Working together
www.rcis.ro

Revista de Cercetare si Interventie sociala

ISSN: 1583-3410 (print), ISSN: 1584-5397 (electronic)

SYSTEMATIC TEST OF EQUIVALENCE PROCEDURE: NEW METHOD TO INVESTIGATE CROSS-CULTURAL VALIDITY

Lawrence H. GERSTEIN

Revista de cercetare și intervenție socială, 2018, vol. 62, pp. 278-293

The online version of this article can be found at:
www.rcis.ro, *www.doaj.org* and *www.scopus.com*

Published by:
Expert Projects Publishing House

expert projects
publishing

On behalf of:
„Alexandru Ioan Cuza" University,
Department of Sociology and Social Work
and
HoltIS Association

REVISTA DE CERCETARE SI INTERVENTIE SOCIALA
is indexed by Clarivate Analytics (Web of Science) -
Social Sciences Citation Index
(Sociology and Social Work Domains)

Systematic Test of Equivalence Procedure: New Method to Investigate Cross-Cultural Validity

Lawrence H. GERSTEIN[1]

Abstract

To conduct valid cross-cultural research, it's essential to reduce bias (e.g., construct, method) and demonstrate equivalence (e.g., construct, linguistic). This paper introduces a new form of equivalence, method equivalence that exists if the same data collection (e.g., self-report scales, interviews) or research methods (e.g., laboratory or field studies) can be used in multiple cultures or countries. Assuming a self-report scale is appropriate to employ, to further establish method equivalence, a culturally valid item response format (e.g., Likert, Forced-Choice, Thurstone) also must be utilized. Further, this paper introduces a new methodology known as the Systematic Test of Equivalence Procedure (STEP) to investigate the cross-cultural validity of constructs, data collection and research methods, scale items, and item response formats in multiple countries or diverse cultures. Additionally, STEP can be used to examine and establish the construct and method equivalence of a measure and reduce the possibility of construct, method, and item bias.

Keywords: equivalence, bias, cross-cultural methodology, Systematic Test of Equivalence Procedure, and method equivalence.

Introduction

For decades, mental health professionals in the United States (U.S.) have conducted cultural and cross-cultural studies abroad. There is a large body of literature, yet many have criticized this scholarship for its potential lack of cultural validity. Some have suggested a Eurocentric or U.S.-centric bias has driven this endeavor shaping and possibly biasing the results that have been discovered (Gerstein, Heppner, Ægisdóttir, Leung, & Norsworthy, 2009; Gerstein, Hurley, & Hutchison, 2015; Leong & Ponterotto, 2003; Leung, 2003; Marsella, 1998; Takooshian, 2003). To address these concerns and to promote more sensitive, appropriate, and relevant research procedures when performing studies worldwide, other scholars (Brislin, 1976; van de Vijver, 1998; Ægisdóttir, Gerstein, & Cinarbas,

[1] Ball State University, College of Health, Department of Counseling Psychology, Social Psychology & Counseling, Muncie, Indiana 47306 USA. E-mail: lgerstein@bsu.edu

2008; Ægisdóttir, Gerstein, Leung, Kwan, & Lonner, 2009) have introduced some conceptual, methodological, and linguistic prerequisites to guide the formulation and proper implementation of cross-cultural research. These scholars also have stressed the critical importance of demonstrating cross-cultural validity in such investigations. Cross-cultural validity refers to establishing the appropriateness and relevance, for example, of specific concepts, methods, and statistics when performing research in different cultures. Stated another way, cross-cultural validity is said to exist when a construct defined in one culture and the method used to study it and statistics employed to assess it are valid in another culture (Gerstein, 2011).

Various types of equivalence (i.e., construct, linguistic, scalar [metric], structural, & measurement) must be present to demonstrate evidence of cross-cultural validity. Perhaps the most fundamental form of equivalence that must be established is construct equivalence. Construct equivalence refers to the fact that the construct of interest has the same basic meaning across the target cultures or countries (He & van de Vijver, 2012; Lonner, 1985; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 1997; Ægisdóttir *et al*., 2008). For example, the construct called conflict resolution would be understood in the same way in the U.S. and Botswana. To demonstrate construct equivalence, however, conflict resolution would *not need* to be operationally defined the same way in the U.S. and Botswana. If conflict resolution was not viewed similarly in these two countries and a study was conducted without addressing this discrepancy appropriately, there would be concerns about construct bias or nonequivalence.

Another form of equivalence that is introduced for the first time in this paper is method equivalence. This refers to the observation that the same data collection (e.g., self-report scales, interviews, scenarios, direct observation) or research methods (e.g., laboratory or field studies) can be used to study the construct of interest, for instance, anxiety, in the two or more targeted cultures or countries. For example, if an investigator wanted to study test-taking anxiety among college students in Jordan and the U.S. s/he would need to determine if it was appropriate to use the same procedure such as self-report measures. If the same procedures could be, and were used, then method equivalence would be demonstrated. If a self-report measure were not appropriate to administer in Jordan because the population of interest could better relate to an interview, using the self-report instrument would introduce what is known as method bias into the study of test-taking anxiety. There is some evidence, in fact, that various research methods (e.g., self-report surveys vs. computer-assisted instruments) lead to different outcomes (van de Vijver & He, 2016). Interestingly, however, there is a dearth of research on this topic even though to demonstrate equivalence it is essential to first establish that the research methods employed are valid and relevant for the individuals from different countries or cultures participating in a study. It is highly recommended, therefore, that investigators conduct projects to assess whether particular data collection procedures (e.g., self-report scales, interviews,

scenarios, direct observation) and research methods (e.g., laboratory versus field study) are more relevant, understandable, and/or preferred by individuals affiliated with various countries and cultures.

Along with determining if the research procedure is appropriate to employ to assess or collect data on a construct in two or more cultures or countries, if a scale is to be administered, to demonstrate method equivalence the researcher also must employ a culturally valid item response format. In other words, it must be determined if the same item response format (e.g., Likert vs. Thurstone scale) can be used. Returning to the test-taking anxiety example, if a Likert item format was not appropriate to use in Jordan because the sample could better relate to a forced choice item format (e.g., yes or no), using a Likert scale also would introduce into the study instrument bias which is a form of method bias. People from different cultures vary in their familiarity with response formats such as forced-choice (He & van de Vijver, 2012) and Likert type formats (Hambleton, 2005). In a study conducted in Lebanon, for example, secondary school and university students were found to be unfamiliar with Likert type scales (Halloun, 2001). Responses to different item formats like phrase versus word-pair formats (Osterlind, Miao, Sheng, & Chia, 2004) and Likert versus forced-choice formats (Javeline, 1999) also have been found to vary by cultures though surprisingly there is very little research on this particular topic. Given the importance of assessing for such differences to establish equivalence, it is understandable that the International Test Commission Guidelines for Translating and Adapting Tests (2015) stated that test developers should provide evidence that the targeted population is familiar with a scale's item response format.

There are other types of method bias that have been observed in cross-cultural and cultural studies. In general, research has indicated that response styles or participants relying on certain aspects of the item response format and not content when answering (Cronbach, 1950) has resulted in method bias when performing cross-cultural studies. There is evidence, for instance, indicating a modesty effect response in Asia (Van de Gaer, Grisay, Schulz, & Gebhardt, 2012), acquiescence (e.g., preference to agree) and midpoint (e.g., preference for neutral) response styles on a Likert scale when conducting studies in Asia (Heine *et al*., 2001; Heine, Lehman, Peng, & Greenholtz, 2002; Lie & Turmo, 2005; Peng, Nisbett, & Wong, 1997) and Germany (Wetzel, Lüdtke, Ingo Zettler, & Böhnke, 2016), and extreme response styles (e.g., preference for strongly agree, strongly disagree) in Germany (Wetzel *et al*., 2016). Moreover, cross-national and within country differences in these response styles have been shown in numerous studies, along with an association between cultural orientation and these styles (For a review see Johnson, Kulesa, Cho, & Shavitt, 2005).

As stated earlier, there are many other forms of equivalence (e.g., linguistic, scalar [metric], structural, & measurement) that must be established when performing cross-culturally valid research that is designed to test hypotheses and/or investigate research questions. If equivalence is not established, there is the

potential that many alternative hypotheses may explain the observed cross-cultural differences (Ægisdóttir *et al*., 2008), and the results might be erroneous (Chen, 2008; Steenkamp & Baumgartner, 1998). Further, if various forms of equivalence are not established, there is a higher potential that different types of bias might be present in the study being performed. Two forms of bias, construct and method were mentioned earlier. Another form is known as item bias where individuals from various cultures or countries, for instance, understand an item differently (He & van de Vijver, 2012).

It should be noted that bias negatively influences equivalence, or stated another way, as bias increases, equivalence decreases (Ægisdóttir *et al*., 2008). When bias is high, observations across cultural groups are not comparable. Obviously, a prerequisite to conducting a valid cross-cultural study is to make sure what is being studied exists, that the construct of interest is functionally equivalent across the cultures (Berry, 1969), and that it can be investigated using the same or properly adapted research methods or procedures including, if appropriate, the item response format to fit each culture/country. For example, in a study conducted many years ago in India that investigated violence between two different religious groups, the researchers first thought to include the concepts of empathy and forgiveness because these were often studied in the U.S. when performing projects on violence between groups. However, in discussions with some Indians in India and in the U.S., it was discovered that these words did not exist in India nor were the concepts perceived and enacted in a similar way as in the U.S. Moreover, investigating these concepts in India using a Likert scale format was thought to be problematic because Indians could not relate to this item response format. If the study had been performed as originally intended, it would have introduced a high degree of construct non-equivalence, and construct and method bias. To reduce this possibility, the study was designed to interview Indians who were members of two different religious groups, and in so doing, gathered qualitative data that enhanced the cultural validity of the project (Shankar & Gerstein, 2007).

Clearly, it is important to demonstrate various forms of equivalence and to reduce bias before, during, and following the completion of performing a cross-cultural study. Some scholars (Matsumoto & van de Vijver, 2011) even have argued that data equivalence must be first conceptually and statistically established. A number of strategies have been introduced and used to assess and enhance equivalence and reduce bias when conducting a cross-cultural study including, for example, employing statistical procedures such as ANOVA, MANOVA, exploratory and confirmatory factor analysis, differential item functioning analysis, multidimensional scaling (He & van de Vijver, 2012; Ægisdóttir *et al*., 2008), simultaneous components analysis (Kiers, 1990), multiple-group SEM invariance analysis, and multiple group mean and covariance structures analysis (Ægisdóttir *et al*., 2008). Other approaches that have been employed to evaluate and strengthen equivalence and reduce bias include instrument procedures like "the dual language split half" (Mallinckrodt & Wang, 2004) and linguistic procedures like decentering,

and convergence (van de Vijver & Poortinga, 2005), application, and adaptation (Ægisdóttir *et al*., 2008) strategies.

While the aforementioned approaches are quite useful and often used to increase various types of equivalence and to decrease different forms of bias, there appears to be no concrete research procedure currently available that is designed to test the relevant validity or equivalence of a construct. Moreover, there is no existing methodology or scale created to investigate the relevance or validity (method equivalence) of employing various research procedures (e.g., self-report surveys, interviews, scenarios, direct observation) and/or item response formats (e.g., Likert, semantic differential, forced choice) when conducting a study in two or more cultures or countries.

This paper addresses this gap by introducing a new cross-cultural methodology known as the Systematic Test of Equivalence Procedure (STEP) to investigate the relevance or cross-cultural validity of constructs, research methods, and scale items and item response formats in multiple countries or diverse cultures. Further, STEP can be used to examine and establish the construct and method equivalence of a measure and reduce the possibility of construct, method, and item bias. Quantitative relevance ratings and qualitative responses linked to scale items and factors in one country/culture are gathered from experts and/or the target participants in another country or culture. Further, experts and/or participants similar to the ones to be involved in the intended study suggest potential additional items for existing factors, along with new factors and their corresponding items. This new procedure yields etic and possible emic items and factors that can then be used to conduct a valid investigation in the targeted culture or country.

It is important to note that STEP is a cross-cultural methodology and based on the assumption that data will be gathered in two or more cultures or countries. The aim of employing this methodology is to determine if a construct exists across countries or cultures, and if it does, to understand if it "behaves" differently or similarly in the two unique contexts. Further, the aim of using STEP is also to determine if the research procedure, scale items, and item response formats to be employed are equivalent in the two or more cultures or countries.

How does STEP work?

There are six steps to STEP (see Figure 1). Each will be described and then illustrated with an example from a research project conducted in India on everyday sexual harassment (Bellare & Gerstein, 2016) to illustrate the particular step. Given the language skills of the participants involved in this STEP study, it was conducted in English. It should be noted, however, that if a translation back-translation process would have been required to create a different language version of the STEP, or for a measure or other research procedure, this process would have occurred before performing the six steps.

1st Step: When conducting research with _____ (e.g., adults) in _____ (e.g., Argentina) is it appropriate to use the following methods? In other words, are these methods culturally appropriate and relevant to use in _____ (e.g., Argentina) with _____ (e.g., adults). Check all that are appropriate.

1. __Self-report quantitative survey instruments
2. __Direct observation strategies
3. __Hypothetical scenarios (cases)
4. __Informants (i.e., collect information from individuals that know a project participant)
5. __Interviews
6. __Questions on an instrument that ask the participant to write a response
7. __Focus groups
8. __Laboratory studies
9. __Field studies
10. __Other method (please describe)
11. Comments you would like to share:

2nd Step: If it is culturally appropriate and relevant to administer self-report quantitative survey instruments to _____ (e.g., adults) in _____ (e.g., Argentina), is it culturally appropriate and relevant to employ any of the following item response formats? Check all that are appropriate.

1. __Likert scale
2. __True-False format
3. __Forced choice format (e.g., multiple choice; rank ordering)
4. __Semantic differential scale
5. __Thurstone scale
6. __Guttman scale
7. Comments you would like to share:

3rd Step1: The _____ scale measures ……. Given this description, using the scale below, please share your perceptions about the relevance of the _____ scale for _____ (e.g., adults in Argentina).

Highly irrelevant					Highly relevant
1	2	3	4	5	6

Rating for _____ (e.g., Argentina adults)_____

Factor 1: _____
This factor is comprised of ___-items (e.g., _____) that assess…….

Highly irrelevant					Highly relevant
1	2	3	4	5	6

Rating for _____ (e.g., Argentina adults)_____

Factor 2: _____
This factor is comprised of ___-items (e.g., _____) that assess…….

Highly irrelevant					Highly relevant
1	2	3	4	5	6

Rating for _____ (e.g., Argentina adults)_____

If you rated any of the factors 1, 2, or 3, please explain your reason(s):

4th Step: Do you think there are other factors or constructs that were not measured by the scale? If you think there are others, please kindly describe each new factor and identify about 8-10 items for each new factor.

5th Step[1]: Item Rating Scale for the _____ Scale.
For each item below, rate how relevant it is for _____ (e.g., adults in Argentina) using the following scale:

Highly irrelevant					Highly relevant
1	2	3	4	5	6

List each item of the targeted scale followed by a place for participants to share their rating.
Item 1……..
Rating _____
Item 2……..
Rating _____

If you rated any of the ____ (number) items on the _____(name of scale) below 1, 2, or 3, please explain your reason(s).

6th Step: Here are the items that are linked to the current ____ (e.g., U.S.) _____Scale factors: Indicate the name of the factors here. Do you think there

were items that were missing that need to be included on the current _____ (e.g., U.S.) factors to assess them for adults in Argentina? If you do, what are they?

Factor (Include all factors on scale)	*Items (Include all items for each factor)*
Indicate Name of Factor 1	Item 1: List item
	Item 2: List item
	Suggested new items:
Indicate Name of Factor 2	Item 1: List item
	Item 2: List item
	Suggested new items:

[1]For the purposes of this illustration, these steps assume a Likert scale response format is a valid and relevant strategy to collect data from the targeted experts and participants. However, when creating these steps for a STEP study, the researcher must first determine which item response format is most appropriate to employ.

Figure 1: The Systematic Test of Equivalence Procedure (STEP)

Steps 1 and 2 involve requesting experts to assess the relevance of different research procedures and item response formats to the countries or cultures being studied. At least three experts in research methods employed in these countries or cultures are needed for this step. In the India study (Bellare & Gerstein, 2016), the experts were four professionals who conduct psychological research in India. For Step 1 (Appropriate/Relevant Methodology), these persons were asked, "When conducting research with college students in India is it appropriate to use the following methods? In other words, are these methods culturally appropriate and relevant to use in India with college students. Check all that are appropriate." The choices they were given and their responses are reported below. The results revealed almost all the experts considered each procedure as relevant when collecting data from Indian college students.

4 responses: Self-report quantitative survey instruments
3 responses: Direct observation strategies
3 responses: Hypothetical scenarios (cases)
3 responses: Informants (i.e., collect information from individuals that know a project participant)
4 responses: Interviews
4 responses: Questions on an instrument that ask the participant to write a response
4 responses: Focus groups
0 responses: Other method (please describe)

The experts also were informed, if they wished, to share comments about these methods. Three comments were reported, "Hypothetical scenarios are excellent

to use with Indian college students," "Direct observation is not well taken," and "Students learn about these methods particularly self report, hypothetical scenarios, and interviews."

For Step 2 (Appropriate/Relevant Item response format), the same experts were asked, "If it is culturally appropriate and relevant to administer self-report quantitative survey instruments to college students in India, is it culturally appropriate and relevant to employ any of the following item response formats? Check all that are appropriate." The options they were given and their responses follow. Again, almost all the experts viewed the item formats as relevant when gathering data from Indian college students.

4 responses: Likert scale
3 responses: True-False format
4 responses: Forced choice format (e.g., multiple choice)
3 responses: Semantic differential scale
3 responses: Thurstone scale
4 responses: Guttman scale

Like Step 1, the experts also were told, if they wished, to share comments about these item response formats. Two comments were reported, "Suggest not having midpoint if using Likert scale" and "Combination of Likert and forced choice formats are good in India."

For Step 3, another group of at least three experts are requested to assess the relevance to their country or culture of the scale and its constructs/factors under investigation. It is also recommended that responses be gathered from at least three persons similar to members of the sample to be recruited for the eventual cross-cultural study focused on the targeted constructs. The experts are selected based on their knowledge about how the constructs of interest are understood in their culture or country. When performing this assessment, the experts (or members of the sample population) respond to a series of items accompanied by 6-point Likert scales (assuming this item format is culturally appropriate) ranging from Highly Irrelevant (1) to Highly Relevant (6). If the participants share low relevance ratings of 1, 2, or 3, they are asked to provide a reason for their rating.

In the Bellare and Gerstein (2016) study being used to illustrate STEP in this article, for Steps 3 to 6, four different professionals with expertise in the area of everyday stranger harassment, and/or sexual harassment or gender-based discrimination residing in India or the U.S. were recruited to assess the cultural relevance of the scale items and constructs included on the Fear of Rape Scale (Senn & Dzinas, 1996). The criteria used in this study to define and select an expert included, at least one publication, presentation, workshop, or seminar in the field of everyday stranger harassment, sexual harassment, or gender-based

discrimination. Responses were not collected in this study from persons similar to members of the sample to be recruited for the eventual cross-cultural study focused on everyday stranger harassment in India.

For Step 3, the India experts were first told, "The Fear of Rape scale measures the fear of being raped by strangers. Given this description, please share your perceptions about the relevance of the Fear of Rape Scale for Indian college students." When doing this, the experts rated relevance on a Likert scale of 1 (Highly Irrelevant) to 6 (Highly Relevant). The mean rating obtained was 5.25 indicating the experts viewed this scale as relevant to administer to Indian college students.

Next, the experts were requested to rate, using the same Likert scale, the Fear of Rape factor measured by the scale. They were told, "This factor is comprised of 31-items (e.g., I am afraid of being sexually assaulted; I think twice before going out for a walk late at night) that reflect the fear of being raped by a stranger and the precautions taken by women to protect themselves from being raped." Again, the mean rating obtained was 5.25 suggesting the experts viewed this factor as relevant for Indian college students.

The last instruction for Step 3, asked the experts, "If you rated the scale or factor three or below, please explain your reason(s)." None of the experts involved in the India study rated the scale or factor three or below.

For Step 4, the same experts and/or members of the sample population employed in Step 3 are told to suggest any additional factors that are not measured by the questionnaire being investigated but should be included when conducting research in their country or culture. They also are informed to identify and describe the new factor(s), and to identify 8 to 10 items that may adequately capture this factor.

In the India study (Bellare & Gerstein, 2016), the experts were asked, "Do you think there are other factors or constructs that were not measured by the Fear of Rape scale? If you think there are others, please kindly describe each new factor and identify about 8-10 items for each new factor." The India experts mentioned no other factors or constructs.

For Step 5, experts and/or members of the sample population are requested to assess the relevance to their country or culture of individual items found on the targeted questionnaire using Likert scales ranging from Highly Irrelevant (1) to Highly Relevant (6). If the participants rate any of the items below 3, they are then asked to provide reasons.

In the India study (Bellare & Gerstein, 2016), the experts were told, "For each (Fear of Rape Scale) item, rate how relevant it is for college students in India using the following scale" with the same anchor points mentioned in the previous paragraph. The experts also were informed, "If you rated any of the items on the Fear of Rape Scale 1, 2, or 3, please explain your reason(s)." What follows are some example items that were rated below 3 and how these items were modified (change appears in italics) to be culturally relevant for college students in India.

Original item: Before I go to bed at night I double check to make sure the doors are securely locked. (Mn expert rating = 2.0)

Expert comment: Students do not have this responsibility. It is their parents or the elders in the household that have the responsibility.

Modified item: Before I go to bed at night I check with my parents or elders to make sure that the doors are securely locked.

Original item: I ask friends to walk me to my car/the subway if it is late at night. (Mn expert rating = 2.75)

Expert comment: There is no subway in India but there are train stations. Also, it is common for friends to walk with each other to a bus stop late at night.

Modified item: I ask friends to walk me to my car/the train station/bus stop if it is late at night.

Original item: If I was driving alone and I had to park my car I would try to park on a well lit street. (Mn expert rating = 2.75)

Expert comment: Most college students do not have access to a car to drive.

Modified item: If I had a car and I was driving alone and I had to park my car I would try to park on a well lit street.

Finally, for Step 6, the same experts and/or members of the sample population that were involved in Steps 3 to 6 are asked to determine if the items loading on the different targeted factors/constructs are relevant to the country or culture being studied and they are also requested to suggest additional items that may be required to measure these factors. In the Bellare and Gerstein (2016) study, for this Step, the experts were told, "Here are the items that are linked to the current USA Fear of Rape Scale. Do you think there are items that were missed that need to be included on the current USA factor to assess it in India for college students? If you do, what are they?" A list of the 31 items on this scale followed this instruction.

The experts recommended two new items (Party/drinking and its impact on fear of rape; Sometimes, fear of rape develops when a girl has had an argument with a guy and it would have hurt his ego and the guy threatens her with rape.). These items, however, were not incorporated into a revised version of the Fear of Rape Scale because they were considered not directly relevant to the scale.

Given the results of the STEP just reported and the modifications to the Fear of Rape Scale adopted, it is now possible to conduct a cross-cultural study whereby this scale can be administered to college students in the U.S. and India. In summary, it was found that a similar data collection procedure (i.e., self-report surveys) and item response format (i.e., Likert scale) could be employed in the U.S. and India thereby demonstrating method equivalence. Though slight changes to some of the Fear of Rape Scale items were recommended to decrease construct and item bias, these modifications did not change the meaning of the items. If new items and/or factors that were indigenous to India had been adopted to assess fear of rape, then it would be much more difficult to analyze the responses collected in a

cross-cultural study. One suggestion to address this challenge would be to analyze (e.g., factor analysis, MANOVA) only responses to the common items and factors found on the Fear of Rape Scale administered to the Indian and U.S. participants. Following this, another analysis (e.g., factor analysis, MANOVA) could focus on both groups of participants' responses to the new items and factors. If the new items and factors were, in fact, unique to one country, in this case India, you would expect differences in responses with higher (or lower) scores for the non-U.S. group (in this case Indian participants) than the original population (in this case U.S. participants).

In conclusion, if the Bellare and Gerstein (2016) STEP project briefly discussed above was not first conducted in India, and instead, Indians' responses to the U.S. version of the Fear of Rape Scale were collected in a cross-cultural study, the interpretations and conclusions drawn about the findings would most likely have been erroneous. As others have argued, without demonstrating equivalence, alternative hypotheses may account for observed cross-cultural differences (Ægisdóttir et al., 2008).

It should be mentioned when an earlier version of STEP was employed in other studies, the experts raised some similar concerns to those expressed in the Bellare and Gerstein (2016) study. In specific, this earlier version of the STEP was used in a study in Korea (Gerstein, 2012) involving the Beliefs About Psychological Services Scale (Ægisdóttir & Gerstein, 2009), another study in Korea (Kim & Gerstein, 2011) focused on the Calling and Vocation Questionnaire (Dik, Eldridge, Steger, & Duffy, 2012), and a study in Hong Kong (Gerstein et al., 2018) with the Teenage Nonviolence Test (Gerstein, Mayton, Hutchison, & Kirkpatrick, 2014). As in the Bellare and Gerstein (2016) study, in the three studies just mentioned, some items were modified based on the solicitation of experts' ratings. Further, in two of these studies (Gerstein et al., 2018; Kim & Gerstein, 2011), the experts also recommended additional constructs/factors to assess other indigenous constructs not measured by the original U.S. versions of the scales.

Conclusion

Conducting cross-cultural research is a complicated and challenging yet critical and urgent endeavor if we are to further identify universal and unique attributes of individuals, groups, cultures, and countries, as well as constructs, data collection procedures, and research methods. Much has been written about the importance of this undertaking, obstacles to performing cross-culturally valid research, and strategies to identify and address these obstacles in the formulation, implementation, and outcome stages of conducting studies. Two essential components to assessing the quality and validity of cross-cultural research are bias and equivalence. Both must be sufficiently addressed to have confidence in the internal and external validity of the obtained results and the corresponding interpretations and implications.

Until now, the concept of method equivalence or first establishing when conducting cross-cultural or cross-national research the validity of different data collection and research methods, and if appropriate, item response scale formats, has not appeared in the literature. Like other forms of equivalence, when formulating and implementing research, investigators must avoid or reduce bias, in this case method bias, and then demonstrate what has been called in this article method equivalence to perform cross-culturally valid research projects.

To pursue the tasks just mentioned, a new cross-cultural methodology, STEP, created to investigate the relevance or cross-cultural validity of constructs, data collection procedures, research methods, scale items, and item response formats in two or more countries or cultures was introduced in this paper. Additionally, STEP is a tool that can be used to examine and establish the construct and method equivalence of a measure and reduce the likelihood of construct, method, and item bias. The basic structure and content of STEP also may be adapted to assess the relevance, validity, and applicability of theories and intervention strategies to be employed with individuals from different countries or cultures.

While using the STEP in previous studies has yielded valuable results to strengthen the cross-cultural validity of these projects, additional research particularly involving data collection procedures other than self-report scales is required to further establish the utility and validity of this tool. STEP studies also need to be performed to determine the utility and validity of this strategy when assessing the relevance or cross-cultural validity of employing different research methods (e.g., laboratory versus field projects) with various cultures or countries. Finally, STEP studies need to be conducted with both experts and representatives of target samples to investigate the reliability of this device. In the meantime, the STEP approach appears to be a pragmatic and useful strategy to collect data when assessing the cross-cultural validity and relevance of constructs, data collection procedures, scale items, and item response formats.

References

Bellare, Y., & Gerstein, L.H. (2016, March). *The Systematic Test of Equivalence Procedure and everyday stranger harassment.* Poster presented at the Great Lakes Regional Counseling Psychology Conference, Indiana University, Bloomington, Indiana.

Berry, J.W. (1969). On cross-cultural comparability. *International Journal of Psychology*, 4, 119-128.

Brislin, R.W. (1986). The wording of translation of research instruments. In W.J. Lonner and J.W. Berry (Eds.), *Field Methods in Cross-Cultural Research* (pp. 137-164). Thousand Oaks, CA: Sage.

Chen, F.F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95, 1005-1018.

Cronbach, L.J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 3-31.

Dik, B.J., Eldridge, B. M., Steger, M. F., & Duffy, R. D. (2012). Development and validation of the Calling and Vocation Questionnaire (CVQ) and Brief Calling Scale (BCS). *Journal of Career Assessment*, 20, 242-263.

Gerstein, L.H. (2011). Research, practice and training across and within cultures: Simply complicated! *Journal of Asia Pacific Counseling*, 1(1), 1-12.

Gerstein, L.H. (2012, August). *A new method to explore cross-cultural validity of instruments*. Symposium Chair at the Annual Meeting of the American Psychological Association. Orlando, Florida.

Gerstein, L.H., Chan, Y., Hutchison, A., Fung, A.L.C., Kinsey, R., & Jeffers, H. (2018). The Teenage Nonviolence Test: Applicability in Hong Kong? *Current Psychology* 37(1), 313-324. doi:10.1007/s12144-016-9514-3.

Gerstein, L.H., Heppner, P.P., Ægisdóttir, S., Leung, S. A., & Norsworthy, K. L. (2009). Cross-cultural counseling: History, challenges, and rationale. In L.H. Gerstein, P.P. Heppner, S. Ægisdóttir, S.A., Leung, & K.L. Norsworthy (Eds.). *International handbook of cross-cultural counseling: Cultural assumptions and practices worldwide* (pp. 3-32). CA: Sage Publications.

Gerstein, L.H., Hurley, E., & Hutchison, A. (2015). The Dynamic-Systemic-Process Model of International Competencies for Psychologists and Trainees. *Revista de Cercetare si Interventie Sociala*, 50, 239-261.

Gerstein, L.H., Mayton, D., Hutchison, A., & Kirkpatrick. D. (2014). The Teenage Nonviolence Test: A factor analytic investigation. *Revista de Cercetare si Interventie Sociala, 44*, 9-19.

Halloun, I.A. (2001). *Student views about science: A comparative study*. Monograph. Educational research center, Lebanese University, Beirut, Lebanon.

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P.F. Merenda, C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, Unit 2. Retrieved from http://scholarworks.gvsu.edu/orpc/vol2/iss2/8

Heine, S.J., Kitayama, S., Lehman, D.R., Takata, T., Ide, E., Leung, C., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology*, 81, 599-615.

Heine, S.J., Lehman, D.R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales: The reference-group problem. *Journal of Personality and Social Psychology*, 82, 903-918. doi: http://dx.doi.org/10.1037/00223514.82.6.903

International Test Commission. (2015). *ITC Guidelines for Translating and Adapting Tests*, Version 1. Retrieved from https://www.intestcom.org/files/guideline_test_adaptation.pdf

Javeline, D. (1999). Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*, 63(1), 1-28.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, 36, 264-277.

Kiers, H.A.L. (1990). *SCA: A program for simultaneous components analysis*. Groningen, Netherlands: IEC ProGamma.

Kim, T.S., & Gerstein, L.H. (2011, August). *Cross-Cultural Validity of Career Calling in Korea: Preliminary Research*. Poster presented at the Annual Meeting of the American Psychological Association. Washington, D.C.

Leong, F.T.L., & Ponterotto, J.G. (2003). A proposal for internationalizing counseling psychology in the United States: Rationale, recommendations, and challenges. *The Counseling Psychologist*, 31, 381-395.

Leung, S.A. (2003). A journey worth traveling: Globalization of counseling psychology. *The Counseling Psychologist*, 31, 412-419.

Lie, S., & Turmo, A. (2005). *Cross-country comparability of students' self-reports. Evidence from PISA 2003* (OECD/PISA, Document: TAG(0505)11). Paris: OECD.

Lonner, W.J. (1985). Issues in testing and assessment in cross-cultural counseling. *The Counseling Psychologist*, 13, 599-614.

Mallinckrodt, B., & Wang, C.-C. (2004). Quantitative methods for verifying semantic equivalence of translated research instruments: A Chinese version of the Experiences in Close Relationships Scale. *Journal of Counseling Psychology*, 51, 368-379.

Marsella, A.J. (1998). Toward a "global-community psychology": Meeting the needs of a changing world. *American Psychologist*, 53, 1282-1291.

Matsumoto, D., & van de Vijver, F.J.R. (2011). *Cross-cultural research methods in psychology* (Eds.). New York, NY: Cambridge University Press.

Osterlind, S., Miao, D., Sheng, Y., & Chia, R.C. (2004). Adapting item format for cultural effects in translated tests: Cultural effects on construct validity of the Chinese versions of the MBTI. *International Journal of Testing*, 4, 61-73.

Peng, K., Nisbett, R.E., & Wong, N.Y.C. (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods*, 2, 329-344.

Senn, C. Y., & Dzinas, K. (1996). Measuring fear of rape: A new scale. *Canadian Journal of Behavioural Science*, 28, 141-144.

Shankar, J., & Gerstein, L.H. (2007). The Hindu-Muslim conflict: A pilot study of peacebuilding in Gujarat, India. Peace and Conflict: *Journal of Peace Psychology*, 13, 365-379.

Steenkamp, J.B.E.M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78-90.

Takooshian, H. (2003). Counseling psychology's wide new horizons. *The Counseling Psychologist,* 31, 420-426.

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology*, 43, 1205-1228.

van de Vijver, F.J.R. (1998). Towards a theory of bias and equivalence. In ZUMA (Centrum fur Unfragen Methoden und Analysen)-Nachrichten Spezial Band 3: Cross-Cultural Survey Equivalence (pp. 41-65). Retrieved from http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten_spezial/zn-sp-3-

van de Vijver, F.J.R., & He, J. (2016). Bias assessment and prevention in noncognitive outcome measures in context assessments. In S. Kuger, E. Klieme, N. Jude, & D, Kaplan (Eds.), *Assessing contexts of learning world-wide extended context assessment framework and documentation of questionnaire material* (pp. 229-253). New York, NY: Springer.

van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage Publications.

van de Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29-37.

van de Vijver, F.J.R., & Poortinga, Y.H. (2005). Conceptual and methodological issues in adapting tests. In R.K. Hambleton, P.F. Merenda, C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J.R. (2016). The stability of extreme response style and acquiescence over 8 Years. *Assessment*, 23, 279-291.

Ægisdóttir, S., & Gerstein, L.H. (2009). Beliefs About Psychological Services (BAPS): Development & psychometric properties. *Counselling Psychology Quarterly*, 22, 197-219.

Ægisdóttir, S., Gerstein, L.H., & Çinarbas, D. (2008). Methodological issues in cross-cultural counseling research: Equivalence, bias and translations. The Counseling Psychologist, 36, 188-219.

Ægisdóttir, S., Gerstein, L.H., Leung, S.A., Kwan, K.L.K., & Lonner, W.J. (2009). Theoretical and methodological issues when studying culture. In L.H. Gerstein, P.P. Heppner, S. Ægisdóttir, S.A., Leung, & K.L. Norsworthy (Eds.) I*nternational handbook of cross-cultural counseling: Cultural assumptions and practices worldwide* (pp. 89-109). CA: Sage Publications.