The first factor explained 30.4 percent and 29.6 percent of the total variance of the English and Japanese versions of the JDI, respectively.

Estimation of the values of item parameters and latent trait through LOGIST 5 was done by using an iterative procedure since we did not have any information about these parameters beforehand. During the process of estimation, the 16th item of the English version, "gives sense of accomplishment", was eliminated, since all the responses to this item were found to be "Yes," and it was impossible to estimate the parameters in terms of this item.

Table 1 shows the result of parameter estimation in terms of the remaining 15 items. Although the Japanese version successfully converged the parameter estimation for both *a* and *b* parameters through the iterative procedure, the English version could not do so for a parameter. The

English version's maximum values for a parameter was 2.0, which is the computation default in the LOGIST 5 program. This result might indicate that the number of American respondents was too small to converge the parameter estimation and that the work situation evaluated by the American respondents was so specific that the JDI did not function as a good stimulus to yield general responses from them.

Table 1

Estimated Item Parameters for English and Japanese Versions of the JDI

		English			Japanese			
	Item	*p*	*a*	*b*	*p*	*a*	*b*	χ^2
1	Fascinating	89	1.07	1.76	.82	1.88	-0.95	.70
2	Routine	.56	1.96	-0.10	.50	1.04	.18	1.11
3	Satisfying	.97	1.17	-2.51	.86	2.95	-1.00	3.47
4	Boring	.76	2.00	-0.61	.81	0.77	-1.29	1.70
5	Good	.98	2.00	-2.83	.94	1.82	-1.71	.23
6	Creative	.91	1.70	-1.39	.85	2.29	-1.04	.41
7	Respected	.97	.13	-15.26	.91	1.85	-1.33	15.43**
8	Pleasant	.97	2.00	-2.53	.80	2.95	-0.78	.94
9	Useful	.98	2.00	-2.83	.95	1.70	-1.87	.29
10	Tiresome	.32	2.00	.54	.14	1.45	1.75	.69
11	Healthful	.80	.69	-1.49	.71	1.27	-0.59	.35
12	Challenging	.94	2.00	-1.70	.94	2.34	-1.58	.14
13	Frustrating	.25	1.29	1.35	.25	2.95	1.05	3.76
14	Simple	.73	1.21	-0.72	.65	1.90	-0.37	.49
15	Endless	.22	2.00	1.42	.11	2.95	1.82	1.08

Note: *a* = discrimination parameter;
b = difficulty parameter;
p = proportion of "Yes" response;
** p<.01 (df=2)

Table 1 also shows the results of statistical testing concerning the equivalence of the items translated. As the results of chi-square tests indicate, it was found that item 7 ("Respected") did not maintain equivalence of translation. This might be explained in two ways: (1) "Respected," translated into Japanese as "Sonkei-sareru", has culturally different meanings in the United States and in Japan; (2) estimation of the parameters themselves was not successful due to technical data-handling problems.

The first reason could be justified as follows. For American employees, "Respected" means that "my work is respected by myself." On the other hand, for Japanese employees, it means that "my work is respected by other people." It is generally said that American people tend to choose their work in the light of their own responsibilities, they tend to work in order to pursue individual interests. Japanese, however, tend to choose their work by considering the entire social situation in which they find themselves. They tend to work not to attain individual goals but to fit in nicely in the work situation and maintain harmony with other people. These culturally differentiated attitudes toward work could have accounted for this result.

The second reason could be explained by saying that unidimensionality of scale was not fully maintained and that maximum likelihood estimation was not adequate for parameter estimation of the small sample. In future, handling the data in such a way as to maintain unidimensionality and attempting parameters estimation by different methods, such as marginal maximum likelihood estimation, would be necessary to obtain satisfactory results.

References

Baker, F. B. (1992) *Item Response Theory: Parameter Estimation Technique*, New York: Marcel Dekker.

Bock, R. D. and Aitkin, M. (1981) "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," *Psychometrika*, 46, 443–59.

Brislin, R. (1970) "Back-Translation for Cross-Cultural Research," *Journal of Cross-Cultural Psychology*, 1, 185–216.

Brislin, R. (1976) "Translation Research and Its Applications: An Introduction." In R. Brislin, ed., *Translation: Applications and Research*, New York: Wiley.

Casagrande, J. (1954) "The Ends of Translation," *International Journal of American Linguistics*, 20, 335–40.

Ellis, B. B. (1989) "Differential Item Functioning: Implications for Test Translations," *Journal of Applied Psychology*, 74, 912–21.

Hambleton, R. K. and Swaminathan, H. (1985) *Item Response Theory: Principles and Applications*, Boston: Kluwer-Nijhoff.

Hulin, C. L. (1987) "A Psychometric Theory of Evaluations of Item and Scale Translations: Fidelity across Languages," *Journal of Cross-Cultural Psychology*, 18, 115–42.

Hulin, C. L., Drasgow, F., and Komokar, J. (1982) "Applications of Item Response Theory to Analysis of Attitude Scale Translations," *Journal of Applied Psychology*, 67, 818–25.

Hulin, C. L. and Mayer, L. J. (1986) "Psychometric Equivalence of a Translation of the Job Descriptive Index into Hebrew," *Journal of Applied Psychology*, 71, 83–94.

Ironson, G. H. (1983) "Using Item Response Theory to Measure Bias." In R. K. Hambleton, ed., *Applications of Item Response Theory*, Vancouver, BC: Educational Research Institute of British Columbia.

Linn, R. L. and Harnisch, D. L. (1981) "Interactions between Item Content and Group Membership on Achievement Test Items," *Journal of Educational Measurement,* 18, 109–18.

Lord, F. M. and Novick, M. R. (1968) *Statistical Theories of Mental Test Scores,* Reading, MA: Addison-Wesley.

Lord, F. M. (1980) *Applications of Item Response Theory to Practical Testing Problems,* Hillsdale, NJ: Erlbaum.

Peng, T. K., Peterson, M. F., and Shyi, Y. (1991) "Quantitative Methods in Cross-National Management Research: Trends and Equivalence Issues," *Journal of Organizational Behavior,* 12, 87–107.

Runder, L. M. (1977) "An Approach to Biased Item Identification Using Latent Trait Measurement Theory." Paper presented at AERA meeting.

Smith, P. C., Kendall, L. M., and Hulin, C. L. (1969) *The Measurement of Satisfaction in Work and Retirement,* Chicago: Rand McNally.