African Journal of Business Management Vol. 5(24), pp. 10038-10043, 14 October, 2011
Available online at http://www.academicjournals.org/AJBM
DOI: 10.5897/AJBM11.687
ISSN 1993-8233 ©2011 Academic Journals

Review

The usefulness of measurement equivalence in psychological evaluation: A meta-analysis

Herbert Kanengoni

University of Fort Hare, School of Business and Enterprise, Private Bag, X1314, Alice, 5700, Republic of South Africa.
E-mail: hkanengoni@gmail.com.

Accepted 15 June 2011

"The greatest social benefit will come from applied research if we can find for each individual the treatment to which he can most easily adapt. This calls for the joint application of experimental and correlational methods" (Cronbach as cited by Witcherts, 2007). Recent developments in experimental industrial psychology concerning the effects of individual differences on test performance and efforts to increase validity, reliability and minimise bias have contributed to the understanding of the nature of group differences in achievement and intelligence test scores. This has been made possible by the introduction of measurement equivalence (ME) techniques that are used in psychometric evaluation.

Key words: Measurement, equivalence, invariance, psychological evaluation.

INTRODUCTION

Recent developments in experimental industrial psychology concerning the effects of individual differences on test performance and efforts to increase validity, reliability and minimise bias have contributed to the understanding of the nature of group differences in achievement and intelligence test scores (Witcherts, 2007). This has been made possible by the introduction of measurement equivalence (ME) techniques that are used in psychometric evaluation.

ME, also referred to as metric equivalence, differs from conceptual equivalence and psychometric equivalence (PE). Conceptual equivalence (CE) refers to similarity in meaning of items across language versions of an instrument, and aims to ensure that different language versions measure the same construct. PE refers to similarity in psychometric properties such as floor and ceiling effects, reliability and constructs validity. CE and PE are prerequisites for ME but do not necessarily ensure ME. To date, few studies have investigated ME for different language versions of instruments (Luo et al., 2003).

The study of measurement invariance (MI) across groups originated from the need to study the fairness of cognitive tests in education and in personnel selection, but the psychometric framework of MI can be quite helpful in the study of various psychological phenomena in other settings. Each of these applications of measurement invariance raises specific statistical and psychometric issues that may direct the future technical advancements in the field. In this paper, focus is on the illustration of how tests of ME may greatly enhance the understanding of experimental results in psychological evaluation since this phenomenon is not common in many African nations. However, an overview of the ME phenomenon is necessary.

There are various groups that are usuallually taken into cognizance in study analyses, for example different cultures, regions or countries. In order to compare relationships between constructs or means across groups, we need certain level of equivalence of the constructs across those groups. The meaning of equivalence is whether or not, under different conditions of observing

Abbreviations: ME, Measurement equivalence; **PE,** psychometric equivalence; **CE,** conceptual equivalence; **MI,** measurement invariance; **ANOVA,** analysis of variance; **FI,** factorial invariance; **SEM,** structural equations modeling; **CFA,** confirmatory factor analysis; **CI,** configural invariance; **SI,** scalar invariance; **PLS,** partial least squares.

and studying phenomena, measurement operations yield measures of the same attribute. Meredith (1993) defined ME as the condition where individuals with equivalent true scores would have the same probability of a particular observed score on an associated test.

ME is defined in many different ways. It is also referred to as MI, but in either case it pertains to the consistency of measurement across some specified group demarcation. The groups can be divided in any fashion: by time, nationality, ethnicity, gender, or any number of other factors. The focus of ME is to evaluate the lack of variance between the measurements used in the context of two groups. Put simply, ME aims to ensure that "the same attribute must relate to the same set of observations in the same way in each group" (Ellis et al., 2006).

According to Roe (2006), measurement procedure is equivalent if it produces measurements of a variable X with identical measurement properties in two or more samples that differ with respect to an attribute *a*. In another definition by Trimble (2007), the term ME is said to refer to the possibility that interpretations of psychological measurements, assessments, and observations are similar if not equal across different populations. In order to compare effect parameters across populations, at least ME or MI of the factor loadings between items and theoretical constructs is needed. Davidov, (2008) argues that a higher level of equivalence is needed for comparing means across groups.

Wicherts (2007) pointed out that since psychological constructs are not directly observable; researchers have to choose operationalisations of these constructs in their experimental studies. The data from psychological experiments are normally submitted to analysis of variance (ANOVA) and the mean effects on the observed scores are often treated as if they accurately reflect mean differences in the underlying construct. However, the experimental manipulation may also affect the measurement characteristics of the measure of the experimental effect, which will result in a violation of ME across conditions. For instance, in a study on the cognitive behavioral treatment of anxiety, a self-report measure of anxiety may show a reactive self-report change (Shadish et al., 2002), thereby rendering the interpretation of the experimental findings problematic.

MI assures that the test measures the same construct across groups; for instance testing of job satisfaction in males and females, task conflict in different industries or time perspective in different countries. MI is a starting point to understand the nature of group differences in test scores. So in order to make a meaningful comparison, instruments must have similar measurement qualities across settings. However, it is important to note that ME is a much more complex issue that cannot be fully resolved (Shadish*,* et. al., 2002).

Psychologists typically like to conclude something about psychological properties based on measurements. When the truth-value of such a conclusion depends on the scale that is chosen (a 'permissible' transformation that leaves intact the qualitative relations), it is called an empirically meaningless or illegitimate inference. This is why conclusions based on parametric statistics performed on ordinal scores are considered uninterruptable in terms of the underlying psychological property. Although, such tests are strictly meaningless, they do allow for unambiguous interpretation of group differences under certain circumstances (Borsboom and Dolan, 2007).

ME is the broader concept that subsumes factorial invariance (FI). Like the point and the plane, it is a mathematical abstraction because in reality there are only approximations. It is incumbent on each measurement theory to provide a methodology for demonstrating ME (Bontempo, 2006).

FORMS OF EQUIVALENCE

Depending on which properties are identical, equivalence can be distinguished in different forms. In the traditional literature several distinctions are used, for instance:

1. Hui and Triandis (1985): Conceptual / functional, operational, item, and scalar equivalence.
2. Van de Vijver and Leung (1997): Construct, structural, measurement unit (scalar) equivalence. In the recent literature the focus is on various degrees of equivalence as assessed by successive structural equations modeling (SEM) on items statistics.
3. Steenkamp and Baumgartner (1998): Configurational, Metric, Scalar, Factor variance, Error equivalence / invariance.
4. Vandenberg and Lance (2000): Covariance, Configural, Metric, Scalar, Unique variances, Factor variances, Factor covariances, Factor means equivalence / invariance

Various techniques have been developed to test measurement invariance. The forms of equivalence and their meanings adopted in this paper are based on a particular recent article by Davidov (2008).

Configural invariance

This is the lowest level of invariance. Configural invariance (CI) requires that the items in the measuring instrument exhibit the same configuration of loadings in each of the different populations. That is, the confirmatory factor analysis (CFA) thus, confirms that the same items measure each construct in all populations in the cross-group study.

CI is supported if a single model specifying which items measure each construct fits the data well, all item loadings are substantial and significant, there are no large modification indices, and the correlations between the factors are less than one. The latter requirement guarantees discriminant validity between the factors (Steenkamp and Baumgartner, 1998).

Measurement invariance

CI does not ensure that the people in different groups understand the items in the same way. The factor loadings may still be different across groups. The test of the next higher level of invariance, 'measurement' or 'metric' invariance requires that the factor loadings between items and constructs are invariant across groups. It is tested by constraining the factor loading of each item on its corresponding construct to be the same across groups. MI is supported if the model cannot be significantly improved by releasing some of the constraints.

Partial measurement invariance

However, for cross-group comparison to be allowed, it is not necessary that all factor loadings are equal. Several scholars have suggested that it is enough to have two equal factor loadings per construct across groups to allow comparison of effects. They termed it partial measurement (metric) invariance (Steenkamp and Baumgartner 1998).

Scalar invariance

A third level of invariance is necessary to allow mean comparison of the underlying constructs across populations. This is often a central goal of cross-group research. Such comparisons are meaningful only if 'scalar' invariance (SI) of the items is ensured. SI guarantees that cross-group differences in the means of the observed items are a result of differences in the means of their corresponding constructs. To assess SI, one constrains the intercepts of the underlying items to be equal across countries

Meaningful comparison of construct means across groups requires three levels of invariance, configural, metric, and scalar. Meaningful comparison of relationships between constructs requires two levels of invariance, configural and metric. Only if all these types of invariance are supported can we confidently carry out comparisons. However, commentators such as Roe (2006) argues that limitation of current methods is that they focus on the measurement instrument and ignore the 'ceteris paribus' clause, that is, they fail to acknowledge that there can be differences in the measurement procedure, and that samples can differ in other attributes. Confidentiality of mailed survey may / may not be trusted. In any country, differences can exist in age, gender, education, job type, industry, culture, language. On the other hand, samples from three different countries may be similar in gender, age, job level, industry and urbanity, and yet differ in tenure, employability, standard of living, quality of life, or ethnic homogeneity – while factor structures appear to be identical. In this case, the conclusion that ME of job attitudes exists may be wrong. We would "find the right result for the wrong reasons" (Roe, 2006).

REASONS FOR DIFFERENCES

In trying to produce results that can be generalised across groups researchers have made effort to curb distortions of results brought about by the differences amongst individuals groups. The differences amongst populations which make research a very difficult process especially cross-group research are in most cases non-modifiable factors such as psychological, sociological and linguistic factors. These are discussed in the following paragraphs.

Psychological

This paper cannot assume that answers to questions are given without any cognitive processing at the side of the subjects. Questions have to be interpreted, personal experiences have to be evaluated, answers have to be generated and expressed. Each of these processes is likely to be influenced by some standard ('frame of reference')

Sociological

All of this inevitably depends on the societal context in which people are embedded and on their life history (in interaction with gender and ethnic features). In addition, the context in which the questions are asked and the way this is done may also play a role.

Linguistic

Moreover, in case of cross-cultural studies there will also be an influence of the language in which questions are posed, the quality of the translations, and the language mastery of the respondents.

WHEN SHOULD MEASUREMENT EQUIVALENCE BE USED?

ME is more prevalent in psychology research than in other disciplines. Interest in ME recently increased because of controversy over issues like IQ score differences among ethnic and economic groups. ME becomes especially important and politically charged in these applications as researchers attempt to determine the fairness of research results that attribute relative strengths or weaknesses among groups. Critics are rightly concerned that constructs have the same meaning for all groups under investigation. ME becomes critical when measuring latent variables. Since, by definition, latent variables cannot be directly measured, any variance in measurement of constructs like depression across groups can have a substantial impact upon other things. An oversight in construct development and measurement could lead to an improper (and potentially dangerous) diagnosis for many variables. As a result, the psychology field in particular has come to realize that ME is a prerequisite for making any meaningful comparisons across groups.

In order to make a meaningful comparison, instruments must have similar measurement qualities across settings. Research on measurement equivalence has typically focused on internal structure of multi-item instruments and external relations. Yet, there are more aspects to consider. ME is a much more complex issue that cannot be fully resolved. Equivalence testing is designed to evaluate the comparability between groups. There are several study designs and statistical methods that can be used to assess the comparability of measurement obtained on two (or more) different occasions. According to Shadish et al., (2002), there are two recommended study designs the randomized parallel groups design and the randomized crossover design. Also the study sample should be representative of the intended group in which the test will be used, particularly in regard to age, gender, race/ethnicity, education, and disease severity. Group comparisons must measure identical concepts across groups to be valid.

A key factor in the use of ME is the determination of when it must be addressed. This reduces to the challenge of having to decide when potential bias matters and when it does not. One situation when it is always worth considering the implications of ME is when the research in question is comparing means between groups. When investigating possible bias the researcher must consider biasing effects relative to the effect being tested to appropriately apply ME concepts (Borsboom and Dolan, 2007). A researcher conducting a cross-group study with political overtones, such as any anticipated results that may shift power in an organisation, would be well advised to formally validate the ME status of the constructs. Any situation where results may be called into question is a potential application of ME assessment.

A comparison of within-group measures is typically not a validity threat. However, the effect of the measurement bias can depend upon the source of the biasing effects. An unexpected latent attribute due to multidimensionality (for instance, multiple common meanings of a word) can introduce bias within a group. In this situation, a measurement equivalent assessment would be appropriate.

IMPORTANCE OF USING MI/ME IN PSYCHOLOGICAL TESTING

ME revolves around the issue of how groups differ in the way the measurement of a psychological construct (for instance, mathematics test score) is related to that construct (for instance, mathematical ability). ME means that measurement bias with respect to groups is absent (Witcherts, 2007). ME with respect to groups is an essential aspect of the fair use of scores of psychological tests and other psychological measurements. The valid and fair use of psychological tests in psychology, education, and other settings requires that tests measure what they are supposed to measure, and that test scores are not affected by irrelevant characteristics associated with membership of demographic groups (for instance, ethnicity and gender).

Whenever groups differ in this relation, we speak of measurement bias. Clearly, measurement bias complicates the comparison across groups of test scores. On the other hand, when evaluation is characterised by the same measurement properties over groups, the paper speak of ME. ME is an interesting ideal, but it does not always arise in real data (Witcherts, 2007).

When psychological and work-related constructs are measured in a cross-cultural context, it is pivotal to establish equivalence of the measures prior to drawing meaningful substantive conclusions about the relative importance of constructs across groups (Beuckelaer et al., (2007). ME with respect to groups is an essential aspect for interpreting group differences in scores of any kind of psychological measurement. Tests for ME enable one to differentiate between group differences in the latent constructs that a certain test is supposed to measure that is real ability differences, and measurement artifacts related to group membership.

Lack of ME in data across populations implies that there is no common basis to compare data across populations. In such case, observed differences on relevant constructs (across groups) might result from measurement artifacts related to the measurement instrument used rather than from true differences across groups. Establishing measurement equivalence enables us to answer a series of important questions such as: Do respondents in different groups use a similar frame-of-reference when answering items used to measure relevant constructs? Do respondents in different countries calibrate the intervals on the measurement scale used in similar ways? Are differences in response styles across groups (for example, the tendency to say 'yes' or to use extreme response categories) partly responsible for observed cross group differences in scores?

Beuckelaer et al. (2007) posited that people's general values influence their work-related goals and values and that; therefore, these values serve as a frame-of-reference against which they define their work-related experiences. As surveys ask for reports on work-related experiences, it can be assumed that individuals with different values will not always use the same frame-of-reference when completing survey items. For instance, an organisational concept such as privacy might have a different conceptual meaning across cultures. In cultures where individuals are strongly embedded in social groups, items related to privacy might be interpreted differently than in cultures wherein individuals are more autonomous. Clearly, such differences in the conceptual domain for interpreting survey items might decrease the ME of global surveys. Therefore, this depicts the importance of establishing ME during these cross group psychological evaluations.

Conversely, similarity in cultural values might increase the use of similar conceptual domains when completing global surveys, resulting in ME of the scales used in the survey. Apart from affecting the conceptual frame-of-reference used, cultural values might also influence how individuals interpret the rating scale (Steenkamp and Baumgartner, 1998). Specifically, prior cross-group research has shown that the differences between the intervals of a rating scale are differently perceived across cultures. In fact, substantial cross-country differences have been found with regard to the tendency to agree with items, regardless of the item content (Van Herk et al., 2004).

Similarly, Poortinga and Verhallen (2004), further argue that there is empirical evidence for cross-country bias due to the respondents' use of extreme responses on rating scales as this bias exists between Korean and American respondents, Japanese and American respondents and French and Australian respondents (Clarke, 2000). Such cross-country differences in response styles produce systematic differences in observed variable means and variances. As a result, the assumption of ME of survey instruments may not be tenable. Although, these prior studies were not conducted in an organisational context, they might have direct implications for organisational surveys because the latter also use rating scales.

ME is also important in psychological evaluation because it defines the nature of latent constructs associated with the observed scores and articulate assumptions that permit estimation of latent psychometric properties from observed data. Consistent with the dictum, "all models are incorrect, some models are useful", and a useful measurement model impose few restrictions, provides a basis to assess construct validity, and provide a means to assess bias and equivalence (Bontempo, 2006).

However, Wicherts (2007) posits that researchers have been slow to investigate the possibility of measurement bias in their between-group research. One motivation for an increased awareness of ME is the increased use of partial least squares (PLS) to analyze differences across groups. Also the true coefficients become confounded with the factor loadings under PLS. Most ME issues that arise in research are seen in a factor analysis context. Potential bias in measurement is usually implicitly assumed away during factor analysis or treated as random error. This is not through any overt action on the part of researchers, but rather because of an unawareness of the importance of addressing ME issues before collecting data.

As previously mentioned by Wicherts (2007), researchers are most likely to need to perform an ME evaluation in the context of a CFA. The sequence is a compilation of procedures recommended by authors of ME articles in all disciplines. Evaluators regularly find gender differences in many constructs which contends that the equivalence of the measures across groups is typically assumed and is seldom checked. As a result, an assessment of ME is necessary for meaningful between group comparisons in psychological evaluation. There are possibly hazards of ignoring ME in psychological evaluations. This prior assessment is needed to ensure that the evaluation is measuring the same constructs.

In psychological evaluation, the concept of ME is critical and should be applied, let alone as a regular technique. This can have the effect of rendering much of our research interpretable. According to Witcherts (2007), it is unfortunate that in many applications MI is assumed to hold without testing for the equality over groups of measurement intercepts. The equality of measurement intercepts across groups is essential for MI, what group differences in intercepts may mean, and how these differences can be detected. If the intercept of a particular subtest is different across groups, this implies that between-group differences on this subtest cannot be solely due to between-group differences in the construct(s) that the subtest is supposed to measure. In other words, an intercept difference indicates measurement bias in the sense there are one or more construct-irrelevant variables causing group differences in test scores.

Considering the widespread use of achievement and intelligence tests in college admission and job selection, and the high stakes involved in their use, stereotype threat effects on test performance may have serious personal and social consequences. There is general agreement on the importance of fair, unbiased, assessment in the sense that individual latent abilities should be measured validly and accurately. This means that measurements of ability should not depend on group membership based on, for instance, ethnicity or sex. Therefore, the absence of measurement bias with respect to groups that is, ME is an essential aspect of valid measurement in psychological evaluation (Millsap and Everson, 1993).

The remainder of this paper is organised as follows. In the next section, I address the question of the possibility of dealing with equivalence. After that, then I present the conclusion.

DEALING WITH EQUIVALENCE

There are various ways that have been put across by researchers and evaluators in order to deal with equivalence. Firstly evaluators are advised to assume that equivalence is present, unless it can no longer be denied. Also the evaluators should demonstrate that equivalence is present. That is test the 0-hypothesis that the difference on a given attribute is zero. In organizational research, the first option is common practice, but two is to be preferred (Wicherts, 2007).

CONCLUSION

Within this approach equivalence will be an exception than a rule, partial rather than complete especially in organisational context. This is by virtue of the fact that there are many cases in which nonequivalence is faced in organisational context. Acknowledging non-equivalence will reduce the meaning and value of comparisons. It will force us to put the universality assumption running through much of theories into question. Although, (re)design and correction can raise equivalence, we have to live with non-equivalence (Roe, 2006). Recent studies suggest that when full or partial measurement invariance is not guaranteed, it may still be the case that constructs are equivalent. Davidov (2008) indicates that the test of ME is too strict and may fail although, cognitive equivalence still holds. An alternative procedure would be to conduct cognitive pretests in the different populations. When these procedures are not feasible, testing for equivalence should be done. Also across-time comparisons are very useful in ensuring ME in psychological instruments. MI assures that the test measures the same construct across groups. MI is a starting point to understand the nature of group differences in test scores.

REFERENCES

Beuckelaer AD, Lievens F, Swinnen G (2007). Measurement equivalence in the conduct of a global organizational survey across countries in six cultural regions. J. Occup. Organ. Psychol., 80: 575-600.

Bontempo DE (2006). Measurement Equivalence/ Invariance of the Developmental Behavior Checklist; Factorial Invariance of Categorical Factor Models. Proceedings from ISSBD. The 19th Biannual Meet. International Society for the Study of Behavioural Development, Melbourne, Australia.

Borsboom D, Dolan CV (2007). Theoretical equivalence, measurement invariance, and the idiographic filter. Measurement, 5(5): 236-263.

Clarke I (2000). Extreme response style in cross-cultural research: An empirical investigation. J. Soc. Behav. Personal, 15: 137-152.

Davidov E (2008). Measurement Equivalence of Nationalism and Patriotism in 34 Countries in a Comparative Perspective Gesis-Central Archive for Empirical Social Research, University of Cologne.

Ellis ME, Aguire-Urreta MI, Nan Sun W, Marakas GM (2006). Establishing the Need for Measurement Invariance in Information Systems Research: A Step-By-Step Example Using Technology Acceptance Research.

Hui CH, Triandis HC (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. J. Cross Cult. Psychol., 16: 131-152.

Luo N, Chew L, Fong K, Koh D, Ng S, Yoon K, Vasoo S, Li S, Thumboo J (2003). Do English and Chinese EQ-5D versions demonstrate measurement equivalence? An exploratory study. Health Qual. Life Outcom., 1: 7.

Meredith W (1993). Measurement equivalence, factor analysis and factorial equivalence. Psychometry, 58: 525-543.

Millsap RE, Everson HT (1993). Methodology review: Statistical approaches for assessing measurement bias. Appl. Psychol. Measure., 17: 297-334.

Roe RA (2006). Proceedings from ITC. ITC Conference, Maastricht University, Brussels.

Shadish WR, Cook TD, Campbell DT (2002). Experimental and quasi-experimental designs for causal inference. Boston, MA: Houghton Mifflin Company.

Steenkamp JEM, Baumgartner H (1998). Assessing measurement invariance in cross-national consumer research. J. Consum. Res., 25: 78-90.

Trimble JE (2007). Cultural measurement equivalence. In C. S. Clauss-Ehlers (Ed.), Encyclopedia Cross-culture School Psychology, New York: Springer.

Van de Vijver F, Leung K (1997). Methods and data analysis for cross-cultural research. London: Sage.

Van Herk H, Poortinga YH, Verhallen TMM (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. J. Cross-Cult. Psychol., 35: 346-60.

Vandenberg RJ, Lance CE (2000). A review and synthesis of the measurement equivalence literature: Suggestions, practices, and recommendations for organisational research. Organ. Res. Method, 3(1): 4-70.

Wicherts JM (2007). Group differences in intelligence test performance. Eur. J. Personal., 21: 763-765.