Journal of Korean Academy of Nursing (2006) Vol. 36, No. 8, 1295 - 1300

Measurement Issues across Different Cultures

Lee, JuHee, PhD, RN, Jung, DukYoo, PhDc, RN

Purpose. The purposes of this methodologic paper are to (1) describe theoretical background in conducting research across different cultures; (2) address measurement issues related to instrument administration; and (3) provide strategies to deal with measurement issues.

Methods. A thorough review of the literature was conducted. A theoretical background is provided, and examples of administering instrument in studies are described.

Results. When applying an instrument to different cultures, both equivalence and bias need to be established. Three levels of equivalence, i.e., construct equivalence, measurement unit equivalence, and full score comparability, need to be explained to maintain the same concept being measured. In this paper, sources of bias in construct, method, and item are discussed. Issues related to instrument administration in a cross-cultural study are described.

Conclusion. Researchers need to acknowledge various group differences in concept and/or language that include a specific set of symbols and norms. There is a need to question the philosophical and conceptual appropriateness of an assessment measure that has been conceptualized and operationalized in a different culture. Additionally, testing different response formats such as narrowing response range can be considered to reduce bias.

Key Words : Measurement issues, Cultures, Equivalence, Bias

INTRODUCTION

An interest in understanding the impact of culture on health has increased (Kao, Hsu, & Clark, 2004) in fields of behavioral and psychological symptoms (King et al., 2005; Shah, Ellanchenny, & Suh, 2005), perception of health (Ng, Yau, Chan, Chan, & Ho, 2005), and quality of life (Corless, Nicholas, & Nokes, 2005; Ng, Lim, Jin, & Shinfuku, 2005). In particular, cross-cultural nursing studies have gradually increased since the mid-1950s (Im, Page, Lin, Tsai, & Cheng, 2005). Many nursing scholars have emphasized the necessity of cross-cultural research in nursing research (Leininger, 1999; Boyle, 2000). However, when conducting research across different cultures, the equivalence of measuring instruments may not be guaranteed across groups. Cultural bias can exist in the content of an instrument, the process of test taking, or the format of an instrument (Waltz, Strickland, & Lenz, 2005). Additionally, there can be culturally preferable ways in answering or difficulties among some cultural groups with the use of negative terms, multiple-choice questions, and sensitive answer sheets. These difficulties eventually can affect the reliability and validity of the instrument (Lee, Jones, Mineyama, & Zhang, 2002). For instance, Byrne and Campbell (1999) tested the Beck Depression Inventory (BDI) for Bulgarian, Canadian, and Swedish high school

1. Full-time Instructor, Yonsei University College of Nursing, Seoul, Korea
2. University of Maryland, School of Nursing, U.S.A., Doctoral candidate
Corresponding author: Jung, DukYoo, PhDc, RN, University of Maryland, School of Nursing, U.S.A., Doctoral candidate
655 West Lombard St. Baltimore. MD 21201, USA
Tel: 1-410-303-3870 Fax: 1-410-706-0344 E-mail: djung001@son.umaryland.edu; dukyoo@hanmail.net
Received March 14, 2006 ; Accepted October 20, 2006

adolescents to explore different response styles regarding degrees of skewness and kurtosis to responses across groups. They reported that Swedes were prone to either acquiescent or socially desirable responding. This tendency reflected a cultural bias as manifested in the reluctance to openly acknowledge any evidence of weakness in terms of depressive symptoms among Swedes.

Therefore, the purposes of this methodologic paper are to describe the theoretical background in conducting research across different cultures, address measurement issues in instrument administration, and provide strategies to deal with measurement issues.

THEORETICAL BACKGROUND

Van de Vijver and Poortinga (1997) explained a theoretical point of view regarding equivalence and bias. These two concepts are opposite each other, i.e., scores are equivalent when they are unbiased. Equivalence is associated with the measurement level at which scores obtained in different cultural groups can be compared. Bias is associated with the presence of factors that challenge the validity of cross-cultural comparisons (Van de Vijver & Poortinga, 1997).

Equivalence

Equivalence is a function of characteristics of an instrument and of the cultural groups involved. Establishing equivalence can support consistency of an instrument (i.e., reliability). In each cross-cultural study, the equivalence should be established and reported. There are three levels of equivalence (Van de Vijver & Poortinga, 1997).

First, when an instrument measures different constructs in two cultures (i.e., when apples and oranges are compared), no comparison can be made (Van de Vijver & Poortinga, 1997). There is no link between scores obtained in one culture and in other groups. This is called 'construct inequivalence' resulting from measurement problems. Constructs such as middle class or depression may have different meanings across cultures. For example, Byrne and Campbell (1999) administered the Beck Depression Inventory to groups of Bulgarian, Canadian, and Swedish participants. Finally, different response patterns, such as degree of skewness and kurtosis of responses, were found across groups due to acquiescent or social desirability of Swedes.

The next level is measurement unit equivalence (Van de Vijver & Poortinga, 1997). In measuring temperature using Kelvin and Celsius scales, the measurement unit is identical in both groups but the origins of the scales are not; subtracting 273 from the temperatures in Celsius will convert these into Kelvin degrees. Another example is when intelligence tests developed in the United States have been administered in Korea as a translated version. The test material may contain various implicit references to the US culture but not to the Korean culture. These references will put Korean subjects at a disadvantage. As a consequence, the interval-level scores in each group are not comparable at the ratio level.

The last level of equivalence is called full-score comparability (Van de Vijver & Poortinga, 1997). It can be achieved when the measurement instrument is on the same ratio scale in each cultural group. The measurement of body length (in centimeters or inches) and weight (in kilograms or pounds) are examples. Scalar equivalence will be achieved when scores on an instrument have the same interval scale across cultural groups.

Bias

Bias is a factor that can threaten the validity of cross-cultural comparisons (Waltz, Strickland, & Lenz, 2005). Poor item translations, inappropriate item content, and lack of standardization in administration procedures can lead to cultural bias (Kao, Hsu, & Clark, 2004; Waltz, Strickland, & Lenz, 2005). There are three types of bias:

Table 1. Type of Equivalence

Type of equivalence	Examples of each equivalence
Construct equivalence	• When an instrument measures different constructs in two cultures, no comparison can be made. • There is no link between scores obtained in one culture and in other groups.
Measurement unit equivalence	• In measuring temperature using Kelvin and Celsius scales, the measurement unit is identical in both groups but the origins of the scales are not.
Full score comparability	• The measurement of body length (in centimeters or inches) and weight (in kilograms or pounds) can be examples. • Scalar equivalence will be achieved when scores on an instrument have the same interval scale across cultural groups.

construct, method, and item bias (Waltz, Strickland, & Lenz, 2005). Following are sources of each bias and strategies to overcome them.

Construct bias

The construct bias will threaten validity when the measured construct is not identical across cultural groups, when there is no social component in instruments, or when there is a lack of overlap in behaviors associated with the construct in the cultures studied (Waltz, Strickland, & Lenz, 2005). For instance, Halbreich and Karkun (2006) reviewed the literature on prevalence of postpartum depression (PPD) and depressive symptom across countries. They found that most cited studies were conducted in Western, economically developed countries. They demonstrated that there is a wide range of reported prevalence of PPD, ranging from almost 0% to 60%. In some countries, such as Singapore, Malta, Malaysia, Austria, and Denmark, there are very few reports of PPD or postpartum depressive symptoms, whereas in other countries (e.g., Brazil, Guyana, Costa Rica, Italy, Chile, South Africa, Taiwan, and Korea), PPD is very prevalent. The authors indicated that one of the factors of variability in reported PPD might be due to differences in perception of mental health.

A poor sampling of a construct in the instrument can also be the reason for construct bias. For example, if items of a measure of coping include interpersonal situations only, the instrument will yield poor insight into intrapersonal coping mechanisms and will not generalize to instruments with a broader or differently focused item pool.

To avoid construct bias, pre-testing the measure to investigate the applicability of the construct and instrument is needed. If the constructs are not identical across cultures or contain dissimilar behaviors, a local survey can be carried out asking informants to describe the construct and its characteristic behaviors or factor scores can be compared across cultural groups (Waltz, Strickland, & Lenz, 2005). For example, if an instrument measuring filial obligation is administered in both a Chinese and an American context, different factor structures may be obtained.

Method bias

Even if a construct is well represented in an instrument, there is no guarantee that there will be no bias in the scores (Waltz, Strickland, & Lenz, 2005). Thus, bias can arise from particular characteristics of the instrument or its administration. Differences in response style such as tendency to acquiescence or interviewer effects such as communication problems can be sources of method bias (Van de Vijver & Poortinga, 1997).

Owen, Johnson, and O'Rourke (1999) found higher nonresponse rates among one or more of the minority groups when compared with non-Hispanic White respondents in four large health-related surveys. African-American respondents usually presented higher item nonresponse rates to health questions, as did males. In both subject groups, higher item nonresponse rates were found among those who had lower incomes, males, and the less educated. More educated and older respondents were more likely to refuse to answer income questions, and more educated respondents were less likely to answer "don't know." The authors concluded that social desirability explained differences across cultures. For example, Hispanics refused to answer questions regarding negative relationships. This unwillingness to report other than positive interactions with family and friends led to higher item nonresponse.

Response procedures can show differential familiarity across cultures. Bernal and colleagues (1997) used Likert-type scales to study the Insulin Management Diabetes Self-Efficacy Scale (IMDSES) of a Puerto Rican population. After the pilot testing, they had to change the response format of the scale to minimize the confusion for the respondents. The IMDSES was changed from a six-point to a four-point response format because more than four points resulted in confusion in this population. It was difficult for the subjects to respond "I agree strongly" or "I disagree strongly" with that statement. Therefore, responses were changed to "I don't feel sure"; "I feel a little sure"; "I feel more or less sure"; and "I feel very sure."

Communication problems from different interviewing skills and language problems between the examiner/interviewer and the examinee/interviewee can be a method bias. It is common in cross-cultural studies that the testing or interview language is the second or third language of interviewers, respondents, or even both. Another communication problem could arise from the use of locally inappropriate modes or other violations of local norms.

Item bias

Item bias refers to the measurement artifacts at item

level. It can be produced by various sources such as incidental differences in appropriateness of the item content (e.g., some items of an educational test are not in the curriculum in one cultural group), inadequate item formulation (e.g., complex wording), and inadequate translation. This type of bias is referred to as Differential Item Function (DIF) (Waltz, Strickland, & Lenz, 2005). Allalouf, Hambleton, and Sireci (1999) indicated that 34% of the items had DIF across languages, and the most problematic item formats were analogy items (65%) and sentence completion items (45%), respectively. Also, they mentioned that the major reasons for DIF are changes in word difficulty and item format, differences in cultural relevance, and changes in content.

Administration of instruments

Van de Vijver and Poortinga (1997) described problems and strategies to improve the instrument application in a cross-cultural study. It is based on a differentiation among test/interviewer, testee/interviewee, the interaction between these two, and response procedures.

First, the tester/interviewer is a potential source of problems. The presence of a culturally different person can affect respondents' behaviors; for example, the presence of an experimenter may influence mother-child interactions.

To solve this problem, the tester or a data collection device such as a video camera can be set up for an advance period before data collection begins. Or, there are two ways to deal with tester effects: a priori and a posteriori techniques. Examples of the former are the establishment of interviewer-interviewee acquaintance and the training of interviewers to alert them to the problem. An a posteriori technique is the measurement of tester characteristics. In studies involving many interviewers, their characteristics may be measured in order to do a statistical correction. For example, interviewers' ages and attitudes can be used as covariates in an analysis of covariance. As usual in cross-cultural studies, a priori and a posteriori techniques are complementary and cannot replace each other, i.e., a covariance analysis cannot make up for poor interviewer training and data collection.

Secondly, the testee/interviewee differences may create an administration problem. Cross-cultural studies often involve highly dissimilar groups. Consequently, groups can be different in many background characteristics, only some of which are relevant to the topic studies. A Likert-type scale format that requires the expression of attitudes, opinions, and feelings on an ordinal or interval scale may have a low ecological validity in some cultures. For example, Lee and colleagues (2002) studied different response styles in Likert scale with Chinese, Japanese, and American subjects applying a 13-question Sense of Coherence scale. Their findings show that the Japanese respondents more frequently reported difficulty with the scale, the Chinese more frequently skipped questions, and both of these groups selected the midpoint more frequently on items that involved a positive emotion than did Americans. In addition, Japanese were more likely to write "0," an option not included on the questionnaire. This finding might result from differential familiarity with the use of Likert scales and possible influence from the virtue of Confucian philosophy, which is prevalent in Asian countries.

To solve this problem, a pilot study or post hoc strategies can be used. Examples are the use of lengthy in-

Table 2. Issues in the Instrument Administration in a Cross-cultural Study

Type of problems	Strategies of problem solution
Tester/Interviewer influences	• Use the tester or a data collection device before data collection begins. • Use a priori and posteriori techniques: · A priori technique- the establishment of interviewer-interviewee acquaintance. · A posteriori technique- the measurement of tester characteristics and statistical correction the differences.
Testee/Interviewee differences	• Use a pilot study or post hoc strategies. • Ask respondents directly how familiar they are with such questions.
Tester-testee/Interviewer-interviewee interaction	• Interviewers need to be trained how to administer the interview and how to do intercultural communications with interviewee. • Openness and clarity in communication are required.
Different response procedures	• Measure the response in more than one method and compare results across methods. • Ask subjects to rate their familiarity with the response procedures and a statistical correction for cross-cultural differences.

structions including various examples and exercises and the application of an instrument in a pilot study in a nonstandard way. For instance, the instrument can be administered to a member of the target group by a researcher or an interpreter with the aim of examining the instrument instead of gathering data about the respondent. All kinds of questions are asked to establish whether the respondents answer the intended question in a meaningful way. Or the researcher can ask respondents directly how familiar they are with such questions. The respondent can be asked for his or her interpretation of an item or the reason for the answer given. When this procedure has been followed for a few respondents, it will help to identify weaknesses in the instrument and, finally, provide the information of bias presence. Additionally, a pilot testing can provide evidence of psychometric properties of reliability and validity. Nunnally and Bernstein (1994) suggested a minimum acceptable reliability coefficient alpha of .70 of a translated instrument.

Thirdly, tester-testee/interviewer-interviewee interaction can create an administration problem such as unambiguous communication. To improve communication accuracy, interviewers need to learn how to administer the interview. They should also be skilled in intercultural communication, and adequate training may be required (Waltz, Strickland, & Lenz, 2005). Also, before administering the measurement, introducing the tester or interviewer is recommended so that he or she is familiar with testee or interviewee (Waltz, Strickland, & Lenz, 2005). This includes openness and clarity in communication, an ability to assume an interviewee's viewpoint, and an inviting, non-evaluative tone of interviewing.

Finally, response procedures can be another source of administration problems. An example of response procedures is the application of a monotrait-multimethod matrix. The response is measured in more than one method, and results are compared across methods. A comparison will explain the cross-cultural differences in the two methods as measured by their effect size. Dissimilar effect sizes point to the influence of response procedures. Subjects can be asked to rate their familiarity with the response procedures applied (e.g., frequency of previous exposure), and a statistical correction for cross-cultural differences in familiarity can be carried out in an analysis of covariance.

CONCLUSION

Culture is associated with specific ethnic groups (e.g., Chinese culture); however, it can be applied to specific religious groups (e.g., Jewish groups), specific regional groups (e.g., culture of the US South), and specific institutions (e.g., military culture). Each of these groups has rules of conduct and a language that includes a specific set of symbols and norms within the group. The researcher should be aware of potential difficulties or bias when choosing or developing instruments across different cultures. Response style is significantly related to the language in which questions are posed. In the future, the study should be directed to test different response formats to reduce bias such as acquiescence and narrow response range.

Thus, researchers always should question the philosophical and conceptual appropriateness of an assessment measure that has been conceptualized and operationalized in a culture that differs from the one in which it is to be used. In addition, employing an existing measure rather than developing a new one in cross-cultural research is important (Waltz, Strickland, & Lenz, 2005). When measures are employed and tested over time, more substantial evidence for reliability and validity is built up rather than within the context of a single study. In conclusion, use of a measure from one culture to another requires attention to the cultural relevance.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *J of Edu Meas*, *36*(3), 185-198.

Bernal, H., Wooley, S., & Schensul, J. J. (1997). The challenge of using Likert-type scales with low-literate ethnic populations. *Nurs Res*, *46*(3), 179-181.

Boyle, J. S. (2000). Transcultural nursing: where do we go from here? *J of Transcultural Nurs, 11*(1), 10-11.

Byrne, B. M. & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *J of Cross-cultural Psychology, 30*(5), 555-574.

Corless, U. B., Nicholas, P. K., & Nokes, K. M. (2005). Issues in cross-cultural quality of life research. Image: *J of Nurs Schol*, *33*(1), 15-20.

Halbreich, H. & Karkun, S. (2006) Cross-cultural and social diversity of prevalence of postpartum depression and depressive symptoms. *J of Affect Disorder*, Feb, 5

Im, E., Page, R., Lin, L., Tsai, H., & Cheng, C. (2005). Rigor in cross-cultural nursing research. *Int J of Nurs Stu, 41*, 891-899.

King, M. et al. (2005). Psychotic symptoms in the general population of England-a comparison of ethnic groups (The EMPIRIC study). *Soc Psych & Psych Epi, 40*(5), 375-81.

Kao, H., Hsu, M., & Clark, L. (2004). Conceptualizing and critiquing culture in health research. *J of Transcultural Nurs 15*(4), 269-277.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Res in Nur & Health, 25*, 295-306.

Leininger, M. (1999). What is transcultural nursing and culturally competent care? *J of Transcultural Nurs, 10*(1), 9.

Ng, T. P., Lim, L. C., Jin, A., & Shinfuku, N. (2005). Ethnic differences in quality of life in adolescents among Chinese, Malay and Indians in Singapore. *Quality of Life Res, 14*(7), 1755-68.

Ng, S. M., Yau, J. K., Chan, C. L., Chan, C. H., & Ho, D. Y. (2005). The measurement of body-mind-sprit well-being toward multidimensionality and transcultural applicability. *Social Work in Health Care, 41*(1), 33-52.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). New York: McGraw Hill.

Owens, L., Johnson, T. P., & O'Rourke, D. (1999). *Culture and item nonresponse in health surveys*. Paper presented at the meeting of The 7th Conference on Health Survey Research Methods. Williamsburg, VA.

Shah, A., Ellanchenny, N., & Shu, G. H. (2005). A comparative study of behavioral psychological symptoms of dementia in patients with Alzheimer's disease and vascular dementia referred to psychogeriatric services in Korea and the U.K. *Int Psychogeriatrics, 17*(2), 207-19.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *Eur J of Psych Assess, 13*, 21-29.

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2005). *Measurement in nursing and health research* (3rd ed.). NY: Springer Publishing Company.