ED 443 867 TM 031 512

AUTHOR	Sireci, Stephen G.; Harter, James; Yang, Yongwei; Bhola, Dennison
TITLE	Evaluating the Construct Equivalence of International Employee Opinion Surveys. Laboratory of Psychometric and Evaluative Research Report No. 379.
INSTITUTION	Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.
PUB DATE	2000-04-25
NOTE	32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Analysis of Covariance; Attitude Measures; *Attitudes; *Construct Validity; *Employees; *Languages; Psychometrics; *Test Format; World Wide Web

ABSTRACT

Assessing people who operate in different languages necessitates the use of multiple language versions of an assessment. However, different language versions of an assessment are not necessarily equivalent. In this paper, the psychometric properties of different language versions on an international employee attitude survey are evaluated. This survey was administered to more than 50,000 employees of a large telecommunications company using both paper-and-pencil and Web administration formats. The structural equivalence of the survey was evaluated across language versions, cultural groups, and administration formats using multidimensional scaling. The statistical equivalence of English, French, and Spanish versions of the survey items was evaluated using analysis of covariance. The results indicate the structure of the survey is consistent across the groups studied, and that the different language versions of the items functioned similarly. The implications of the results for future research in this area are discussed. (Contains 3 figures, 4 tables, and 38 references.) (Author/SLD)

Evaluating the Construct Equivalence of International Employee Opinion Surveys[1,2]

Stephen G. Sireci[3]
University of Massachusetts at Amherst

James Harter, Yongwei Yang, and Dennison Bhola
The Gallup Organization

[1] Laboratory of Psychometric and Evaluative Research Report No. 379. Amherst, MA: School of Education, University of Massachusetts.

[2] Paper presented at the annual meeting of the National Council on Measurement in Education as part of the symposium "Solving Problems in International and Cross-lingual Assessment" (L. Cook, chair), April 25, 2000, New Orleans, LA.

[3]The authors thank Denise McLain for her comments on an earlier version of this paper, and Ron Hambleton, H. Swaminathan, Mike Jodoin, and Bruno Zumbo for their thoughts regarding polytomous DIF detection procedures.

Abstract

Assessing people who operate in different languages necessitates the use of multiple language versions of an assessment. However, different language versions of an assessment are not necessarily equivalent. In this paper, we evaluate the psychometric properties of different language versions of an international employee attitude survey. This survey was administered to over 50,000 employees of a large telecommunications company using both paper-and-pencil and web administration formats. We evaluated the structural equivalence of the survey across language versions, cultural groups, and administration formats using multidimensional scaling. The statistical equivalence of English, French, and Spanish versions of the survey items was evaluated using analysis of covariance. The results indicated the structure of the survey was consistent across the groups studied and that the different language versions of the items functioned similarly. The implications of the results for future research in this area are discussed.

Evaluating the Construct Equivalence of International Employee Opinion Surveys

Many contemporary businesses are international, which requires an international workforce. In such companies, the employee population comprises individuals who speak different languages and are from different cultures, which poses special problems for valid measurement of employee attitudes, skills, and opinions. The process of evaluating employees and organizational systems must be sufficiently flexible to accommodate different languages and cultures. At the same time, for valid comparisons to be made across such individuals, the measurement properties of the evaluation instruments must be consistent across all linguistic and cultural groups.

In this paper, we discuss psychometric issues related to cross-lingual assessment and present the results of a series of studies conducted on different language versions of an employee attitude survey. The survey was administered worldwide to employees of an international telecommunications company in both paper-and-pencil and computerized formats. The purposes of the analyses were to evaluate the equivalence of the survey across its different language versions as well as across the two administration media. Issues of both construct and item equivalence were studied.

Before describing our study, we discuss some of the important psychometric issues to be addressed in cross-lingual assessment. Drawing from the literature in this area, we demonstrate some empirical analyses that can be conducted to evaluate the equivalence of different language versions of an assessment.

Research on the Equivalence of Multiple Language Versions of an Assessment

Although providing multiple language versions of an assessment is commendable, it is well known that mere translation of an instrument into alternate languages does not guarantee the different language versions are equivalent, or even comparable (Geisinger, 1994; Hambleton, 1994; Sireci, 1997; van der Vijver & Poortinga, 1997). To facilitate measurement equivalence across different language versions of an assessment, great care must be taken in adapting the instruments and statistical analyses must be conducted to evaluate the comparability of final products. The Guidelines for Adapting Educational and Psychological Tests (Hambleton, 1994) provide guidance regarding the translation/adaptation process (see also Hambleton & Patsula, 1998), and encourage test developers to conduct statistical analyses to check cross-lingual equivalence. In particular, these guidelines recommend comparing descriptive statistics, such as score reliability estimates and standard errors of measurement, dimensionality (factor) analyses, and analyses of item bias (differential item functioning).

Van der Vijver and Tanzer (1998) provided further guidance to cross-cultural researchers for evaluating translated instruments. In providing their taxonomy of bias and equivalence in cross-cultural assessment, they discussed three levels of equivalence and three levels of bias. The first level of equivalence is construct equivalence, which signifies the same construct is measured by instruments in all cultural groups (i.e., an "etic" situation in the Hui & Triandis, 1989 terminology). The second level of equivalence is measurement unit equivalence, which occurs when the assessments are measuring the same construct using a common metric, but the origin of the metric differs, such as in the case of the Farenheight and Celsius temperature scales. The third level of equivalence is scalar equivalence, which occurs when all assessments are measuring the same construct using the "same measurement unit and same origin" (p. 266).

Construct equivalence is most often established through rational analysis and familiarity with the cultural groups being assessed. The primary issue to be resolved is whether the construct to be measured exists in all cultures and can be measured in an equivalent manner. Measurement unit equivalence and scalar equivalence are more difficult to establish. Therefore, many test specialists and cross-cultural researchers have stressed the need to ensure that construct, method, and item bias do not exist in different language versions of an assessment (e.g., Geisinger, 1994; Hambleton, 1993, 1994; Sireci, 1997, in press; van der Vijver & Poortinga, 1997). For example the Guidelines for Adapting Educational and Psychological Tests developed recently by the International Test Commission stipulate:

> Instrument developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the instrument, and (2) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations. (Hambleton, 1994, p. 232)

The first requirement relates to construct equivalence, while the second requirement relates to differential item functioning (DIF). Both lack of construct comparability and DIF can lead to test bias, which implies that inferences derived from test scores are not equivalent across groups.

In this report, we evaluate the comparability of an employee attitude survey, which was developed and administered by Gallup for a large telecommunications company. Descriptive statistics for three different language versions of the survey are presented. In addition, we compare the structure of the survey across groups who differ with respect to native language and ethnicity. Structural equivalence is an important aspect of construct equivalence and is a prerequisite for more detailed analysis of test comparability, such as analyses of DIF. Following the structural analyses, we also present the results of cross-lingual DIF studies.

Method

Instrument

The employee attitude survey (EAS) comprised 50, five-point Likert-type items. The survey was available in eight languages (English, French, German, Italian, Japanese, Mandarin, Portuguese, and Spanish) and in both paper-and-pencil and web-based formats. The EAS contained five sub-scales: Basic Needs, Management Support, Team Work, Growth, and "other." Each of the 50 items was linked to one of these five sub-scales. It was administered to 51,108 employees of a large telecommunications company during April and May 1999. These employees came from 47 different countries. For the purposes of this study, we focused on the three largest volume language versions of the survey: English, French, and Spanish. To evaluate the cross-lingual, cross-cultural, and cross-format functioning of the survey, we distinguished between employees who took the paper and web versions of the survey, as well as those who responded to the same language version, but were from different countries (e.g., employees who spoke English, but were from Canada, Ireland, the U.S., etc.). Due to some of the statistical procedures used in this study, the minimum sample size for a cultural group to be included was 500.

Survey translation steps

The items composing the EAS were originally developed in English. Professional translation teams were hired (separate translation teams were used for each language version of the survey) and met with Gallup scientists to review the content and psychological meaning and intent of the items. Following this meeting, the items were translated into each language. The translations were reviewed for content and accuracy by a bilingual Gallup editor and by a bilingual human resource specialist, who was local to the translation. When discrepancies between the editor and translation team arose, they were resolved via phone conversations and written correspondences between the editor and the translation team. Back-translation was not used for the survey discussed in this paper, but it has recently been added as a further quality control check for current international surveys developed by Gallup.

Survey Participants

The remaining survey participants (n=40,595) were partitioned into seven language/cultural groups: U.S. (English), U.K. (English), Ireland (English), Canadian-English, Canadian-French, French (other), and Spanish. Non-Canadian employees who responded to the French versions of the survey were predominantly from France (over 99%). Employees who responded to the Spanish versions were all from Latin America, predominantly from Mexico (82%) and Columbia (8%). Within each language group, sub-groups were created based on whether the respondents took the web or paper version of the survey.

The sample sizes for the non-Canadian French, Irish, and Spanish groups were not large enough to derive separate matrices for the web and paper versions. For the non-Canadian French and Irish groups, only those who responded to the web version of the survey were included. For the Spanish group, only those who took the paper version of the survey were included. Thus, the analyses involved a total of eleven groups (see Table 1). Across all respondents, 31% were female and 69% were male.

[Insert Table 1 Here]

Analyses

Descriptive statistics

Means, standard deviations, reliability estimates (coefficient alpha), and standard errors of measurement were computed for the total sample and for selected sub-groups differing with respect to language and medium of administration.

Principal components analysis

Principal components analysis (PCA) was used to ascertain a rough estimate of the overall structure of the EAS, irrespective of language version, cultural group, or administration format. In this analysis, the entire data set was analyzed (i.e., collapsing across all language versions and administration formats). The purpose of this analysis was to get a general idea of whether a single, dominant dimension underlied the 50 items, or whether other dimensions were present.

Multidimensional scaling analyses

Weighted multidimensional scaling (MDS) analyses were used to more comprehensively examine the structure of the EAS. MDS is preferable to PCA and other forms of factor analysis when the purpose of the analysis is to discover both the gross and subtle structures of an assessment (Davison, 1985; Davison & Skay, 1991; Sireci, 1998). MDS scales stimuli, such as survey items, along one or more continuous dimensions. It is considered inferior to factor analysis for the purpose of deriving factor scores for respondents. However, it is ideal for evaluating the relationships among survey items with respect to dominant and weaker dimensions. An extremely attractive feature of weighted MDS is that the structure of an assessment can be evaluated simultaneously across multiple groups. Given that employee survey scores are rarely, if ever, reported at the respondent level, group-level analyses are particularly important in this context.

In this study, we used weighted MDS to determine if the structure of the survey was consistent across groups defined by the language version and administration format (web versus paper) of the survey. In weighted MDS models, a common dimensional structure is derived simultaneously for all groups, and weights indicating the salience of each dimension for each group are provided. Differences among the groups with respect to survey structure are reflected in the group weights, which "stretch" or "shrink" the multidimensional space to best fit the data for each group.

The INDSCAL weighted MDS model (Carroll & Chang, 1970) was used for all MDS analyses. This model specifies a weighted Euclidean distance formula to scale the survey items:

$$d_{ijk} = \sqrt{\sum_{a=1}^{r} w_{ka} (x_{ia} - x_{ja})^2} \qquad [1]$$

where: d_{ijk}=the Euclidean distance between stimuli (e.g., survey items) *i* and *j* for group *k*, w_{ka} is the weight for group *k* on dimension *a*, x_{ia}=the coordinate of stimulus *i* (i.e., survey item *i*) on dimension *a*, and *r*=the dimensionality of the model. A common structural space, called the stimulus space, is derived for the stimuli. The "personal" distances for each group are related to the common stimulus space by the equation:

$$x_{kia} = \sqrt{w_{ka}}\, x_{ia} \qquad [2]$$

where x_{kia} represent the coordinate for stimulus *i* on dimension *a* in the personal space for group *k*, w_{ka} represents the weight of group *k* on dimension *a*, and x_{ia} represents the coordinate of stimulus *i* on dimension *a* in the common stimulus space.

Differences in dimensional structure across groups are reflected in the group weights (i.e., w_{ka}). The larger a weight on a dimension (*a*), the more that dimension is necessary for accounting for the variation in the data for the specific group (*k*). In the INDSCAL model used here (all analyses were implemented using the ALSCAL program in SPSS, version 8.0), the weights range from zero to one. A weight of zero indicates the dimension is completely irrelevant to the data for the group. A weight of one indicates the MDS coordinates on that dimension completely account for the variation in the data for that group. Using simulated data, Sireci, Bastari, and Allalouf (1998) found that when structural differences exist across groups on one or more dimensions, one or more groups will have weights near zero, while other groups will have noticeably larger weights. They concluded non-equivalence of the structure of an assessment across groups should be obvious via inspection of the MDS weights.

To conduct the MDS analyses, Pearson correlations were computed among the 50 items. Separate inter-item correlation matrices were computed for each group (11 matrices total). MDS models fit distances to dissimilarity data, not to similarity data. Therefore, the correlations were transformed to dissimilarities before MDS analysis using the transformation suggested by Davison (1985):

$$\delta_{ij} = \sqrt{2-2r_{ij}} \qquad [3]$$

where δ_{ij}=the dissimilarity between item *i* and *j*, and r_{ij}= the tetrachoric correlation between items *i* and *j*.

Although weighted MDS models can evaluate test structure simultaneously across all groups, most MDS models do not provide statistical tests of structural equivalence (cf. Ramsay, 1982). Rather, descriptive fit indices are used to evaluate data-model fit. The STRESS index represents the square root of the normalized residual variance of the monotonic regression of the MDS distances on the transformed proximities. Thus, lower values of STRESS indicate better fit. The R^2 index reflects proportion of variance of the transformed proximities accounted for by the MDS distances. Thus, higher values of R^2 indicate better fit. Recent applications of weighted MDS have illustrated its advantages for evaluating structural equivalence across cultural groups (Day & Rounds, 1998; Day, Rounds, & Swaney, 1998) and across different language versions of a test (Sireci, Bastari, & Allalouf, 1998; Sireci, Fitzgerald, & Xing, 1998).

DIF analyses

The DIF analyses aimed toward identifying items that functioned differently across their different language versions. Several statistical methods for evaluating DIF are available, but most are designed for items that are scored dichotomously (see Camilli & Shepard, 1994 or Holland & Wainer, 1993, for comprehensive descriptions of these methods). Several methods are also available for items with multiple categories, such as Likert-type items used on many surveys. A summary of these methods is presented in Table 2. In this study, we used analysis of covariance (ANCOVA) to study translation DIF. Language group was used as the grouping (independent) variable, score on the item was used as the dependent variable, and total EAS score was used as the covariate. Thus, a separate ANCOVA was run for each item. Although this method is less sophisticated than other methods listed in Table 2, it has several attractive features. First, we could look at the functioning of the item across all three language groups in a single analysis. Second, it is quick and easy to implement using standard statistical software (all ANCOVAs were conducted using SPSS). Third, it is less labor intensive in comparison to the other methods. Given that when translation problems occur in cross-lingual assessment they tend to produce relatively large performance differences across groups (Sireci, Xing, & Fitzgerald, 1999), we believe this method is appropriate for identifying items that involve modest to severe translation problems.

[Insert Table 2 Here]

There are three general limitations of the ANCOVA method for evaluating DIF: (a) the procedure involves making three model assumptions, (b) only conditional mean differences will be detected (i.e., uniform DIF), and (c) there are no standard criteria for how big of a difference constitutes meaningful DIF. With respect to model assumptions, the ANCOVA model assumptions are linearity, homoscedasticity, and homogeneity of regression. Linearity refers to the assumption that the relationship between the total EAS score (minus the studied item) and the item score is linear. This assumption can be evaluated by graphing the relationship and checking for non-linearity. The homoscedasticity assumption states that the variance about each regression line is equivalent along the entire line. This is a common assumption in regression analysis (of which ANCOVA is a special type). The homogeneity of regression assumption states the regression slope for predicting the item score from the total score (minus the studied item) is the same for all three language groups. This assumption is also testable.

Given the large sample sizes used in the DIF analyses (minimum n=1,419 for these analyses), statistical significance is an insufficient criterion for identifying DIF. Instead, we used an effect size criterion based on the proportion of variance in the (covariate-adjusted) item score data accounted for by language group membership. In the context of logistic regression analysis, Zumbo (1999) suggested using an R-squared cutoff of .13 for flagging items for DIF. Gierl and McEwen (1999) and Jodoin (1999) stated that R-squared effect sizes of .035 and .07 roughly corresponded to the Educational Testing Service's categories of "moderate," and "large" DIF, respectively. In this study, we took the more conservative suggestion. Items were identified as functioning differentially across language versions if the eta-squared statistic associated with the ANCOVA was equal to or greater than .07.

Results

Descriptive Statistics

In addition to presenting the sample sizes, Table 1 presents descriptive statistics for the eleven groups who differed with respect to language, culture, and/or survey format. Across all language groups, employees who took the web version of the survey reported higher overall satisfaction than employees who took the paper version. The difference across the web and paper versions was largest for the U.K. English group (about 13 points, which was almost half a standard deviation). The respondents from Ireland (web version) exhibited the highest mean satisfaction score (188.19), and the U.K. English employees who took the paper version had the lowest mean satisfaction score (168.04). The coefficient alpha reliability estimates were high across all groups, ranging from .945 (non-Canadian French) to .970 (U.S. English paper). The standard errors of measurement were similar across all groups, ranging from 5.48 (Ireland, web) to 6.51 (Spanish, paper). The standard deviations were also similar, ranging from 26.11 (Ireland) to 32.27 (U.S. paper).

Principal components Analysis

The PCA results suggested a large general satisfaction factor and several much smaller factors underlied the data. Seven factors had eigenvalues greater than one. The first factor accounted for 36.5% of the variance in the data, while the second factor accounted for 6% of the variance. This sharp drop in percentage of variance accounted for indicates the dominance of the first factor. All 50 items had positive loadings on this factor. The smallest loading was .24 (item 11), the next smallest loading was .41. After the second factor, the percentage of variance accounted for dropped to 3% for the third and fourth factors, and to 2% for the other four factors with eigenvalues greater than one. These smaller factors are not interpreted here, since the MDS results are more revealing regarding the subtleties of the EAS structure. In general, our conclusion from the PCA analyses was that a dominant general factor exists, but several smaller factors are also present. However, these conclusions are based on an aggregate analysis, which cannot be generalized to the various linguistic and format versions of the survey.

MDS Analyses

Identifying the dimensionality

Two- through six-dimensional MDS models were fit to the data. The criteria of model fit to the data and interpretability were used to select the appropriate MDS solution. As is par for the course, the more dimensions fit to the data, the better the model will fit the data. Therefore, in evaluating model fit, the deceleration of improvement in fit is evaluated across solutions ranging from lower to higher dimensionality. The STRESS and R^2 indices for the MDS solutions are presented in Table 3. A substantial increase in fit is seen between the two- and three-dimensional solutions, then, the improvement in fit seems to taper off. However, the STRESS index dropped from .16 to .14 between the five- and six-dimensional solutions, which warranted further study to see if the improvement in fit was substantive. As described below, we found all dimensions in the six-dimensional solution to be readily interpretable. In addition, the data from each of the eleven groups studied were adequately fit by this solution (R^2 ranged from .64 to .89 across the groups). Therefore, the six-dimensional solution was taken as the best representation of the structure of the survey. This solution accounted for 78% of the variance in the (transformed) item dissimilarity data. The percentages of variance in the data accounted for by the dimensions was 25%, 18%, 10%, 9%, 9%, and 8%, for the first through sixth dimensions, respectively.

[Insert Table 3 Here]

Interpreting the MDS solution

In interpreting the solution, we looked for clusters in the MDS space that corresponded to the five sub-scales composing the survey (Basic Needs, Management Support, Team Work, Growth, Other). We also looked at the content of the survey items that had extreme positive or negative coordinates on each dimension. The first dimension corresponded to the "Management Support" sub-scale. This dimension distinguished between items involving motivational factors such as recognition and praise (e.g., "In the last seven days, I have received recognition or praise for doing good work") from those that had nothing to do with employee motivation (e.g., "[the company] is effectively delivering value to our customers").

The second dimension distinguished between items measuring business objectives and efficiency (e.g., "I can clearly explain ... the business strategy of [our company]...," "The changes made... will help us achieve ... business objectives") from those relating directly to the employee (e.g., "Conditions at work allow me ... to balance my work and personal life"). This dimension was labeled "business objectives." The third dimension corresponded to perception of the company overall and was labeled "global satisfaction." Items measuring overall satisfaction (e.g., "...how would you rate your overall satisfaction with [this company]...") had large coordinates (of similar sign) on this dimension.

The fourth dimension corresponded roughly to the "Basic Needs" sub-scale. Seven of the eleven items measuring this sub-scale had large coordinates (in the same direction) on this dimension. The four other items dealt with managerial components of support, which is different from material support. This dimension distinguished between items measuring satisfaction with things that are provided to employees (e.g., "I have the materials and equipment I need to do my work right") from other items. The fifth dimension was an interpersonal relationships dimension that scaled items according to their relevance to interpersonal relationships at work. This dimension distinguished between items such as "I have a best friend at work" and items dealing with issues such as compensation and workplace efficiency.

The sixth dimension corresponded somewhat to the "Team Work" sub-scale. Six of the ten items measuring this sub-scale had relatively large coordinates in the same direction. Closer inspection of the content of the items suggested this dimension measured working with others, particularly in work groups. Thus, this dimension was labeled "work groups."

In general, the observed structure of the survey data was consistent with the hypothesized structure. Three of the five sub-scales were represented in the solution (Basic Needs, Management Support, Team Work) and all dimensions were relevant to important features of employee satisfaction that the survey was designed to measure. A two-dimensional sub-space of the six-dimensional MDS solution is presented in Figure 1. Some clusters of items corresponding to the survey sub-scales are evident, particularly for the Management Support and Basic Needs sub-scales.

[Insert Figure 1 Here]

Evaluating the equivalence of the structure across groups

The MDS dimensions interpreted in the previous section describe the structure that best fits all groups simultaneously. Given the cross-cultural nature of the present research, the consistency of this structure across all studied groups is of greater interest. Thus, we turn now to inspection of the weights for each of the groups on each of the MDS dimensions.

In interpreting the group weights, it seems sensible to search for differences across: (a) the web and paper versions of the survey, (b) the different language versions of the survey, and (c) groups who speak the same language but came from different countries. The group weights are presented in Table 4. None of the groups have weights near zero on any of the dimensions, which suggests that there are no major structural differences across any of the groups.

[Insert Table 4 Here]

Web versus Paper differences. Figure 2 presents a two-dimensional weight space that illustrates the largest structural differences across groups distinguished by survey type. Although both dimensions appear relevant to all groups (minimum weight on either dimension is about .20), the "business objectives" dimension accounted for more of the variation in the web survey data, relative to the "global satisfaction" dimension, and vice-versa. Although the magnitude of the difference in the relative weightings of these dimensions by these two groups is small, it is noteworthy and could help explain the differences in total satisfaction score noted across the web and paper groups.

[Insert Figure 2 Here]

Language and cultural group differences. Figure 3 presents a two-dimensional sub-space of the group weights where the groups are labeled according to their language and country of origin. This two-dimensional space was selected because these two dimensions had the largest differences across the different language groups. As can be seen in the figure, the differences across the group weights are minor, and there is no pattern consistent with the different languages or cultures. The weights for the French and Spanish versions do not stand out from those for the English versions, and the weights for groups from the same country are no closer to one another than are the weights from different countries. These results suggest the structure of the survey is consistent across these languages and cultures. If there are any structural differences across the groups studied, they appear to be related to medium of the survey, rather than language version. Thus, these results suggest the translation and adaptation of the EAS was successful.

[Insert Figure 3 Here]

Differential item functioning analyses

Fifty ANCOVAs were run to assess DIF for each of the 50 survey items. The language group effect sizes (eta-squared) ranged from zero to .05, with a mean of .008 and a standard deviation of .01. None of the items reached the .07 criterion for DIF. However, three items had effect sizes larger than .03: items 6 (.035), 11 (.031), and 12 (.049). These three items were from three different sub-scales, and did not cluster together in the MDS solution. For items 6 ("My manager, or someone at work, seems to care about me as a person") and 12 ("In the last six months, someone has talked to me about my progress"), employees who responded to the English version tended to agree more strongly than those who responded to the French or Spanish versions. The opposite pattern occurred on item 11 ("I have a best friend at work").

Given the interpersonal/social aspects of these items, some potential for differential translation exists. This hypothesis will be followed-up by bilingual translators (these analyses were not available at the time of this writing), but it should be noted that they could also be due to chance variation.

Discussion

When he formally introduced the summated rating scale method for measuring attitudes, Likert (1932) explicitly pointed out that cultural differences might preclude construct equivalence. As he put it, "because a series of statements form a unit or cluster when used with one group of subjects,...it does not follow that they will constitute a unit on all other groups of persons with the same or different cultural backgrounds" (p. 53). In discussing the context specificity of assessment scores, Messick made a similar point: "the extent to which a measure displays the same properties and patterns of relationships in different population groups and under different ecological conditions becomes a pervasive and perennial question" (p. 15). Therefore, studying the construct equivalence of different language versions of an assessment is a fundamental validity issue whenever translated or adapted assessment instruments are used, or whenever the assessment involves different cultural groups.

The Standards for Educational and Psychological Testing (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999) and the Guidelines for Adapting Educational and Psychological Tests (Hambleton, 1994) require testing agencies to demonstrate the comparability of different language versions of an assessment whenever comparisons are made across people who take the assessment in different languages. For international businesses, it is important to compare the attitudes, skills, and opinions of employees who operate in different languages. Therefore, measurement equivalence across languages is critical.

The results of the present study support the cross-lingual equivalence of the EAS. The internal consistency reliability estimates and standard errors of measurement were consistent across all studied groups. Furthermore, MDS analysis of the structure of the survey indicated the structure was consistent across the different language versions of the assessment, and preliminary analyses of DIF did not identify major item translation problems.

It is interesting that the mean scores were noticeably different across employees who took the web and paper versions of the survey. It could be that employees who could not access the web version of the survey come from locations that have fewer computer resources, which could explain their lower satisfaction. However, given the slight structural differences noted across the web and paper surveys, it is possible that the web and paper versions are not necessarily equivalent. It could be that some survey items were mishandled when translating them from paper to the computer or vice versa. On the other hand, it is also possible that this difference is a function of the types of employee groups who were surveyed using each mode, and perhaps not the mode itself. For example, it could be that employees who responded to the web version were closer to the actual business objectives and had more interaction with them (e.g., managers), which made this dimension account for more of the variation in their data. These and other hypotheses should be followed up, since they are likely to inform future development of Gallup surveys.

This study looked at data from a subset of the survey population and so the results may not generalize to other groups, such as employees who took the test in German, Italian, Mandarin, or Portuguese. Future research should look beyond global differences across the groups and focus on individual items. Given the multidimensionality noted in the structure of the survey, it may be better to conduct DIF analyses within each sub-scale, rather than using total survey score as the conditioning variable. However, the large general factor noted in the principal component analysis, supports the use of a unidimensional covariate.

As recent research in cross-lingual assessment has indicated, translating assessment material is a difficult endeavor. This study provides preliminary evidence that Gallup's translation/adaptation process is effective. However, further research is needed on other international surveys. Since Gallup's translation procedures closely followed the Guidelines stipulated by the International Test Commission (Hambleton, 1994), we recommend these guidelines to other measurement specialists who need to develop multiple language versions of their assessments. In addition, we also support the International Test Commission's guidelines urging cross-cultural researchers to empirically evaluate the equivalence of test and item scores across original and adapted instruments. The results of the present study illustrate the utility of MDS for assessing construct equivalence, as well as the utility of ANCOVA for assessing DIF in Likert-type items. Future research should also evaluate the generalizability of these statistical methods across different types of assessment data.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Brown, P. J. (1996). Using differential item functioning analysis to determine differential item interpretations of survey questions. Unpublished doctoral dissertation. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Camilli, H., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage.

Carroll, J. D., & Chang, J. J. (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika, 35, 283-319.

Chang, H., Mazzeo, J., & Roussos, L. (1993, April). Detecting DIF for polytomously scored items: An adaptation of Shealy-Stouts SIBTEST procedure. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Davison, M. L., (1985). Multidimensional scaling versus components analysis of test intercorrelations. Psychological Bulletin, 97, 94-105.

Davison, M.L., & Skay, C.L. (1991). Multidimensional scaling and factor models of test and item responses. Psychological Bulletin, 110, 551-556.

Day, S. X., & Rounds, J. (1998). Universality of vocational interest structure among racial and ethnic minorities. American Psychologist, 53, 728-736.

Day, S. X., Rounds, J., & Swaney, K. (1998). The structure of vocational interests for diverse racial-ethnic groups. Psychological Science, 9, 40-44.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304-312.

Gierl, M. J., & McEwen, N. (1999). Consistency among statistics methods and content review for identifying differential item functioning. Unpublished manuscript. University of Alberta.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. Social Indicators Research, 45, 153-171.

Holland, P.W., & Wainer, H. (Eds.). (1993). Differential item functioning. Hillsdale, New Jersey: Lawrence Erlbaum.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology, 20, 296-309.

Jodoin, M. G., (1999). Reducing Type I error rates using an effect size measure with the logistic regression procedure for DIF detection. Unpublished master's thesis, University of Alberta, Edmonton, AB, Canada.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27, 307-327.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140, 44-53.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Messick, S. (1989b). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.) (pp. 13-103). Washington, D.C.: American Council on Education.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. Journal of Educational Measurement, 30, 107-112.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. Journal of the Royal Statistical Society, 145, 285-312.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. Educational Measurement: Issues and Practice,16,12-19.

Sireci, S. G. (1998). Gathering and analyzing content validity data. Educational Assessment, 5, 299-321.

Sireci, S. G., (in press). Evaluating cross-lingual test comparability using bilingual research designs. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.) Adapting educational and psychological tests for cross-cultural assessment. Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S. G., Bastari, B., & Allalouf, A. (1998). Evaluating construct equivalence across adapted tests. . Laboratory of Psychometric and Evaluative Research Report No. 340. Amherst, MA: University of Massachusetts, School of Education.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). Adapting credentialing examinations for international uses. Laboratory of Psychometric and Evaluative research report no. 329. Amherst, MA: University of Massachusetts, School of Education.

Sireci, S.G., Xing, D., & Fitzgerald, C. (1999, April). Evaluating translation DIF across multiple groups: Lessons learned from the Information Technology industry. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Swaminathan, H. (2000, April). Methodological issues in cross-linguistic comparisons. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thissen, D, Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

van der 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 der Vijver, F. & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 47, 263-279.

Welch, C. J. (1990). Procedures for extending item bias detection techniques to polytomously scored items. Unpublished doctoral dissertation. Iowa City, IA: The University of Iowa.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. Journal of Educational Measurement, 32, 163-178.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Table 1

Descriptive Statistics for Selected Groups

Language Group	Medium	N	Mean Survey Score	Standard Deviation	Coefficient Alpha	Standard Error of Measurement
Ireland English	Web	513	188.19	26.11	.956	5.48
U.S. English	Web	16,152	184.39	29.56	.965	5.53
	Paper	1,271	177.45	32.27	.970	5.59
U.K English	Web	3,433	181.37	28.37	.957	5.88
	Paper	1,952	168.04	29.48	.955	6.25
Canadian English	Web	10,799	183.49	28.58	.962	5.57
	Paper	2,415	173.89	30.39	.963	5.85
Canadian French	Web	765	187.66	28.81	.961	5.69
	Paper	727	180.40	29.77	.962	5.80
French (Other)	Web	1,149	169.76	26.84	.952	5.88
Spanish	Paper	1,419	179.83	27.78	.945	6.51
Total		**40,595**				

Table 2

Comparison of Selected Polytomous DIF Detection Methods

Method	Sample References	Strengths	Limitations
Analysis of Covariance	Current paper.	Can be conducted using standard statistical software packages. Can look at multiple groups simultaneously. Easy to implement.	Cannot detect non-uniform DIF. Strong model assumptions. No research base to date.
IRT Likelihood Ratio	Thissen, Steinberg, & Wainer (1988, 1993)	Extremely powerful. Can detect both uniform and non-uniform DIF.	IRT model assumptions must be met. Requires relatively large sample sizes. Time consuming—multiple calibration runs required to isolate DIF.
Poly-SIBTEST	Chang, Mazzeo, & Roussos (1993)	Can assess DIF amplification and cancellation.	Relatively small research base. Ability to detect non-uniform DIF is unclear.
Generalized Mantel-Haenszel	Welch, 1990; Welch & Miller, 1995	Relatively modest sample size requirements.	Cannot detect non-uniform DIF. Less powerful than other polytomous methods.
Loglinear Modeling	Mellenbergh (1982); Kelderman & Macready, (1990)	May detect both uniform and nonuniform DIF. No linearity or distribution assumptions.	Requires relatively larger sample sizes. Treats continuous variables as categorical.
Polytomous Logistic Regression; Ordinal Log. Regress.	Swaminathan & Rogers, (1990); Brown, 1996; Zumbo (1999)	May detect both uniform and nonuniform DIF	Requires specialized software. Relatively small research base.
Logistic Discriminant Function Analysis	Miller and Spray (1993)	May detect both uniform and nonuniform DIF	Specialized software required. Relatively small research base.
Confirmatory Factor Analysis	Swaminathan (2000)	Can incorporate multiple variables into analysis. Can also investigate common factor structure.	Can only detect uniform DIF. Requires multivariate normality and linear relationships among variables.

Table 3

Fit Statistics For MDS Solutions

Dimensions	**STRESS**	**R^2**
2	.30	.60
3	.23	.69
4	.19	.72
5	.16	.75
6	.14	.78

Table 4

MDS Group Weights

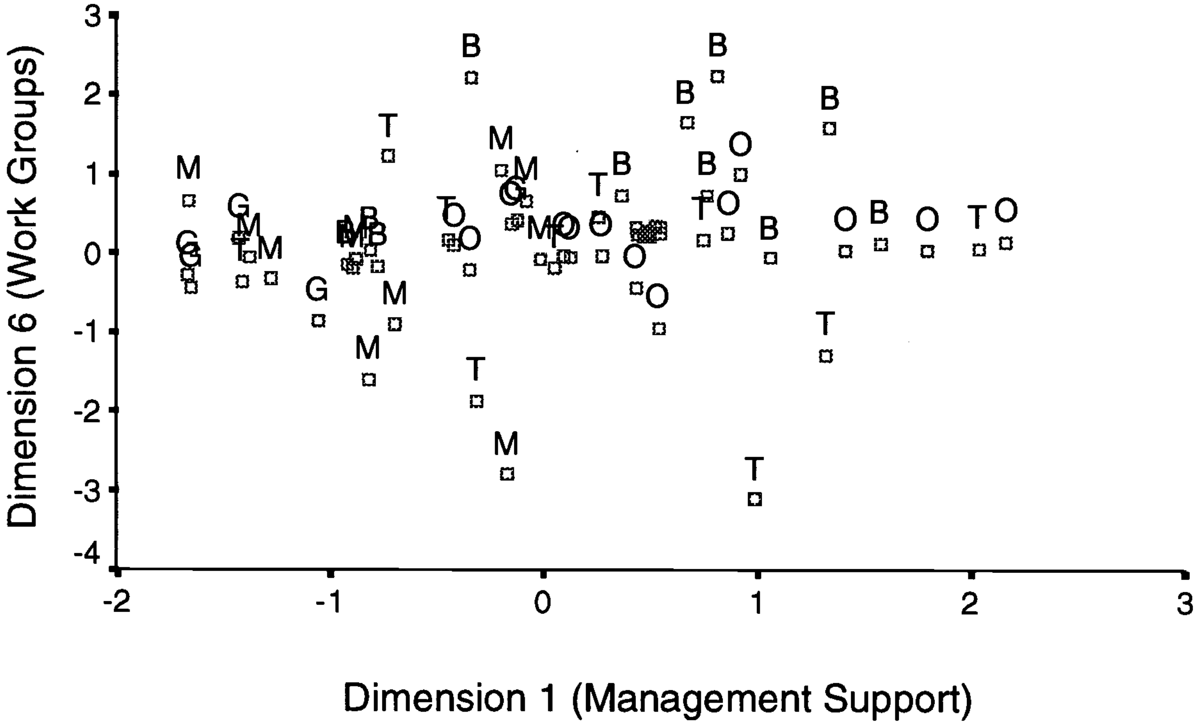
Group	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Weirdness[a]
U.S. (Paper)	.43	.34	.40	.34	.25	.32	.15
U.S. (Web)	.48	.50	.26	.26	.43	.28	.19
Spanish (Paper)	.50	.27	.35	.34	.16	.26	.20
French (Web)	.55	.42	.19	.34	.29	.26	.15
Can. English (Paper)	.52	.31	.39	.27	.28	.36	.16
Can. English (Web)	.49	.56	.26	.24	.40	.24	.21
Can. French (Paper)	.44	.27	.43	.28	.20	.25	.21
Can. French (Web)	.47	.50	.27	.33	.30	.21	.13
U.K. (Paper)	.55	.31	.36	.30	.22	.33	.15
U.K. (Web)	.49	.56	.21	.28	.36	.26	.19
Ireland (web)	.50	.41	.30	.33	.22	.21	.11
% Variance Accounted for[b]	.25	.18	.10	.09	.09	.08	

[a] The weirdness index indicates how similar the pattern of weights for the group is to the average weights for all groups. A value near zero reflects a weight pattern similar to the average across all groups. A value near one reflects weighting a single dimension very high, and all others very low.

[b] These statistics indicate the percentage of variance in the transformed dissimilarities that is accounted for by the item coordinates on the dimension.

Figure 1

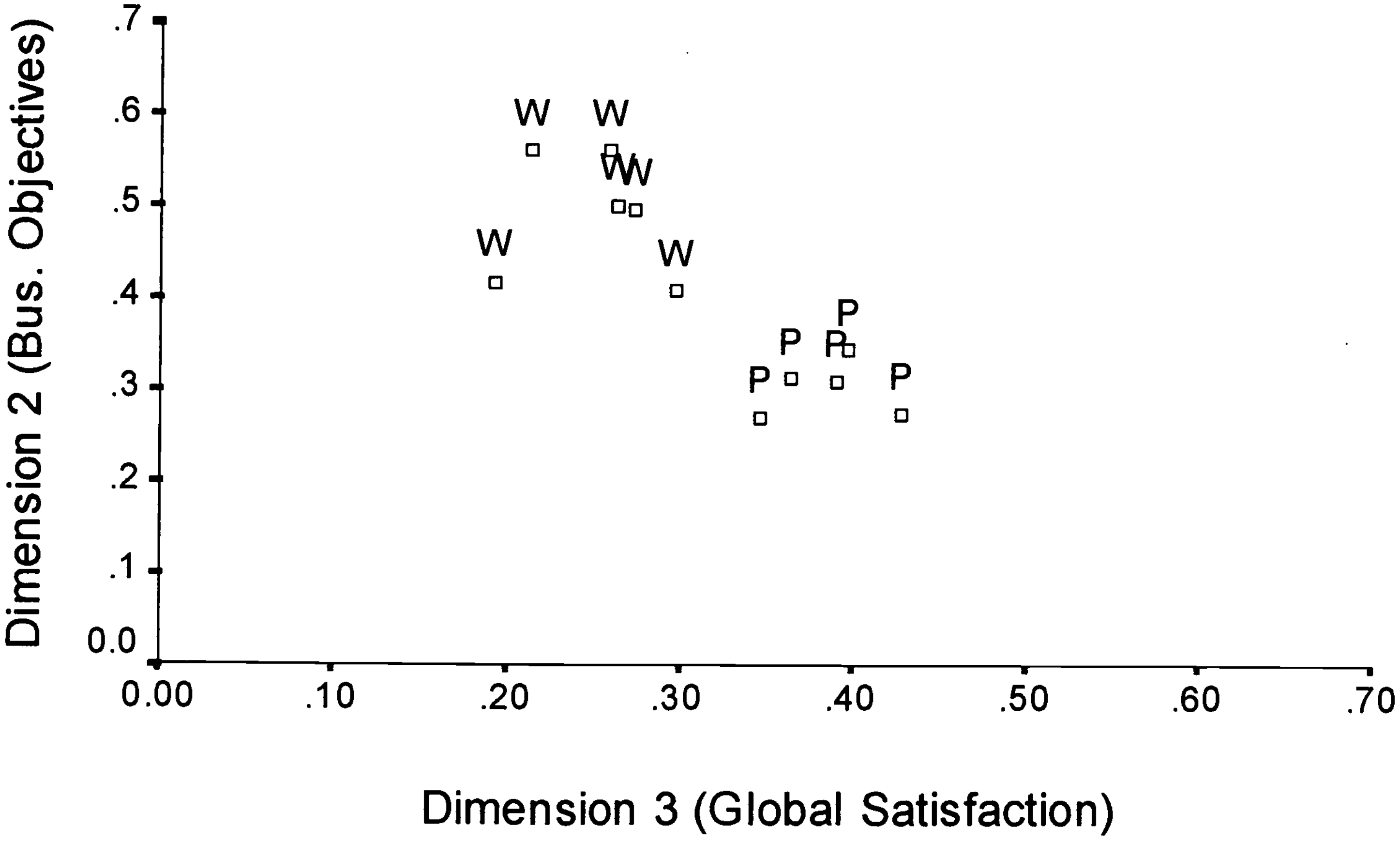
Survey Sub-space



B=Basic Needs, M=Mgmt. Support, T=Team Work, G=Growth, O=Other

Figure 2

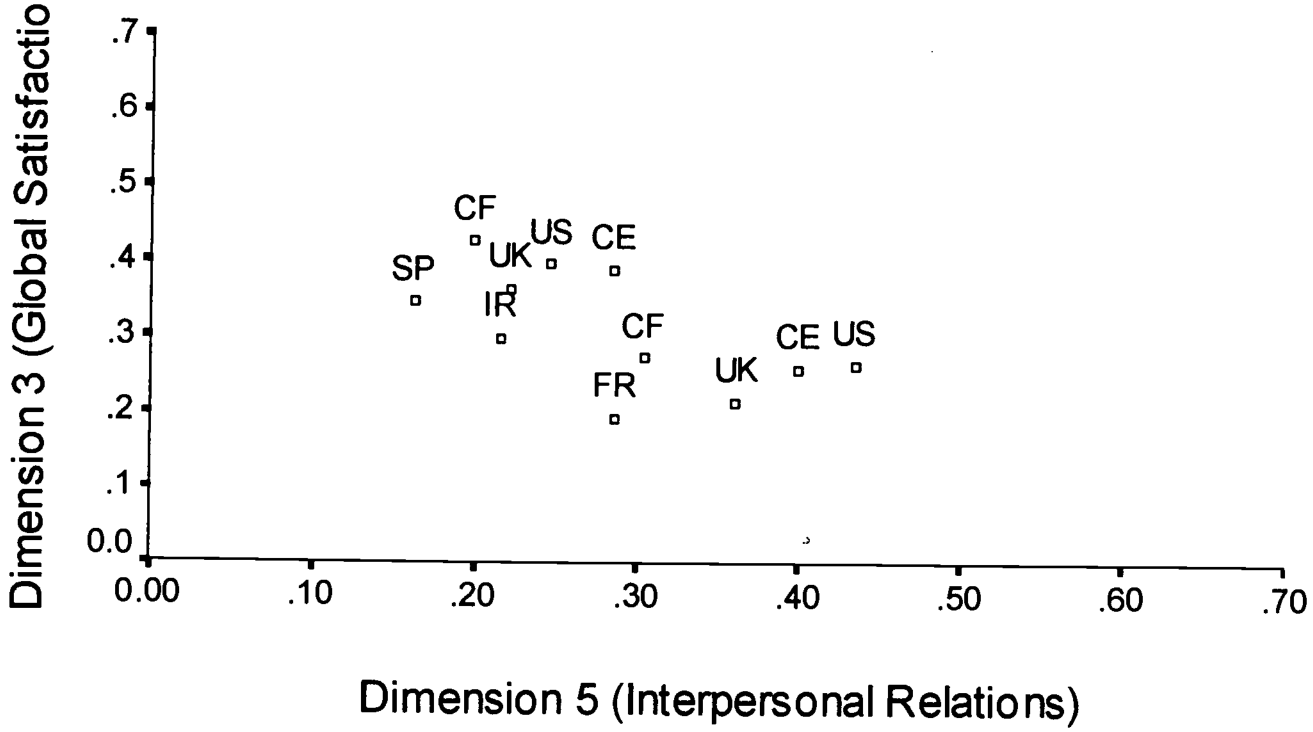
Web/Paper Group MDS Weights



P=Paper Suvey, W=Web Survey

Figure 3

Language Group MDS Weights



CE=Can. English, CF=Can. French, FR=French, IR=Ireland (English)

UK=United Kingdom, US=United States, SP=Spanish

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Evaluating the Construct Equivalence of International Employee Opinion Surveys

Author(s): Stephen G. Sireci, James Harter, Yongwei Yang, Dennison Bhola

Corporate Source: NCME? UMASS?

Publication Date: 4/25/00

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☑	Level 2A ↑ ☐	Level 2B ↑ ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Stephen G. Sireci Assoc. Prof.	
Organization/Address: School of Ed UMASS, Amherst, MA 01003-4140	Telephone: 413 545 0564	FAX:
	E-Mail Address: sireci@acad.umass.edu	Date: 6/5/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com