A cross-cultural analysis of the end-user computing satisfaction instrument: A multi-group invariance analysis

Xiaodong Deng a,*, William J. Doll b, Said S. Al-Gahtani c, Tor J. Larsen d, John Michael Pearson e, T.S. Raghunathan b

a Oakland University, 2200 North Squirrel Road, Rochester Hills, MI 48309, USA
b The University of Toledo, Toledo, OH, USA
c Department of Administrative Sciences, King Khalid University, Abha, Saudi Arabia
d President’s Office, Norwegian School of Management, Oslo, Norway
e Southern Illinois University, Carbondale, IL, USA

1. Introduction

The globalization of business has highlighted the need to understand the effectiveness of information systems that span different nations and cultures. Global organizations have a growing need to utilize IT to achieve economies of scale, coordinate global operations, and facilitate collaborative work across distributed locations and diverse cultures. However, cultural differences can make the difference between success and failure in the adoption and implementation of IS. Despite the importance of cross-cultural studies, no studies of the robustness or measurement equivalence of any of the established information system instruments across national cultures exist. A key problem is thus in ensuring valid cross-cultural comparisons of system effectiveness.

Global IT management and cross-cultural research in MIS often involve comparing samples from two or more cultures systems on system success, e.g., user satisfaction/acceptance or system use. A methodological consideration in conducting cross-cultural comparisons is on ruling out alternative explanations for differences and, thus, enhancing the interpretability of results. Any observed differences in mean scores across samples give rise to many explanations. Researchers or IT managers may question whether results are true differences due to culture [39] or merely due to measurement artifacts [38]. To make valid comparisons, the instruments must provide equivalent measurement.

Without this, observed scores from different nations or cultures may not be directly comparable. Examining whether scales are comparable is thus necessary to improve our understanding of how culture affects IS success and better manage technology in a global context.

User satisfaction has become a pervasive measure of the success or effectiveness of information systems for both managers and researchers. Originally developed by Doll and Torkzadeh [12] to measure a user’s satisfaction with a specific application, the end-user computing satisfaction (EUCS) instrument (12-item summed scale and five factors) has been widely used and cross-validated [18,36,43]. While the original instrument is relatively old, the item-factor loadings for the 12-item instrument have been remarkably stable. Gelderman [20] found...
that EUCS was a good predictor of an application’s impact on organizational performance and, thus, a useful surrogate for system success.

The EUCS instrument has been validated for today’s popular enterprise wide applications [41] that employ standardized software modules with user customizable interfaces. Such applications (e.g., SAP, Oracle, and PeopleSoft) are commonly used by global firms’ users in a variety of national cultures.

The EUCS construct is defined as a second-order latent factor consisting of five first-order latent factors (i.e., content, format, accuracy, ease of use, and timeliness). The five first-order latent factors and their structural weights define the meaning of the second-order EUCS construct. The structural weights indicate how central each of the first-order latent factors is to the meaning of the construct [33].

The EUCS instrument’s 12-item scale was recently found to provide equivalent measurement (e.g., item-factor loadings) across a wide variety of application types, respondent positions, and hardware platforms [16]. However, the structural weights (loadings of first order factors on second-order EUCS factor) for accuracy and ease of use were found to be variant across user groups and types of applications. Thus, the study found that the EUCS 12-item summed scale and the five component factor scales provided robust (equivalent) measurement across a variety of target populations.

Studies of the cross-cultural measurement equivalence of IS instruments can complement the research of those focusing on how culture affects the design, adoption, implementation, and on-going use of global IS. Tan and Gallupe [42] argued that instruments that are developed and validated in the traditional IS field can be and should be applied to and validated across the global context. The wisdom of the global IT manager’s decision to allocate resources to diagnosis and intervention may depend upon robust instruments like EUCS. If the instrument is robust, low user satisfaction scores may indicate a training or implementation process that is poorly designed for a particular location.

Despite the wide use of the EUCS instrument, its validity and measurement equivalence across different cultures has not been examined. Studies that explore the measurement equivalence across national cultures will enhance interpretability of research. Based on five independently collected samples from the US, Taiwan, Western Europe, India, and Saudi Arabia, we used multigroup invariance analysis to explore the degree to which the EUCS instrument provided measurement equivalence across national cultures/world regions.

2. The importance of measurement invariance across culture

If the Director of MIS of an international firm can compare user satisfaction scores for an application across cultures, he or she may be able to identify problems areas where user satisfaction is low, diagnose the nature and cause of the problem, and take corrective action. Yet, there are many well-known problems associated with conducting cross-national research [21].

When designing cross-cultural studies to evaluate IS, both researchers and practitioners need to know whether user satisfaction has equivalent measurement across national cultures in their firm or sample. Evaluation and diagnostic instruments that provide equivalent measurement across national cultures are particularly valuable to multinational firms who seek to leverage technology across countries/cultures. Such studies later work in benchmarking [15].

The global management of IT requires comparative studies of both the systems development process and the post-implementation evaluation of IS. The development process is complex and highly context dependent. Ives and Jarvenpaa [28] contended that the context of the development process depended on the firm’s global business strategy, the IS platforms in each country, international data sharing regulations, and cultural differences. Studying the systems development process involved behavioral observations of a dynamically changing process (e.g., user participation, stages, and methods) that may have different meanings across cultures. Cultural studies may therefore be challenging and sometimes the interpretation of results may not be equivocal.

The ongoing evaluation of user satisfaction with a system occurs in a more stable context (routine work flow); the system has often had to be adapted to meet local requirements, business goals, hardware platforms, or infrastructures; users have had to gain experience in using the systems to meet their needs. In this more stable context, lack of measurement equivalence may be attributed to culture.

A fundamental, unresolved issue with multinational research is whether similarities or differences are, in fact, real [4]. Standardized instruments must provide equivalent (invariant) measurement across national cultures (equal true scores) if comparative statements across cultures are to have substantive import. Drasgow and Kanfer [19] argued that without equivalent measures, observed scores from different cultures were on different scales and, therefore, were not comparable.

For example, to compare EUCS scores from India and the US, the 12 items comprising the EUCS instrument must have the same amount of trait or true scores in both countries. The formula relating trait and error for each item of the 12 items is

\[ x = \lambda \xi + \epsilon \]

where \( x \) is the observed score for the item on a 1–5 Likert scale, \( \lambda \) is slope of the regression of the observed score on the true score, \( \xi \) is the true score, and \( \epsilon \) is the error term. For observed scores in India and the US to be equivalent, \( \lambda \) values for each of the 12 items must be statistically equivalent for both India and the US (i.e.,

\[ \lambda_{1, \text{India}} = \lambda_{1, \text{US}}, \lambda_{2, \text{India}} = \lambda_{2, \text{US}}, \ldots, \lambda_{12, \text{India}} = \lambda_{12, \text{US}}. \]

If this rather stringent condition is not met, we do not know whether an observed difference for the 12-item summed scale of the EUCS instrument is real or a measurement artifact of cultural differences between the countries.

If the \( \lambda \)s for India are systematically smaller than those for the US, the mean score on the 12-item summed EUCS scale would tend to suggest that the respondents from Indian respondents have lower user satisfaction than US respondents, even if both groups have the same true satisfaction level. Correlations between EUCS 12-item scores and correlates such as usage (e.g., hours, extent, etc.), training, or support would also appear to be lower in the Indian sample than in the US sample. MIS managers with global responsibility might mistakenly conclude that satisfaction has a lower correlation with usage in India than in the US. These managers might also mistakenly conclude that user training or support is not as effective at improving satisfaction in India. However, these lower correlations would be an artifact of cultural differences in measurement. These issues were discussed further by Bollen [6].

2.1. Research on the EUCS instrument

The EUCS instrument has been used in different cultural contexts: by Igbaria and Tan [27] in Singapore, by Doll and Torkzadeh [13] in the US, by McHaney et al. [37] in Taiwan, by Gelderman in the Netherlands, and by Al-Gahtani and King [1] in Great Britain. However, we know little about the accuracy of the EUCS for cross-cultural system evaluations in enterprise wide applications across the globe. The question of cross-cultural measurement invariance is not whether national culture affects
mean scores, but rather whether it affects the trait (true scores) in the scale’s items. Differences in true scores (e.g., item-factor loadings) between cultures mean that the observations are in different frame of reference and therefore not comparable. Without knowledge of whether the instrument provides equivalent measurement across cultures, it is not possible to determine whether observed differences in user evaluations are due to culture or to a non-equivalent measurement instrument.

In 2004, Doll et al., using a US sample, demonstrated the stability of the 12-item-factor loadings and thus the 12 item summed scale for measuring user satisfaction. The 12-item EUCS instrument was found to provide equivalent item-factor loadings across subgroups representing type of application, position of respondent, mode of development, and platform (personal computer versus mainframe). This stability across subgroups within a single culture was the stimulus to test for measurement invariance across cultures.

In their study, Doll et al. found differences in structural weights for accuracy across population subgroups representing respondent position and type of application. This suggested that the meaning of user satisfaction could be context sensitive and differ across population subgroups.

2.2. National culture as a frame of reference for user satisfaction

Smith et al. [40] suggested that the frame of reference of an individual in evaluating an application will be shaped by his or her life experiences. Measurement nonequivalence may result from these differing frames of reference. Zmud et al. [44] argued that all instruments could be viewed as located on a continuum reflecting the extent to which the construct is linked with an experientially based context, such as a national culture. If significant construct-context interaction is present, item-factor loadings vary between subgroups or contexts. Thus, the cultural context may, in part, influence the construct’s meaning or how it is scaled.

Hofstede [25] argued that national cultures represent fundamental differences in the way people perceive and interpret the world; they are frames of reference that are grounded in lifetime experience – family, friends, school, etc. – of individuals as they grow and interact with others who share the same values. Thus culture can introduce a bias in user evaluation of IS.

National culture dimensions may represent a frame of reference that introduces a bias in IS evaluation. Dawar et al.’s [11] research on information exchange and culture suggested that national culture (power distance and uncertainty avoidance) bore an inverse relation to a society’s general openness to objective sources of new information. Since cultures low in uncertainty avoidance take a more empirical approach to understanding and knowledge [24], they may respond more favorably when evaluating ISs than cultures high in uncertainty avoidance. Cultures with high power distance scores tend to be hierarchical. Empirically driven sources of knowledge may be evaluated more favorably in cultures that are not oriented toward absolute perceptions of authority.

2.3. The EUCS measurement instrument

Fig. 1 shows EUCS as a single second-order latent construct with five first-order latent factors. Confirmatory studies by Doll and Xia [14], Doll et al. in 1994, Kim and McHaney [31], and McHaney et al. have validated this model. Studies of the instrument’s test-retest reliability by Hendrickson and Glorfeld [23], McHaney and Hightower, and Torkzadeh and Doll have reported good stability and reliability, as indicated by Cronbach’s alpha values above 0.90. Since this second-order measurement model has been supported by previous studies that utilized and tested the EUCS instrument, we chose the second-order measurement model [8] to test the invariance of the instrument across national cultures/world regions.

In Fig. 1, the five arrows leading from end-user computing satisfaction to the five first-order latent factors depict the structural weight, which can be viewed as regression coefficients in the first-order factors on the higher-order factor. Marsh and Hocevar explained that these structural weights indicated the centrality or importance assigned to the first-order factors in scaling the second-order factor. The weights can be used to derive the second-order EUCS score from the weighted average of the first-order factor scores.

The arrows leading from the first-order latent factors to the measurement items are the item-factor loadings, which can be viewed as regression coefficients in the regression of observed variables on their corresponding latent factor. These loadings indicate the extent that the item captures the trait of its latent factor. In different contexts or sub-populations, the item-factor loadings may vary, reflecting the cultural context in which the evaluation is made.
loadings may be different, suggesting that the first-order factors may have different meanings across subgroups.

2.4. The hypotheses

The most common method for scaling EUCS is a simple aggregation (or average) of the scores for the 12 items. Substantial differences in item-factor loadings between groups would indicate that EUCS may have to be scaled differently (different items or different item-factor loadings) and comparisons across groups would not be possible. If item-factor loadings are invariant across population subgroups, EUCS scores obtained by summing the 12 items are taken in the same context and, therefore directly comparable across subgroups. Therefore, we proposed the hypothesis:

H1. The 12 items of the EUCS instrument have equivalent item-factor loadings on their corresponding first-order latent factors across cultures (i.e., US, Western Europe, Saudi Arabia, India, and Taiwan).

A less frequently used and more complicated, but potentially more accurate method for scaling EUCS is to use the second-order factor score. This is computed by multiplying each first-order factor score by its corresponding structural weight and summing the weighted factor scores across the five factors. This score does not use arbitrary weights. The weights reflect the centrality or importance of the first-order factors to the second-order EUCS factor. If the structural weights differ across subgroups, EUCS factor scores may provide a more accurate measure of satisfaction for each subgroup, but the scores may not be comparable across subgroups. The invariance of EUCS second-order factor scores requires that the structural weights as well as the item-factor loadings be equivalent across national cultures. Thus, we proposed a second hypothesis:

H2. The five first-order latent factors have equivalent structural weights on the second-order EUCS factor across cultures (i.e., US, Western Europe, Saudi Arabia, India, and Taiwan).

H2 is based on the assumption that the item factor loadings are equivalent. Thus H2 can not be tested if the item factor loadings are not equivalent across national cultures.

3. Research methods

Given the many costs associated with collecting the sample data, strict confirmatory analysis is not commonly used [9]. We used structural equation modeling in a model generating mode to explore which item-factor loadings and structural coefficients of the EUCS measurement model were equivalent across countries or regions.

Mullen recommends multi-group invariance analysis as the preferred approach to diagnosing measurement equivalence in cross-cultural research. Data analysis using this approach is based on independent covariance matrices of subgroups; it does not require mixing independent samples. In examining the measurement equivalence of the EUCS instrument across population subgroups, two sets of parameters were of special interest:

- the equivalence of item-factor loadings for the 12 items; and
- the equivalence of the structural weights of the five first-order factors on the second-order EUCS factor.

Testing for equivalence is a particularly demanding test of robustness. In some cases, minor differences may not be critical to the interpretation of research results.

We used individual users within national cultures to infer the influence of cultural differences. While we recognize that individuals will have differences in personal values, attitudes towards power distance, etc., multi-group invariance analysis was used to test for differences between groups in item-factor loadings which represented the amount of trait in each item and can be interpreted as the regression coefficient or the slope of the regression of true scores on observed scores. The most plausible cause of variance between groups is national culture rather than individual differences, which may cause differences in individual mean scores on the EUCS instrument, but are unlikely to explain differences in the amount of trait in each item.

The primary objective of our research was not to isolate any individual or subculture differences when applying the EUCS instrument across cultures. Global IT managers who wish to make comparisons between user satisfaction scores will not be able to remove all sources of variation. What they need is some reasonable assurance that differences in scores are not an artifact of national culture.

Hofstede [26] addressed this issue. He argued that in order to be a meaningful subject for the study of culture, a group needed only be reasonably homogenous with regard to the cultural characteristic studied. He contended that there are within culture differences or subcultures, and argued that data should therefore be collected to include samples from all subcultures or subgroups. Following Hofstede, if the national samples are reasonably representative of their culture, one can test for differences in item-factor loadings across sample groups.

We therefore followed an established modeling approach that has been developed and used in a number of studies in various disciplines to test invariance of second-order measurement models [34].

3.1. The measures

Each study used the items proposed by Doll and Torkzadeh in 1988. The order of the 12 items was randomized in the questionnaire. A five point scale was used: 1 = almost never; 2 = some of the time; 3 = about half of the time; 4 = most of the time; and 5 = almost always. The questionnaires were in English for India, Taiwan, US, and Western Europe subgroups.

For the Saudi Arabia subgroup, the questionnaire was in Arabic. All instrument items were translated professionally from English to Arabic using two groups of three English professors in each group applying the Back-Translation method [7] which converged after three rounds. Item E2 of the EUCS (Is the system easy to use?) was substituted by item EO16 of Davis’s [10] EOU scale (I would find the application easy to use). A seven point scale was used for this group: 1 = never; 2 = occasionally; 3 = about half the time; 4 = most of the time; and 5 = almost always.

3.2. The samples

The data were collected by independent researchers using the standardized 12-item EUCS instrument. Each independent researcher provided a covariance matrix for one national culture or world region. Table 1 shows sample sizes for each subgroup.

<table>
<thead>
<tr>
<th>Groups</th>
<th>n</th>
<th>χ²</th>
<th>d.f.</th>
<th>p-Value</th>
<th>NNFI</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>India (n = 375)</td>
<td>131.11</td>
<td>49</td>
<td>0.000</td>
<td>0.96</td>
<td>0.97</td>
<td>0.067</td>
<td></td>
</tr>
<tr>
<td>Saudi Arabia (n = 1190)</td>
<td>244.05</td>
<td>49</td>
<td>0.000</td>
<td>0.97</td>
<td>0.98</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>Taiwan (n = 342)</td>
<td>198.69</td>
<td>49</td>
<td>0.000</td>
<td>0.90</td>
<td>0.92</td>
<td>0.095</td>
<td></td>
</tr>
<tr>
<td>US (n = 618)</td>
<td>235.47</td>
<td>49</td>
<td>0.000</td>
<td>0.95</td>
<td>0.96</td>
<td>0.079</td>
<td></td>
</tr>
<tr>
<td>Western Europe (n = 123)</td>
<td>89.51</td>
<td>49</td>
<td>0.000</td>
<td>0.94</td>
<td>0.96</td>
<td>0.083</td>
<td></td>
</tr>
</tbody>
</table>
Data were collected from two local organizations who were heavily export oriented with 51 and 324 responses, respectively. The first organization was a manufacturing firm and the second a financial services organization. Senior VPs within both organizations supported the research. The questionnaires were sent to the employees from the Senior VP’s office with a supporting letter to major users of applications with operational or strategic import. Ninety-one percent of the respondents were male and 9% were female. Sixty-eight percent of the respondents had a bachelor’s degree and 27% has a master’s degree or higher. The respondents were using many different applications including a variety of specialized spreadsheet and database applications.

3.4. The Saudi Arabia sample

In Saudi Arabia, both public sector and private sector organizations were contacted using direct mail to request their participation in the study. A formal letter from vice rector of King Khalid University was sent asking for cooperation. This letter was sent to 136 public and private oil, manufacturing, merchandizing, and services organizations. A mix of international and local organizations agreed to distribute the Arabic questionnaire. Several semi-governmental companies which constituted a large portion of the sample could be considered international (Saudi Aramco, Saudi Telecom, Saudi Electric, and Saudi Arabian Airlines).

Organizations who agreed to participate provided the name of a contact person, usually the MIS directors, who identified the firm’s major applications and users. Furthermore, a liaison was established via the contact persons in these companies and the project representatives in the main regions of the country. For organizations that did not respond, the researcher made a follow up via telephone calls, letters, and e-mail. Eventually, 56 organizations participated in the study where 66.4% were public and 33.6% private. The usable responses were 1190 out of 1900 for a response rate slightly over 62%.

Approximately 79% of the respondents were male and 21% were female. The survey respondents were distributed by functional area approximately as follows: 37% personnel, 23% accounting/financial analysis, 21% operations (e.g., production, inventory control, scheduling), 12% marketing/sales, and 7% other work-related applications.

This high response rate can be, in part, explained by the survey being a part of a Saudi government project financed by the first government research department - King Abdul-Aziz City for Science and Technology (KACST).

3.5. The Taiwan sample

The target population was knowledge workers—specifically individuals whose primary work related activities were information-based and required the use of IT to complete them. Representatives from 25 companies participated. The companies were a diverse group including the public sector, manufacturing, consulting, healthcare, transportation and finance. Each representative distributed 20 questionnaires to a randomly selected group of knowledge workers throughout their organization. The respondents were Taiwanese working primarily for multi-national firms.

Data were collected from over 60 firms, about half of those initially contacted. Typically, the Directors of MIS were asked to identify their major applications and the users who directly interacted with each. Major applications were defined as those that were of operational or strategic importance to the firm.

Questionnaires were distributed to end-users through inter-office mail with a cover letter describing the survey as a university-based research project. Six hundred and eighteen (618) responses were obtained. About half of these came from manufacturing firms with the remainder being almost equally distributed among retail, government agencies, utilities, hospitals, and educational institutions. The firms were local rather than international.

The sample represented a variety of applications including accounts payable, accounts receivable, budgeting, CAD/CAM, customer service, service dispatching, engineering analysis, process control, work order control, general ledger, manpower planning, financial planning, inventory, order entry, payroll, personnel, production planning, purchasing, quality, sales analysis and forecasting, student data, and profit planning. Although gender was not specifically asked, the respondents entered their name. Classification of gender by first name suggested that the sample was approximately 35% female.

3.6. The Western Europe sample

A manufacturing division within an international corporation with headquarters in Oslo, Norway, agreed to participate. The division has factories and sales offices worldwide. This sample was limited to employees in ten western European countries. The respondents were users of an Enterprise Resources Planning application (ERPA). At the time of data collection, users had completed their training and had been using the ERPA for at least 4 months.

The division project manager identified and contacted 233 possible respondents and distributed the questionnaire via e-mail. Responses were mailed directly to the researchers, guaranteeing anonymity. After one reminder, 123 completed questionnaires were received, for a response rate of 53%. The final sample consisted of 34 middle managers and 89 super-users; 87 were male and 36 were female.

3.7. The United States sample

Data were collected from 123 completed questionnaires and 233 possible respondents and distributed the questionnaire via e-mail. Responses were mailed directly to the researchers, guaranteeing anonymity. After one reminder, 123 completed questionnaires were received, for a response rate of 53%. The final sample consisted of 34 middle managers and 89 super-users; 87 were male and 36 were female.

3.8. Confirmatory factor analysis

Confirmatory factor analysis was conducted using LISREL VIII [30], a statistical tool for analyzing covariance matrices according to systems of structural equations. The hypothesized second-order measurement model was assessed for model-data fit and item-factor loadings for each of the five subgroups. The examination of goodness-of-fit within subgroups was essential for assessing the issue of congeneric or conceptual equivalence and identifying where, if anywhere, the model did not achieve adequate fit.

Item-factor loadings are estimates of the validity of the observed variables (items). The larger the item-factor loadings – as compared with their standard errors and expressed by the corresponding t values – the stronger is the evidence that the
measured variables represent the underlying constructs. Bagozzi and Yi [3] suggested that item-factor loadings should exceed 0.60.

3.9. Invariance analysis methods

Confirmatory factor analysis (CFA) models of factorial invariance allowed us to examine explicitly the structure of a measurement model or its individual parameters for equivalence across subgroups or conditions. When parallel data existed for more than one group, CFA provided a particularly powerful test of the equivalence of solutions across the multiple groups [2,17]. Tests of factorial invariance across multiple groups involved a hierarchical ordering of nested models. Any two models are nested if the set of parameters estimated in the more restrictive model is a subset of the parameters estimated in the less restrictive model. When models are nested, the difference between the two models can be tested by subtracting the two $\chi^2$ values and testing this value against the critical value associated with the difference in degrees of freedom [32].

This $\chi^2$-test is powerful and, where the hypothesis of equal item-factor loadings or equal structural weights between subgroups is not rejected, Marsh [35] contended that this test provided strong support that observed differences between subgroups were due to chance. If the $\chi^2$ difference was significant, the subjective fit indexes were examined to see how much they decline as invariance constraints are imposed. Small decreases in the subjective fit indexes would suggest that the differences in factor loadings or structural weights are not substantial and are unlikely to effect the interpretation of research results. To evaluate whether model-data fit declines substantially as invariance parameters are imposed, the researchers will examine changes in non-normed fit index (NNFI), comparative fit index (CFI), root mean square error of approximation (RMSEA), and expected cross-validation index (ECVI).

3.10. The sequence of invariance analysis

In multi-group invariance analysis, Joreskog [29] argues that tests of invariance across different dimensions should be done independently. Each hierarchical sequence begins with the least restrictive model in which only the form of the model – the pattern of fixed and nonfixed parameters – is invariant across groups. This initial baseline model is “totally noninvariant” in the sense that there are no between-group invariance constraints on estimated parameters (i.e., “no invariance” model).

This baseline model provides a basis of comparison for all subsequent models in the invariance hierarchy. If it is not able to fit the data, then none of the more restrictive models in the hierarchy will be able to do so. Bollen argues that the failure of the baseline model for a particular subgroup to achieve “good fit”, or at least adequate fit, on subjective fit indexes may preclude any further tests of invariance involving this subgroup. Poor fit in a subgroup suggests that the instrument may not measure the phenomena adequately in this subgroup—a new instrument or measurement model may have to be developed for this particular subgroup.

Next, the equivalence of factor loadings across groups in each dimension is tested (i.e., tau-equivalency). In evaluating measurement models, the primary concern is usually about whether each item is a good measure of its latent construct. Factor loadings are examined first because the equivalence of factor loadings is the minimal condition for “factorial invariance.” Bollen noted that the equality of factor loadings is generally of a higher priority than the equality of other parameters. Bentler [5] suggested testing first for invariance of factor loadings because, without such invariance, it would be difficult to argue that the factors were the same. If the factors are not the same, it may be meaningless to test for the invariance of other parameters.

To test for equal factor loadings, an equal item-factor loading constraint is added to the baseline model, creating a nested or more restrictive model that is a subset of the baseline model. Thus, the significance of $\chi^2$ differences between these two models provides a test of the hypotheses of equal item-factor loadings. If the hypothesis of equivalent item-factor loadings is rejected, alternative hypotheses may have to be formulated and examined to determine the source and nature of the variance in item-factor loadings.

Finally, if the hypothesis of equal item-factor loadings is not rejected, we move on to test for the equality of the structural weights (gammas) across subgroups. This nested or more restrictive model is a subset of the model specifying equal item-factor loading. Thus, the $\chi^2$ difference between these models provides a test of whether all five structural weights are equivalent across the subgroups. A p-value $>0.05$ indicates that the null hypothesis (i.e., no difference in structural weights across subgroups) is not rejected. If the null hypothesis is rejected, we examine the structural weights, generate alternative hypotheses, and explore these alternative hypotheses to identify what structural weights are equivalent across subgroups.

4. Results

Chi-square statistics and subjective goodness-of-fit indices for each of the five subgroups were given in Table 1. A proper solution is obtained for each subgroup. The subjective fit indices for each of the five subgroups suggested adequate model-data fit. In all five subgroups, NNFI and CFI scores were above 0.90 and 0.92, respectively. RMSEA was below 0.08 for all subgroups except for Taiwan (0.095) and Western Europe (0.083). All subgroups have RMSEA scores below 0.1.

The item-factor loadings in the five subgroups indicated that the items were generally good measures of their corresponding first-order latent factors (see Table 2). Of the 60 item-factor loadings, the only one below 0.60 was for item E2 in the Saudi Arabian subgroup (0.32). The structural weights depicted in Table 3 indicated how central each first-order factor was to the EUCS construct. Of the 25 structural weights, the only loading below 0.60 was for the “ease of use” factor (0.49) in the Western European subgroup. Thus, we concluded that, in general, the measurement model adequately fit the data for each subgroup and that we could proceed with the invariance analysis for all subgroups.

### Table 2

<table>
<thead>
<tr>
<th>Groups</th>
<th>Content</th>
<th>Accuracy</th>
<th>Format</th>
<th>Ease of use</th>
<th>Timeliness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1</td>
<td>C2</td>
<td>C3</td>
<td>C4</td>
<td>A1</td>
</tr>
<tr>
<td>India</td>
<td>0.76</td>
<td>0.82</td>
<td>0.78</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>0.79</td>
<td>0.86</td>
<td>0.78</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>Taiwan</td>
<td>0.77</td>
<td>0.75</td>
<td>0.70</td>
<td>0.68</td>
<td>0.84</td>
</tr>
<tr>
<td>US</td>
<td>0.85</td>
<td>0.85</td>
<td>0.79</td>
<td>0.79</td>
<td>0.91</td>
</tr>
<tr>
<td>Western Europe</td>
<td>0.89</td>
<td>0.91</td>
<td>0.72</td>
<td>0.83</td>
<td>0.96</td>
</tr>
</tbody>
</table>
The invariance analysis results are shown in Table 4. The left side reports the invariance models (numbered 1–10), the model descriptions, and their corresponding fit statistics ($\chi^2$, degrees of freedom, NNFI, CFI, RMSEA, and ECVI). The right side shows the change in degrees of freedom between the nested models, the change in degrees of freedom between the nested models, and the significance level of the change in $\chi^2$ per degree of freedom. It contains a five subgroups analysis, a four subgroups analysis, and a three subgroups analysis. Each starts with an equal pattern model that served as a baseline.

### 4.1. Results for five subgroups invariance analysis

Table 4 reports the $\chi^2$, degrees of freedom, and fit indices for the factor loading invariant model (Model 2) for the five subgroups analysis. To test the hypothesis that the 12 items of the EUCS instrument has equivalent item-factor loadings on their corresponding first-order latent factors across cultures, we examined the differences in $\chi^2$ and degrees of freedom between the factor loading invariant model (Model 2) and the baseline equal pattern model (Model 1). The $\chi^2$ difference of 77.8 for 28 degrees of freedom is significant at $p = 0.0000$. Thus, the hypothesis of equivalent item-factor loadings across India, Saudi Arabia, Taiwan, the United States, and Western Europe respondents is rejected.

While it is rejected, an examination of the subjective fit indexes suggested that they did not decline much or at all as the item-factor loading invariance constraint was imposed. These minor changes in subjective fit indices suggest that, while the hypothesis of equivalent item-factor loadings is rejected, the variance may not affect the interpretation of some substantive research results (e.g., testing differences in mean scores).

We were testing measurement hypotheses rather than substantive hypotheses. Since the hypothesis of equal item-factor loadings across the five subgroups was rejected, we did not continue to test H2.

While the differences in item-factor loadings were significant, the test results did not indicate the source and nature of these differences. Item-factor loadings do help identify where the item-factor loadings differ: the item-factor loading (0.32) of E2 on the "ease of use" factor in the Saudi Arabia subgroup is substantially lower than those in other subgroups. To identify whether this item alone accounted for the rejection of H1, we explored whether the 11 items (exempting E2) had equivalent item-factor loadings on their corresponding first-order latent factors across cultures. Specifically, we explored whether, with the item-factor loadings of E2 ($\lambda_{E2}$) for the “ease of use” factor set free across all five subgroups, the other 11 items have equivalent loadings on their corresponding first-order factors across the five subgroups.

### 4.2. Rationale and results for four subgroups invariance analysis

We proceeded with a four subgroups invariance analysis because we wished to evaluate whether the EUCS instrument provided equivalent measurement across national cultures/world regions. Although the five subgroups analysis identified E2 as variant across subgroups, it did not confirm whether the variance in E2’s item-factor loading was limited exclusively to the Saudi Arabian subgroup or, whether E2 also had significant differences in item-factor loadings across other subgroups. While we could not test the equivalence of the five structural weights (i.e., H2) across the five groups, we wished to evaluate whether the EUCS...
instrument had equivalent structural weights across the US, Western Europe, India, and Taiwan. Therefore, we examined whether the 12 items of the EUCS instrument had equivalent item-factor loadings on their corresponding first-order latent factors across the US, Western Europe, India, and Taiwan and whether the five first-order latent factors had equivalent structural weights on the second-order EUCS factor across the US, Western Europe, India, and Taiwan.

To explore whether the 12 items of the EUCS instrument had equivalent item-factor loadings on their corresponding first-order latent factors across four cultures, we examined the differences in $\chi^2$ and degrees of freedom between the factor loading invariant model (Model 5) and the baseline equal pattern model (Model 4). The $\chi^2$ difference of 25.4 for 21 degrees of freedom is non-significant ($p = 0.2303$), suggesting that item-factor loadings across the four national cultures were equivalent.

The equivalence of item-factor loadings indicated that the first-order factors had the same meaning across the four subgroups. Since the first-order factors had the same meaning across the four subgroups, we could explore whether each first-order factor was equally important or central to the meaning of the second-order EUCS factor across the four cultural subgroups.

For EUCS factor scores to be in the same scale and thus comparable across subgroups, both item-factor loadings and structural weights of the five first-order factors had to be equivalent. To explore whether the structural weights for the five first-order factors were equivalent across the US, Western Europe, India, and Taiwan, we examined the differences in $\chi^2$ and degrees of freedom between the factor loading and structural weight invariant model (Model 6) and the baseline factor loading invariance model (Model 5). The $\chi^2$ difference of 26.0 for 15 degrees of freedom was significant at $p = 0.038$. Thus, the results suggested that the structural weights across the four cultures were not equivalent. While the structural weights across the four cultures were found not equivalent, the subjective fit indices were similar for both Models 5 and 6 (see Table 4). Thus, the differences may not be substantial enough to affect the interpretation of some substantive research results.

While the structural weights were statistically different across the four subgroups, the results did not indicate which factors or how many factors had different structural weights across subgroups. An examination of the structural weights suggested that major difference in structural weights (0.38), exempting the Saudi Arabia subgroup, occurs for the “ease of use” factor (i.e., 0.87 for India versus 0.49 for Western Europe). We thus explore whether the four first-order latent factors (content, accuracy, format, and timeliness) had equivalent structural weights on the second-order EUCS factor across the US, Western Europe, India, and Taiwan.

We examined the differences in $\chi^2$ and degrees of freedom between Model 7 (the factor loadings and structural weight invariant, except $\gamma_{koi}$ free) and Model 5 (the baseline factor loading invariance model). The $\chi^2$ difference of 19.4 for 12 degrees of freedom was non-significant ($p = 0.0793$). This suggested that content, accuracy, format, and timeliness of EUCS had equivalent structural weights. While each of these four information products factors had different structural weights in each subgroup, these weights were not statistically different across the US, Western Europe, India, and Taiwan subgroups.

4.3. Rationale and results for three subgroups invariance analysis

The four groups analysis identified the ease of use factor as having a different structural weight, but the analysis freed the ease of use structural weight across all four subgroups. Thus, it did not conclusively isolate this variance in structural weights as occurring only in the Western European subgroup. For example, the four groups analysis did not examine whether the structural weight in Table 3 for ease of use in India (0.87) was significantly different from Taiwan (0.71). Therefore, we explored whether the 12 items of the EUCS instrument had equivalent item-factor loadings on their corresponding first-order latent factors across the US, India, and Taiwan and whether the five first-order latent factors have equivalent structural weights on the second-order EUCS factor across the US, India, and Taiwan.

We first made sure the prerequisite of item-factor loadings equivalence was met in the three subgroups analysis. We then compared Model 8 (equal pattern) with the nested or more restrictive Model 9 (factor loading invariant). The difference in $\chi^2$ of 16.9 for 14 degrees of freedom was non-significant ($p = 0.2615$). The finding of no significant differences in item-factor loadings enables us to further compare Model 9 (factor loadings invariant) with nested Model 10 (factor loadings and structural weights invariant), yielding a change in $\chi^2$ of 16.3 for a change of 10 degrees of freedom—non-significant ($p = 0.0914$). The results indicated that all five first-order factors had equivalent structural weights across the US, India, and Taiwan samples.

5. Discussion

This exploratory study of the EUCS instrument had several limitations. First, the countries or world regions included in the study were based on national/world region samples that were gathered by independent researchers using different sampling strategies. Second, the samples did not represent all national cultures or world regions. Third, conclusions related to measurement equivalence may be subgroup specific. An instrument that is invariant across one set of national cultures may not be equivalent across others.

Despite these limitations, these results suggest that the 12-item EUCS instrument may be robust across cultures. We found that all 12 item-factor loadings appeared to be equivalent across the cultures examined. This suggested that: the 12-item summed scale might be used to make comparisons between samples gathered in different national cultures, and that there was no evidence that the measures of the component factors were not comparable across the national cultures examined.

All the structural weights were not equivalent across all regions. While content, format, accuracy, and timeliness showed no significant differences, ease of use had a significantly lower structural weight in the Western European subgroup compared to its value in the US, Taiwan, or India. This suggested that the importance or centrality of ease of use to the meaning of user satisfaction may differ across national cultures.

The results have implications on the use of 12-item summed scale or the second-order EUCS factor in comparing user satisfaction between samples gathered in different cultures. Managers may find the 12-item summed scale preferable because of its comparability and ease of calculation. For the nations/world regions examined, the 12-item summed scale is appropriate.

6. Conclusions

In global firms, the increasing use of enterprise wide applications has stimulated the need to monitor how effectively IT is being implemented and used in different parts of the world. The results of our study suggest that the meaning of user satisfaction may differ between cultures. Thus, researchers should be cautious when comparing user satisfaction scores across cultures, especially when using second-order EUCS factor scores. We recommend that
the 12-item summed scale or component factors (content, accuracy, format, ease of use, and timeliness) be used because no evidence of non-equivalent measurement was found across the cultures in our study.

Acknowledgements

Said S. Al-Gahtani wishes to thank King Abdul-Aziz City for Science and Technology (KACST), Saudi Arabia for providing financial support for this research project and the administration of King Khalid University for their help throughout the project phases.

References


Xiaodong Deng is an Associate Professor of Management Information Systems at Oakland University. He received his PhD in Manufacturing Management and Engineering from The University of Toledo. His research has appeared in Decision Sciences, Information and Management, Journal of Management Information Systems, Information Resources Management Journal, and Journal of Intelligent Manufacturing. His research interests are in post-implementation information technology learning, information systems benchmarking, and information technology acceptance and diffusion.

William J. Doll is a Professor of MIS and Strategic Management at the University of Toledo. Dr. Doll holds a doctoral degree in Business Administration from Kent State University and has published extensively on information system and manufacturing issues in academic and professional journals including Management Science, Communications of the ACM, MIS Quarterly, Academy of Management Journal, Decision Sciences, Journal of Operations Management, Information Systems Research, Journal of Management Information Systems, Omega, and Information & Management. Dr. Doll has published extensively in a variety of topics including computer integrated manufacturing, executive steering committees, top management involvement in MIS development, strategic information systems, information System downsizing, and end-user computing. He has developed instruments to measure a variety of constructs including user involvement in systems development, time-based manufacturing practices, system-use, impact of technology on work, and end-user computing satisfaction.

Said S. Al-Gahtani is an Associate Professor of Computer Information Systems in the Department of Administrative Sciences at King Khalid University, Abha, Saudi Arabia. He has a B.Sc. in Systems Engineering from King Fahad University of Petroleum & Minerals, M.Sc. in Computer Sciences from Atlanta University, Atlanta, Georgia, and a PhD in Computer-Based Information Systems from Loughborough University, Loughborough, UK. His research interests include the user acceptance of information technologies, the modeling of IT acceptance, end-user computing, and

**Tor J. Larsen** holds a MA in Systems Thinking from the University of Lancaster, England in 1975 and earned his Ph.D. in Management Information Systems (MIS) from the University of Minnesota in 1989. He is professor in Knowledge Management and holds the position of Senior Vice President at the Norwegian School of Management. He has served as associate editor for MIS Quarterly. Dr. Larsen has published widely, for example in Information & Management, Journal of MIS, and Information Systems Journal. He is a member of AIS and IFIP WG8.2 and WG8.6. In 2006 he was program co-chair for the WG8.6 Working Conference on the “Transfer and Diffusion of IT for Organizational Resilience.” Dr. Larsen’s research interests are in the areas of managers’ use of information, knowledge management, innovation, diffusion, representation, and innovation outcome.

**John Michael Pearson** is an Associate Professor of Information Systems at Southern Illinois University @ Carbondale. Before coming to SIUC, Dr. Pearson served as department chair (Business Computer Information Systems) at St. Cloud State University. Dr. Pearson has presented several papers at regional, national and international conferences. He has published articles in Communications of the ACM, Information & Management, Journal of Strategic Information Systems, Journal of Information Systems, Journal of Computer Information Systems, Decision Support Systems, Review of Business, Journal of Internet Commerce, Information Resources Management Journal and Public Administration Quarterly. His research interests are in the areas of organizational planning, technology adoption, e-commerce, management of quality, and IT project management.

**T.S. Ragu-Nathan** is Professor of Information Systems and Operations Management in the College of Business Administration at the University of Toledo. He holds a Ph.D. in Management Information Systems from the University of Pittsburgh. Has published in several journals including Information Systems Research, Decision Sciences, OMEGA: International Journal of Management Science, Journal of MIS, Journal of Information Systems, and Journal of Strategic Information Systems. His current research interests are in information systems strategy, quality issues in information systems, and use of information technology in manufacturing, Supply Chain Management, and E-Commerce.