Jurnal Pengurusan 11 (1992) 15-24

Reliability Assessment for Cross-cultural Measurement Comparability

Nik Rahimah Nik Yacob
Ismail Rejab

ABSTRACT

The problem with measurement comparability in cross-cultural studies arises when researchers attempt to attribute differences in obtained scores to cultural influences. It is a common approach among researchers in marketing to use a standardised instrument cross-culturally to ensure comparability. The author suggests the performance of reliability assessments as a basis for determining the appropriateness of using standardised instruments although the evidence of comparability is inconclusive.

ABSTRAK

Masalah perbandingan pengukuran dalam kajian antara budaya timbul apabila penyelidik menganggap bahawa perbezaan dalam skor yang diperhatikan adalah berpunca dari pengaruh budaya. Satu pendekatan yang lazim dilaksanakan oleh penyelidik-penyelidik pemasaran bagi kajian antara budaya adalah dengan menggunakan instrumen pengukuran standard. Pengarang menyarankan penggunaan analisis reliabiliti bagi menentukan kesesuaian penggunaan instrumen pengukuran standard walaupun penilaian tersebut tidak memberi bukti-bukti yang muktamad bahawa hasil pengukuran adalah sama antara budaya.

INTRODUCTION

In most cross-cultural studies in marketing, the issue of measurement comparability between cultures is largely ignored. The neglect is often critical because the findings of similarities and differences between cultures can be an artifact of construct measurement rather than actual similarities and differences. Cross-cultural studies are especially prone to the challenge of measurement incomparability because they involve different cultural context and the use of different languages (Henry 1961; Verba 1971; Winter & Prohaska 1983).

The social science literature is replete with prescribed approaches for assuring measurement comparability. However, the most common approach taken by marketers (e.g. Narayana 1981; Green, Cunningham & Cunningham 1973; Permut

1977; Nagashima 1970) in cross-cultural surveys involves the use of a standardised instrument which is translated and back translated to ensure fidelity between the original and the translated instrument. It is the purpose of this paper to evaluate the viability of such an approach by assessing the reliability of an instrument which had been administered to subjects from two different cultures, American and Malaysian. The Arousal Seeking Tendency instrument by Mehrabian and Russell (1974) for measuring the construct, optimum stimulation level, was selected because it had been demonstrated to have high reliability and validity among American subjects (Raju 1980).

MEASUREMENT COMPARABILITY

Zelditch (1971: 273) explains the term comparability by stating:

> Let 1, 2, 3, ... n be units in each of which the process phi takes place. Then 1 is comparable to 2, 3, . . . n if and only if (a) there exists a variable V common to each of them and (b) the meaning of V is the same for all of them.

Satisfying the first requirement is a relatively trivial problem. The existence of a universal concept that cuts across cultural barriers is sufficient to assure that the requirement is met. However, satisfying the second requirement is anything but trivial. Because of different cultural contexts, conditions, and languages, assuring meaning invariability is a challenging task for any cross-cultural researchers.

According to Zelditch (1971), requirements (a) and (b) together give rise to prerequisites such as units should be constant, procedures uniform, and points of reference invariant. However, the fact that a common framework of concepts is required does not necessitate empirical universals as the foundation of comparative analysis.

As a result, Verba (1971) suggests the use of measures which are functional equivalents to the theoretical concept. A concept may be operationalized or indexed by a variety of items, and different items may be the most appropriate indicators in different settings. For example, a measure of wealth can be in the form of annual income, number of heads of cattles owned, or number of acres of land owned. Sears (1961) views this point as the existence of many interchangeable operational definitions to the concept in which the cultural context serves as determinant of the appropriateness of the operational definition. The contextual grounding of measures is critical in most cases for assuring meaning invariability (Przeworski & Teune 1966; Verba 1971; Winter & Prohaska 1983; Zelditch 1971).

Przeworski and Teune (1966) recommends the seeking of equivalence when identical indicators do not provide sufficient scale reliability and when it is reasonable to expect on empirical grounds that the meaning or structure of a concept differs from culture to culture. These authors contend that indicators have identical cross-cultural validity with respect to a given concept and a given set of cultures

when the indicators are intercorrelated in a pooled, cross-cultural analysis. Equivalent indicators, on the other hand, are specific to a culture but are correlated with the set of indicators common to all cultures in the investigation.

When identical indicators are used, problems in translation become critical. Ervin and Bower (1953) identified the sources of translation distortion as stemming from (1) differences in the meaning of words, (2) differences in syntactical contexts, and (3) differences in the cultural context of the readers or hearers. Ervin and Bower (1953: 602-603) detailed the following procedure as a method of arriving at meaning equivalence:

1. In country of origin, the author of the questionnaire draws up an exhaustive set of notes while formulating the questions. These notes explain in detail each question and word used, and include synonyms and alternative phrases wherever possible.
2. In the country where the translation is to take place, the text and the notes are given to two translators who, without consulting each other, try to arrive at the best possible translation.
3. A third translator then takes both translations, as well as the explanatory notes, and without consulting the original text, indicates which of the translation seems to him to reproduce best the content and structure of the explanatory notes.
4. Finally, the original text, the two translations, and the choice of the third translator are compared, in order to decide definitely on the wording to be used.

The above procedure is applicable particularly when the main direction of the study is in the hands of researchers from the foreign country or bilingual persons with sufficient knowledge of the foreign language. However, when an established instrument is utilised the problem of translation lies in the conformity between the translated and the original version without the aid of detailed notes on its development. In this case the procedure of back translation by an independent translator is extremely useful for checking translation errors and locating language difficulties and meaning discrepancies (Ervin & Bower 1953).

In sum, the issue of measurement comparability does not provide a researcher with easy solutions with respect to when to use and when not to use identical indicators. There appears to be no sure way of knowing in advance which approach is going to be more useful. For some constructs or concepts, the use of identical indicators cross-culturally may be feasible and valid while for some others equivalent indicators have to be utilised. A possible solution lies in a posteriori evaluation of the instrument with regards to its reliability coefficient.

Nunnally (1978) states that the major use of reliability coefficients is in communicating the extent to which the results obtained from a method are repeatable. If the obtained results from several cultures achieve reliability coefficients of 0.70 or higher (Nunnally 1978: 245), the instrument yields sufficient consistency of what is measured cross-culturally though insufficient to establish its validity. However, a reliability assessment coupled with item analysis will provide an adequate indication of the appropriateness of using identical indicators. If low

reliability coefficients are obtained in other cultures, this would signify to the researcher that some of the indicators are nonfunctioning and thus would entail a major redesigning of the instrument.

THE AROUSAL SEEKING TENDENCY INSTRUMENT

The Arousal Seeking Tendency instrument by Mehrabian and Russell (1974) is a self-report measure of an individual's optimum stimulation level. Optimum stimulation level (OSL), according to Raju (1980: 272) is "a property that characterizes an individual in terms of his general response to environmental stimuli." Generally, individuals with high OSLs will have a greater tendency to explore new stimuli and situations because of a higher need for environmental stimulation. By contrast, individuals with low OSLs tend to feel more comfortable with familiar situations and stimuli and withdraw from new or unusual ones (Raju 1980).

The instrument has forty items scored on nine-point Likert scales ranging from "very strongly disagree" (- 4) to "very strongly agree" (+4). It incorporates preference for arousal caused by five major factors: arousal from change, arousal from unusual stimuli, arousal from risk, arousal from sensuality, and arousal from new environments. One half of the items are positively worded and the other one-half negatively worded. The total score of the items, after appropriate scale reversals, indicates the optimum stimulation level of the respondent.

This instrument was selected for several reasons. In the first place, the concept of optimum stimulation level is very important, especially in understanding consumer exploratory and variety seeking behaviour. The reliability test of the instrument in another culture will not only expand the utility of the instrument (if an adequate reliability coefficient is realised) but will also provide a preliminary evidence on the impact of culture. Further, the instrument has been utilised in consumer research by Goodwin (1980) and Raju (1980). The scale has also been demonstrated to have high reliability and validity (Raju 1980). The Kuder-Richardson reliability was found to be 0.87 and the test-retest reliability was 0.88, with a period of four to seven weeks between test and retest (Raju 1980). This findings would serve as a basis for comparison for the newly obtained results.

METHODOLOGY

The English version of the questionnaire consisting of the arousal seeking tendency measure and several questions on demographic characteristics was first drafted. After it was edited and refined, the instrument was given to an independent translator for the development of the Malay version. The process of back translation into English was then performed by another translator. The degree of conformity between the translated instruments was compared and discrepancies resolved

through a discussion between the researchers and the translators. For American subjects, the English only version of the questionnaire was provided. However, for the Malaysian subjects, both the English and Malay versions of the instrument were provided because of their proficiency in both languages.

The sample consisted of 120 business students at the University of Colorado, Boulder campus and 114 business students at the Universiti Kebangsaan Malaysia. A homogeneous sample was taken to keep constant such factors as age and educational level that can possibly confound the findings (Calder, Phillips, & Tybout 1981). Standardized instructions were read to both groups of subjects.

Reliability and item analyses were performed on the data using the SPSSx package. Separate assessments were made for American and Malaysian subjects. A t-test was also performed on the mean scores of the arousal seeking tendency scale.

RESULTS

Results of the reliability and item analyses are given in Tables 1 and 2. Table 1 features the item-total correlation of the arousal seeking tendency scale and the reliability coefficients (Cronbach's alpha) for both the American and Malaysian subjects. As evident from Table 1, all forty items of the scale for the American subjects appear to function well in that all item-total correlations are positive and sizeable. The lowest item-total correlation for the American subjects is 0.1613 for Item 1. As a result, a high reliability coefficient of 0.8821 is obtained for the Americans. This result is even better than that reported by Raju (1980) based on the Kuder-Richardson reliability (0.87) and the test-retest reliability (0.88). There are two major reasons for the above statement. First, the Kuder-Richardson reliability and the Cronbach's alpha are based on internal consistency of items in the scale and therefore are directly comparable (Hopkins & Stanley 1981). Second, the test-retest reliability is always higher than a reliability coefficient derived from the internal consistency approach (Hopkins & Stanley 1981).

For Malaysian subjects, item-total correlations in Table 1 are consistently lower than item-total correlations for the Americans. In fact, three items (7, 23, and 28) exhibit negative item-total correlations. When correlations are very low or negative, the items are probably miskeyed, nonfunctional, or intrinsically ambiguous (Hopkins and Stanley 1981, Nunnally 1978, Thorndike 1982). A common suggestion to overcome this problem is either to revise the items or delete them from the scale. In this case , deletion appears to be a more viable approach because of a posteriori evaluation of the instrument. After deleting Items 7, 23, and 28, Cronbach's alpha increased slightly from 0.7074 to 0.7334. It should be noted that even without item deletion, Nunnally's (1978) standard of reliability of 0.70 or higher is met.

TABLE 1. Item-total correlation of the Arousal Seeking Tendency Scale

Item	Item-total correlation	
	American	Malaysian
1. I seldom change the pictures on my walls.	.1613	.0943
2. I am not interested in poetry.	.2406	.1088
3. It is unpleasant seeing people in strange weird clothes.	.3360	.1479
4. I am continually seeking new ideas and experiences.	.5257	.2881
5. I much prefer familiar people and places.	.4981	.3428
6. When things get boring I like to find some new and unfamiliar experience.	.4519	.2104
7. I like to touch and feel a sculpture.	.3957	-.0070
8. I don't enjoy doing daring foolhardy things just for fun.	.4041	.2423
9. I prefer a routine way of life to an unpredictable one full of change.	.5134	.4726
10. People view me as quite an unpredictable person.	.4686	.0849
11. I like to run through heaps of fallen leaves.	.3397	.0792
12. I sometimes like to do things that are a little frightening.	.3647	.1949
13. I prefer friends who are reliable and predictable to those who are excitingly unpredictable.	.4380	.1242
14. I prefer an unpredictable life full of change to a more routine one.	.6156	.4216
15. I wouldn't like to try the new group-therapy techniques involving strange body sensations.	.2626	.1291
16. Sometimes I really stir up excitement.	.2019	.0447
17. I never notice textures.	.3907	.1424
18. I like surprises.	.4186	.2879
19. My ideal home would be peaceful and quiet.	.3429	.1889
20. I eat the same kind of food most of the time.	.3514	.3246
21. As a child, I often imagine leaving home just to explore the world.	.3502	.2803
22. I like to experience novelty and change in my daily routine.	.6076	.2532
23. Shops with thousands of exotic herbs and fragrances fascinate me.	.2662	-.0324
24. Designs and patterns shoud be bold and exciting.	.2423	.1171
25. I feel best when I am safe and secure.	.4226	.0495
26. I would like to hold a job as a foreign correspondent of a newspaper.	.3272	.1299
27. I don't pay much attention to my surroundings.	.2967	.0620
28. I don't like the feeling of wind in my hair.	.3588	-.0213
29. I like to go somewhere different nearly every day.	.5069	.2441

(continued)

TABLE 1 (continued)

30. I seldom change the decor and furniture arrangement at my place.	.3112	.2512
31. I am interested in new and varied interpretations of different art forms.	.3943	.3067
32. I wouldn't enjoy dangerous sports such as mountain climbing, airplane flying, or sky diving.	.2955	.4311
33. I don't like to have lots of activity around me.	.2847	.4493
34. I am interested in only what I need to know.	.3919	.1295
35. I like meeting people who give me new ideas.	.2892	.2329
36. I would be content to live in the same house the rest of my life.	.4112	.2315
37. I like continually changing activities.	.5561	.3247
38. I like a job that offers change, variety, and travel even if it involves some danger.	.6093	.5452
39. I avoid busy, noisy places.	.3017	.1395
40. I like to look at pictures that are puzzling in some way.	.4304	.1916
Cronbach's alpha	.8821	.7074
Cronbach's alpha after Items 7, 23, and 28 were deleted from the scale	.8770	.7334

The lower reliability coefficient and item-total correlations for Malaysian subjects in comparison with American subjects can be attributed to two different sources. It can be traced partly to the lesser variability in the obtained scores of Malaysian subjects with a standard deviation of 24.49 compared to the standard deviation of 31.93 for the Americans (Table 2). According to Nunnally (1978), the standard error of measurement is considered to be a fixed characteristic of any measurement tool regardless of the sample of subjects under investigation. Further, the size of the reliability coefficient is directly related to the standard deviation of obtained scores for any sample of subjects. With standard error of measurement held at 11.20, the lower standard deviation of obtained scores for Malysian subjects is expected to decrease reliability to 0.7908 from 0.8770. This value represents the maximum realizable reliability coefficient for Malaysian subjects given the lower standard deviation of obtained scores and that the standard error of measurement is constant for the two groups of subjects.

The assumption that the standard error of measurement would be the same for the two groups may be questionable if the means of the two groups are extremely different with respect to the trait in question (Nunnally 1978). From Table 2, the scale means for the Americans and Malaysians differ significantly ($p < 0.0001$) indicating that the assumption of equal standard error of measurement might be untrue. As a result, a reliability coefficient of other than 0.7908 would be expected

for Malaysians, particularly when there are other sources of errors that would further influence the coefficient.

TABLE 2. Group statistics before and after scale shortening

Group	Statistics	Before	After
American	Scale mean	43.75	42.27
	Standard deviation	34.55	31.93
	Standard error of measurement	11.86	11.20
	Number of cases	120	120
Malaysian	Scale mean	20.86	18.55
	Standard deviation	24.67	24.49
	Standard error of measurement	13.35	12.65
	Number of cases	114	114
Statistical significance of group means		<.0001	<.0001

Since the study result yields a reliability coefficient of 0.7334 forMalaysians, another potential source of error could result from the items themselves or the translation of the items. What appeared to be clearly understood among the Americans tended to generate confusion and ambiguity among the Malaysians. This problem resulted from the American or Western context of some of the items. For example, the item "I like to run through heaps of fallen leaves" is inappropriate in an environment where trees are evergreen. However, even with the above problems reliability coefficients for both the Americans and the Malaysians exceed the standard set by Nunnally (1978) for most basic research.

DISCUSSION

In most cross-cultural studies, the issue of measurement camparability arises when researchers attempt to attribute differences in obtained scores to cultural influences. For example, the significant difference in the mean optimum stimulation level between Americans and Malaysians could easily have been attributed to cultural differences (Table 2). However, this conclusion is rather shortsighted even with the use of a standardised instrument (Wallendorf & Reilly 1982; Arnould 1982). Unless further analysis is performed on the measurement scale, the differences could easily have been an artifact of poor instrument translation, inappropriate contextual setting, and/or technical or procedural concerns.

Credibility of a cross-cultural finding can be enhanced with at least a reliability assessment of the measurement instrument. Since reliability is a necessary but an

insufficient condition for validity, the reliability test serves as a prerequisite for the establishment of measurement comparability. At the very least the test would reveal to the researchers whether sufficient level of consistency has been achieved cross-culturally and if not, which items in the scale would require amendments or deletion. Providing evidence for measurement comparability is similar to the situation with construct validation. It is not an all or nothing situation. Construct validation raises the question of "Are we measuring what we think we are measuring?" In assuring measurement comparability an additional issue besides the above question is raised. This issue pertains to the question "Is the meaning of the construct the same cross-culturally?" Besides establishing the construct validity within a theoretical framework, some contextual comparisons are necessary to ensure measurement comparability.

REFERENCES

Arnould, E.J. 1982. Fancies and glimmers: Culture and consumer behavior. *Advances in Consumer Research* 10: 702-704.

Calder, B.J., Phillips, L.W., & Tybout, A.M. 1981. Designing research for application. *Journal of Consumer Research* 8: 197-207.

Ervin, S. & Bower, R.T. 1953. Translation problems in international surveys. *Public Opinions Quarterly* 16: 595-604.

Green, R.T., Cunningham, I.M. & Cunningham, W.H. 1973. Cross-cultural consumer profiles: an exploratory investigation. *Advances in Consumer Research* 1: 136-144.

Goodwin, S.A. 1980. The impact of stimulus variables on exploratory behavior. *Advances in Consumer Research* 7: 264-269.

Henry, W.E. 1961. Projective tests in cross-cultural research. In *Studying Personality Cross-Culturally*, Ed. Bert Kaplan. New York: Harper and Row, 587-596.

Hopkins, K.D. & Stanley, J.C. 1981. *Educational and Psychological Measurement and Evaluation*. Sixth edition. Englewood Cliffs, New Jersey: Prentice-Hall.

Nagashima, A. 1970 (Jan.). A comparison of Japanese and US attitudes toward foreign products. *Journal of Marketing* 34: 68-74.

Narayana, C.L. 1981 (Summer). Aggregate images of American and Japanese products: Implications on international marketing. *Columbia Journal of World Business*, 31-35.

Nunnally, J.C. 1978. *Psychometric Theory*. New York: McGraw-Hill.

Permut, S.E. 1977 (Fall). The European view of marketing research. *Columbia Journal of World Business*, 94-103.

Przeworski, A. & Teune, H. 1966. Equivalence in cross-national research. *Public Opinion Quarterly* 30: 551-568.

Raju, P. S. 1980 (Dec.). Optimum stimulation level: Its relationship to personality, demographics, and exploratory behavior. *Journal of Consumer Research* 7: 272-282.

Sears, R.R. 1961. Transcultural variables and conceptual equivalence. In *Studying Personality Cross-Culturally*, Ed. Bert Kaplan. New York: Harper and Row, 445-455.

Thorndike, R.L. 1982. *Applied Psychometrics*. Boston: Houghton-Mifflin.

Verba, S. 1971. Cross-national survey research: The problem of credibility. In *Comparative Methods in Sociology*, Ed. Ivan Vallier. Berkeley: University of California Press, 309-356.

Wallendorf, M. & Reilly, M.D. 1982. Distinguishing culture of origin from culture of residence. *Advances in Consumer Research* 10: 699-701.

Winter, L.G. & Prohaska, C.R. 1983 (Fall). Methodological problems in the comparative analysis of international marketing systems. *Journal of the Academy of Marketing Science* 11: 417-432.

Zelditch, M.Jr. 1971. Intelligible comparisons. In *Comparative Methods in Sociology*, Ed. Ivan Vallier. Berkeley: University of California Press, 267-307.

Jabatan Pemasaran
Fakulti Pengurusan Perniagaan
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor D.E.