Short title: MEASUREMENT INVARIANCE STUDY

Full title: Measurement Invariance Study of the Training Satisfaction Questionnaire (TSQ)

Date of postage: May 27, 2011

Abstract

This article presents an empirical measurement invariance study in the substantive area of satisfaction evaluation in training programmes. Specifically, it (I) provides an empirical solution to the lack of explicit measurement models of satisfaction scales, offering a way of analysing and operationalizing the substantive theoretical dimensions; (II) outlines and discusses the analytical consequences of considering the effects of categorizing supposedly continuous variables, which are not usually taken into account; (III) presents empirical results from a measurement invariance study based on 5,272 participants' responses to a training satisfaction questionnaire in three different organizations and in two different training methods, taking into account the factor structure of the measured construct and the ordinal nature of the recorded data; and (IV) describes the substantive implications in the area of training satisfaction evaluation, such as the usefulness of the training satisfaction questionnaire to measure satisfaction in different organizations and different training methods. It also discusses further research based on these findings.

Keywords: invariance, satisfaction, evaluation, training method

Problem Statements

Lack of explicit measurement models of satisfaction scales

Although different evaluation issues have been described in many specialized reports (Barron, 1997; Birnbrauer, 1996; Lewis, 1996) the same degree of systematization and specialist literature is not available for already-validated specific instruments, or with respect to the analytic techniques to be used in evaluating satisfaction. Indeed, the specialist literature has not defined empirically and systematically the main theoretical dimensions to be taken into account when evaluating satisfaction in training programmes. Furthermore, training evaluation models present general definitions of constructs that can be translated into different empirical definitions of their dimensions (Kirkpatrick, 1999; Phillips, 1996a, 1996b; Shelton & Alliger, 1993).

In practice, differences in training programme performance are normally analysed by means of cross-group comparisons of participants' satisfaction, for example, tests of group mean differences (Averns, Maraschiello, van Melle, & Day, 2009; Hopper & Johns, 2007; Vanderberg & Lance, 2000). This type of analysis involves comparing levels of satisfaction among participants in different, specific aspects of the training process, such as content, trainers, objectives, methodology or utility (Basarab & Root, 1992; Kirkpatrick, 1999; Phillips, 1990). Measurement instruments, such as rating scales, are frequently developed in order to evaluate satisfaction (Konradt, Andreßen, & Ellwart, 2009; Ybema, Smulders, & Bongers, 2010).

The design of measurement instruments is based mainly on the idiosyncrasy of the organization and the nature of the context. Moreover, the absence of validated theoretical models with empirical definitions of their substantive dimensions has resulted in a lack of

well-defined measurement models. This means that it is difficult to study the construct validity of evidence obtained from existing rating scales, as their theoretical foundations or item inclusion criteria are not available (Gillespie, Denison, Haaland, Smerek, & Neale, 2008).

Invariance studies are needed

Accordingly, and based on a theoretical measurement model of satisfaction, it is necessary to direct efforts toward the study of invariance when measuring satisfaction as part of training programme evaluation. Doing so would provide a way of analysing and operationalizing the substantive theoretical dimensions usually found in the literature (e.g. Basarab & Root, 1992; Kirkpatrick, 1999; Phillips, 1990), as well as their possible generalization by recording data from samples or sub-samples of different populations (Díaz & Sánchez-López, 2004). This would reduce the probability of threats to validity such as the inadequate explication of constructs or their possible interactions with different organizational contexts (Shadish, Cook, & Campbell, 2002; reviewed by Chacón & Shadish, 2008). An adequate explanation of the constructs involved entails integrating the adequacy, meaning and utility of the inferences derived from scores obtained on measurement instruments (Messick, 1994). In this regard, the study of invariance measurement would support the correspondence between data and theory, help to increase theoretical coherence, and promote the usefulness of the scale in different contexts or situations (i.e. across organizations or training methods) (Bejar & Doyle, 1981; Masters & Hyde, 1984; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009; Vanderberg & Lance, 2000).

Ordinal measures considered as categorical

Another problem with the evaluation of satisfaction is that the measures used and their traditional forms of data analysis usually show low sensitivity in detecting differences between responses. Subjects tend to be assigned to the same assessment categories, despite being at different points on the assessment continuum, and this produces an important decrease in data variability. As a consequence, the results do not detect aspects that should be improved, unless these are particularly significant (Thayer, 1991).

However, analytic steps have been taken to avoid this lack of sensitivity when ordinal variables are used (Millsap & Tein, 2004). This enables a more appropriate study of construct validity with regard to one of its main aspects, the dimensionality of the instrument (Menjares, Michael, & Rueda, 2000).

Given that ordinal scales have neither a point of origin nor a measurement unit it is meaningless, when analysing subjects' responses at the item level, to calculate the means or variance-covariance (Holgado, Chacón, Barbero, & Vila, 2010). In order to study the association between these variables the only useful information is the number of cases in each cell of a bivariate contingency table. If, in this case, Pearson correlations are used to analyse the degree of association between ordinal variables, the values obtained will be lower because Pearson correlations reduce the magnitude of the coefficients obtained among observed variables (since the categorization reduces variability). As a consequence, the factor loadings obtained when factoring the correlation matrix will also be reduced, as there is not only a random error but also a category error effect (DiStefano, 2002; Saris, Van Wijk, & Scherpenzeel, 1998). Problems of estimation may therefore arise (Guilley & Uhlig, 1993).

However, if the subjects were able to be situated along the latent continuum, without category restrictions, the scores obtained would be different (Flora, Finkel, & Foshee, 2003; Jöreskog, 2001; Maydeu & D'Zurilla, 1995). In a Monte Carlo simulation study that

examined the influence of the number of categories, the cell probabilities, the population correlation (ρ) and sample size, Jöreskog and Sörbom (1996) found that polychoric correlations, a technique for estimating the correlation between two theorized normally distributed continuous latent variables from two observed ordinal variables (Holgado el al., 2010), were the most consistent and robust estimator. Use of the PRELIS and LISREL programs enabled data obtained from an ordinal scale to be analysed by estimating a matrix of polychoric correlations developed from categorical data and computing the asymptotic variance-covariance matrix for the estimation (Jöreskog & Sörbom, 1996).

General Objectives and Hypotheses

With the aim of providing useful solutions to these two criticisms (i.e. a lack of explicit measurement models of satisfaction scales that are directly related to empirical definitions of the dimensions included in the theoretical models, and the consequences of considering the effects of categorizing supposedly continuous variables), we present, based on a second-order factor satisfaction measurement model (Holgado, Chacón, Barbero, & Sanduvete, 2006), an empirical measurement invariance study of participants' satisfaction in different organizations and in different training methods (Meade, Michels, & Lautenschlager, 2007; Vanderberg & Lance, 2000). Our approach takes into account both the factor structure of the measured construct and the ordinal nature of the recorded data. We then discuss the substantive consequences as regards the evaluation of training satisfaction.

In the context of invariance studies our general hypotheses are as follows: (I) the measurement model remains stable across different organizations and in different training methods; and (II) inadequate decisions can be made when conducting cross-group

comparisons based on the observed scores obtained from item dimensions with different patterns of factor loading.

Specifically, we used the same Training Satisfaction Questionnaire (TSQ) (Holgado et al., 2006) to collect data from three different public organizations: a provincial council (PC), a university training centre for administrative and service staff (UT), and a regional sports institute (RS). Two different training methods were being applied in these organizations: *online* (via internet) and *traditional* (method involving direct contact between trainers and participants). These were the natural conditions of the study, which did not imply the presence of a completely randomized factorial design (3x2).

Theoretical and Methodological Framework

Multi-group confirmatory factor analysis

Multi-group confirmatory factor analysis with polychoric correlations was used to examine whether the measurement model remained stable across the three different organizations and the two training method groups (Del Barrio, Carrasco, & Holgado, 2006; Vanderberg & Lance, 2000). Briefly, in this analysis the initial null hypothesis is that given different groups the variance-covariance matrix is equal across groups. Rejection of this hypothesis implies the non-equivalence of the groups, such that we then need to search for the source of non-invariance (Jöreskog, 1971). When looking for evidence of multi-group invariance, researchers seek to answer one of the following five questions (Byrne, 1998): (I) Do the items that form a particular measurement instrument operate equivalently across different populations? (II) Is the factorial structure of an instrument equivalent across populations? (III) Is a causal structure invariant across different groups? (IV) Are the latent

means of a construct different across populations? (V) Is the factorial structure of a measurement instrument equivalent across independent samples of the same population?

Fit indices

In these analyses, categorical estimators may not be a viable alternative if the models have a large number of observable variables or sample sizes are small (Bollen, 1989a). Although, in theory, it is necessary to test the assumption of bivariate normality before calculating the polychoric correlation, this correlation is fairly robust with respect to such a violation (Coenders, Saris, & Satorra, 1997). It is therefore necessary to find alternative indices for detecting the lack of invariance. Jöreskog (2001) proposes using the root mean square error of approximation (*RMSEA*) as a fit index, as when its values are no greater than .1, parameter estimation is not significantly affected, even when the variables do not show bivariate normality. Chen, Sousa and West (2005), Chen (2007) and Kim (2005) discuss the use of the *RMSEA* and the comparative fit index (*CFI*), where for the latter a value above .95 is considered to be indicative of a good fit. Garver and Mentzer (see Hoe, 2008) recommend using the non-normed fit index (*NNFI*), where a value higher than .9 is considered to indicate a good fit. Finally, MacCallum and Hong (see Kim, 2005) propose the use of the goodness-of-fit index (*GFI*) and the adjusted goodness-of-fit index (*AGFI*), where values higher than .95 are indicative of a good fit in both indices.

Methods of estimation

When using Likert-type items and investigating the relationship between them by means of structural equation models the methods of estimation employed become particularly

important (DiStefano, 2002). The most popular among estimators based on normal distributions is the maximum likelihood (ML) method, as it finds consistent and asymptotically unbiased parameters (Bollen, 1989a).

However, if the variables are ordinal the relationships between them should be analysed using polychoric correlations, along with the asymptotic variance-covariance matrix as a weighting element in the estimation. In this process the weighted least squares (WLS) method, a particular case of the generalized least squares (GLS) procedure, is recommended when sample size is large but there are not too many variables in the given model (at least 12; Jöreskog & Sörbom, 1996).

In this regard, previous studies found that WLS showed a small bias in estimating parameters and this bias was reduced as sample size increased (DiStefano, 2002). Furthermore, when using WLS, GLS, unweighted least squares (ULS) and ML for polychoric correlations, it was shown that the factor loadings from WLS and ML were the closest; however, the standard errors for the estimated factor loading from WLS were the smallest (Bollen, 1989a). In conclusion, when using factor analysis to test a measurement model the scale used to measure the observable variables must be taken into account (Flora et al., 2003; Jöreskog, 2001; Maydeu & D'Zurilla, 1995).

Hypotheses in the Invariance Study

Specifically, three different hypotheses were tested in the invariance study (Byrne, 1998; Del Barrio et al., 2006):

(I) *The proposed second-order factor model is suitable across different organization and training method groups*. The aim is to test whether the measurement and structural model is common to the different organization and training method groups.

(II) *The pattern of factor loading is invariant across these groups*. Focusing on the measurement model the aim is to test the invariance pattern coefficients across organization and training method groups.

(III) *The structural model is equivalent in the defined groups*. To test the invariance of the structural model we firstly focus on the relationship between the construct and the different factors across groups, and then on the construct itself.

Method

Participants

The sample was purposive and comprised a total of 5,272 responses obtained during the year 2007 in three different organizations: 1,968 subjects were drawn from staff of the PC, 1,630 from the UT and 1,674 from the RS. Participants were able to attend once only those training activities related to their work. Thus, each case of the sample represents another independent subject. The TSQs were filled in just after finishing the training activity. In order to obtain a large sample size we sought to guarantee anonymity and, therefore, no demographic variables were recorded.

The data were obtained from 70 training events in the PC (13 online and 57 traditional), 81 in the UT (4 online and 77 traditional) and 68 in the RS (22 online and 46 traditional). Each of these training events had various standard series. Overall, 1,099 respondents participated in an online course, while the other 4,173 received traditional training.

The average duration of training events was 23 hours, with a range of 3 to 300 hours. Their content was varied (for example, law, sport services, quality management, libraries and

financial services), mainly because the professions and functions of participants were quite different (for example, police, psychologists, teachers, firemen, and gardeners).

In all three organizations professionals were trained in order to improve their skills and the quality of their work. Differences between them included the fact that the training programme implemented in the RS was newer than those in the other two organizations, and also that the PC invested a large amount of resources in 2007 in order to improve its training programme.

As regards the different training methods a clear advantage of the online method over the traditional one was that participants could organize their performance in a more flexible way. However, some aspects of the online method needed to be improved, mainly because this training method was relatively new; for example, further work was required in relation to how the trainers provided follow-up to participants, or in terms of adapting the training to participants without any knowledge of how to use computers.

Instruments

The steps to develop the TSQ were as follows (Holgado et al., 2006):

(I) *Search for items*. We reviewed available items from satisfaction questionnaires, mainly those used by different training organizations in Spain. This yielded 72 items, which were grouped into three dimensions: *Objectives and content*; *Method and training context*; and *Usefulness and overall rating*.

(II) *Content validity study*. Expert judges (specifically, 20 training centre managers and trainers from different universities and private training firms) were used to carry out a content validity study. The judges evaluated each item with respect to its representativeness (the extent to which the specific item represents the dimension to which it is assigned) and utility (the extent to which the specific item is useful for measuring satisfaction with respect to the

dimension to which it is assigned). Finally, each item was quantified through an index of congruence (Osterlind, 1998). A total of 21 items presented indices higher than .6 on both aspects (representativeness and utility).

(III) *Pilot study*. The psychometric properties of these 21 items were then studied after gathering data from 123 participants who attended UT training programmes.

(IV) *Final selection*. Of these 21 items we chose the 12 items with adequate psychometric properties (discrimination and item reliability from classical test theory) and the highest quantitative congruence indices. The resulting TSQ (Holgado et al., 2006) comprised 12 items scored on a 5-point scale (1 = *totally disagree*, 5 = *totally agree*). The final second-order measurement model of the construct *satisfaction* (SAT) comprised three factors (see Fig. 1): *Objectives and content* (F1), measured by items 1–4; *Method and training context* (F2), measured by items 5–7; and *Usefulness and overall rating* (F3), measured by items 8-12. In the present study the *structural model* is defined by the relationships between latent dimensions (SAT, F1, F2 and F3). Specifically, two parameters define these relationships: gamma (γ), which refers to the relationships between the second-order factor (SAT) and the first-order factors (F1, F2 and F3), and phi (φ), representing the variance-covariance of SAT. The *measurement model* comprises the relationships between factors (F1-F3) and items (ITEM 1-ITEM 12), as defined by the parameter lambda (λ). The specific content of each item is shown in Table 2.

[Insert Fig. 1]

(V) *Study of psychometric properties in the original sample (Holgado et al., 2006)*. The internal consistency, as measured by Cronbach's alpha coefficient, was .888. The average discrimination index was .674. In accordance with the theoretical development of the scale, a second-order factor model was tested and provided evidence of construct validity.

(VI) *Considering effect indicators instead of causal indicators.* Based on the scale development process and our empirical evidence we consider items as effect indicators (Bollen & Bauldry, 2011) for the following reasons: (I) According to the previous content validity study (representativeness), any change in the latent variable should lead to changes in observed indicators. (II) Dropping a single indicator does not affect the relationships between the remaining indicators or their relationships with the latent variable. In this regard, one might consider effect indicators with roughly the same reliability and validity as being interchangeable or non-essential. (III) Effect indicators that are positively associated with a latent variable should all have positive correlations with one another. Here we have positive correlations among effect indicators that are positively associated with the latent variable, with no low or negative significant correlations. If we had found low or negative correlations among a set of indicators with positive relationships to the latent variable, then this would have been evidence of either poor effect indicators or causal indicators.

Procedure

The data obtained were stored in SPSS 15.0 files. The internal consistency using Cronbach's alpha coefficient and the average discrimination index were calculated.

The matrix of polychoric correlations, using PRELIS (application included in LISREL 8.71), was estimated from different sub-samples (Flora et al., 2003) because items were considered as categorized continuous variables from a normal multivariate distribution (Holgado et al., 2010).

To justify the use of the matrix of polychoric correlations it was necessary to test the assumption of bivariate normality, calculating the percentage of tests that rejected the null

hypothesis of bivariate normality for each pair of correlations, assuming a nominal level of 5% and using the Bonferroni correction.

In addition, and following Jöreskog (2001), the percentage of correlations whose *RMSEA* was less than .1 was reported.

The theoretical invariance model was tested using a multi-group confirmatory factor analysis (CFA) (Bollen, 1989b). As polychoric correlations were being used the recommended methods of estimation were WLS and robust WLS because, in large samples and with fewer than 20 indicators, these methods provide consistent estimators (Flora & Curran, 2004; Holgado et al., 2010; Jöreskog & Sörbom, 1996).

Before studying the structure of invariance across different groups it was necessary to study the model proposed in Holgado et al. (2006) in each subgroup using the program LISREL, specifying as free every parameter in the PC, UT, RS, traditional and online groups separately, and studying the fit between this proposed measurement and structural model and the data collected here.

Specifically, the proposed second-order factor model (with operationalized substantive dimensions) was obtained from a previous content validity study, and was tested and shown to be adequate.

Hypothesis I: testing the validity of the second-order factor model (baseline model).

All parameters of the structural and measurement model were estimated simultaneously, conducting multi-group analyses without invariance constraints being imposed. The aim of these analyses was to obtain evidence that the model was common across different *organization* and *training method* groups. Indirectly, we tested whether different reports obtained from these two types of group were equivalent.

Hypothesis II: testing for the invariant pattern of factor loading

The following analyses focused on the measurement model. Multi-group analyses were performed, imposing equality constraints on pattern coefficients of lambda (λ_x) in order to test the invariance of the measurement model of the groups (Brown, 2006).

According to Ying and Fan (2003), increasing the constraints on a model leads to poorer model fit and, therefore, the different constraint models studied should be compared with a *baseline model* in order to assess the effect of these invariance constraints on the fit/misfit model. In this context other authors such as Browne and Du Toit (1992) and Cheung and Rensvold (2002) have suggested, respectively, using the root deterioration per restriction (*RDR*) statistic and changes in the comparative fit index (CFI).

As a first step the pattern coefficients of all the factors were constrained to be invariant across *organization* and *training method* groups. Subsequently, the invariance of each factor was analysed separately (maintained as equal across groups), before finally checking, again separately, the equivalence of item scores (maintaining equal across groups the parameter related to the specific item in question).

Hypothesis III: testing for invariance of the structural model.

The following analyses focused on the structural second-factor model, the aim being to assess equivalent relationships between the theoretical constructs (Vanderberg & Lance, 2000).

Firstly, the comparison models were obtained, constraining all lambda parameters (λ) to be equal across groups (Byrne, 1998). All gamma parameters (γ) were then constrained to test

the invariant structure across groups between the construct *satisfaction* and the three factors studied (Del Barrio et al., 2006). Finally, the factor phi (φ), related to the construct *satisfaction* was constrained to be equal across groups.

Results

Data Analysis

In the sample used in this study the internal consistency using Cronbach's alpha coefficient was .917 and the average discrimination index was .691.

Assumption of Bivariate Normality

Bivariate normality was tested using the matrix of estimated polychoric correlations. As the TSQ comprised 12 items a total of 66 correlations (12 × 11/2) were obtained. Results for all of them showed that this assumption was rejected at the significance level of $\alpha = .05/66 = .00075$ using the Bonferroni correction, which corresponds to a X^2 value of 38.52 with 15 degrees of freedom. Despite this, the *RMSEA* value was significantly lower than .1 in all cases. These results support the use of the matrix of polychoric correlations as the basis for the factor analyses.

The Proposed Model in Each Group

Fit indices presented in Table 1 show appropriate results in every group (the *organizations* PC, UT and RS; and the traditional and online *training methods*). However, the significance

of X^2 shows discrepancies with respect to the other indices, probably because this index is influenced by sample size.

[Insert Table 1]

In sum, the model fit for the different groups is considered to be adequate, and the model can thus be regarded as providing a reasonable representation of the data from groups. The standardized solutions for groups referring to *organizations* (PC, UT and RS) are presented in Table 2 and the correlations between factors in Table 3. Similarly, standardized solutions for different *training methods* (traditional and online) are shown in Table 4, with the correlations between factors given in Table 5.

[Insert Table 2]

[Insert Table 3]

[Insert Table 4]

[Insert Table 5]

High correlations and standardized solution values provide further evidence of the model's adequacy with respect to the data obtained from the different groups.

Hypothesis I: Testing the Validity of a Second-Order Factor Model (Baseline Model)

Table 6 shows the goodness-of-fit indices for the variables *organization* and *training method* separately (general *baseline* models 1 and 2, respectively). The values obtained support the existence of this common structure.

[Insert Table 6]

Hypothesis II: Testing for an Invariant Pattern of Factor Loading

The results (see Table 6) support the hypothesis of invariant pattern coefficients. Although the increment in X^2 (ΔX^2) compared to the baseline models was significant across both *organization* groups (comparing *1Baseline* model with model *1λF1,F2,F3*, where all the pattern coefficients of lambda (λ_x) were imposed as equal across organization groups) and *training method* groups (comparing *2Baseline* model with model *2λF1,F2,F3*, where all the pattern coefficients of lambda (λ_x) were imposed as equal across training method groups), the results for the other fit indices (*ECVI, RMSEA*, *GFI*, *CFI, NFI*, *NNFI* and *RDR*) were appropriate and enable us to assume equivalence across groups in the measurement model (Marsh, 1994). In all likelihood, the significance in ΔX^2 on this occasion was also due to the large sample size.

To obtain more details and to ensure that the significant ΔX^2 did not imply a lack of invariance, λ_x parameters were constrained as invariant in each factor across both *organization* groups (models *1λF1* meant that λ_x parameters were constrained in F1; *1λF2*, constrained in F2; and *1λF3*, constrained in F3) and *training method* groups (models *2λF1*, *2λF2* and *2λF3*). Except for the significant ΔX^2, the indices presented acceptable values.

Finally, to obtain even more detail an equal λ_x was imposed on each item across groups. Non-significant ΔX^2 were now found in models *1λIT1, 1λIT5* and *1λIT10* (where λ_x was

imposed as equal on items 1, 5 and 10 across *organization* groups), and in *2λIT1, 2λIT2*, *2λIT5*, *2λIT9*, *2λIT10* and *2λIT11* (where λ_x was imposed as equal on items 1, 2, 5, 9, 10 and 11 across *training method* groups).

Taking into account the acceptable values obtained for most of the models with the other fit indices (*ECVI*, *RMSEA*, *GFI*, *CFI*, *NFI*, *NNFI* and *RDR*) it can be concluded that the only measures which presented substantial differences across groups were those obtained in the item referring to the quality of the documentation given (IT8) across *organization* and *training method* groups, since models *1λIT8* and *2λIT8* revealed a non-fit in the results obtained, especially for *ECVI*, *RMSEA*, *GFI*, *NNFI* and *RDR*.

Hypothesis III: Testing for Invariance of the Structural Model

The comparison models for the *organization* and *training method* groups were, respectively, *1λF1,F2,F3* and *2λF1,F2,F3* (see Table 7).

[Insert Table 7]

As regards *organization*, goodness-of-fit indices (*ECVI*, *RMSEA*, *GFI*, *CFI*, *NFI*, *NNFI* and *RDR*) present acceptable values. Therefore, the invariance of the structural model across groups was confirmed, taking into account the influence of sample size on significant increases in X^2.

With respect to *training method*, the increase in the value of the X^2 test was not significant in the different models (*2γF1,F2,F3* and *2φ*). Given this result and the values of the descriptive fit indices (which coincide with those obtained in the comparison model

2λF1,F2,F3) the invariant hypothesis regarding the structural pattern coefficients cannot be rejected either.

Discussion and Conclusions

This invariance study in the area of training satisfaction evaluation sought to confirm that construct explication was adequate, and that the measurement model remained stable across different comparison groups. Furthermore, and due to the type of scale presented, polychoric correlations were used to analyse the relationship between the ordinal variables in this case. Both issues are crucial in a construct validation process.

Fit indices were used to determine whether the model adequately reproduced the relationships between variables, in this case in the substantive area of satisfaction evaluation in training programmes. It should be noted, however, that we do not seek to justify an absolute solution, but rather present a possible model that is consistent with the data and with our theoretical framework. Indeed, our main concern was to provide a useful decision-making tool for practitioners to use when making comparisons based on satisfaction measures in the real world. In this regard, we report necessary information that is rarely available in published papers and in an ongoing substantive area of research. This information could, for instance, be used to make satisfaction comparison inferences, effect size estimations or measurement error corrections.

The preliminary analysis revealed that the underlying structure fitted the data across groups, thus confirming that the different *organization* and *training method* groups were equivalent.

In relation to general hypothesis I the results obtained across *organization* and *training method* groups provide empirical evidence of invariance in the measurement model. This is

suggested by the optimal results obtained in the different descriptive fit indices used for the hierarchical confirmatory factor analysis models, despite the significant X^2 tests usually found due to sample size.

More specifically, the three hypotheses of the invariance study which involved increasingly restrictive models were confirmed: (I) the previously-tested second-order factor model is suitable across different *organization* and *training method* groups; (II) the pattern of factor loading is invariant across these different groups; and (III) the structural model is equivalent across the defined groups. The study of covariance matrix invariance is important to ensure that the lower-order factor errors are equivalent across the groups (Chen et al., 2005).

These results support the usefulness of the TSQ for measuring satisfaction in different organizations and in relation to different training methods.

As regards general hypothesis II, in both *organization* and *training method* groups the parameter related to item 8 (*The documentation given out was of good quality*) was not the same for different groups. Nevertheless, the scores obtained in different organizations and training methods may be comparable, except for those obtained on item 8. This finding supports the need to carry out an invariance study before conducting single-measure comparisons. Organizational decisions about the quality of documentation should not be made on the basis of the observed scores on item 8.

In this specific case, differences in scores across groups could be due to real differences across organization and training method groups, as previously described in the *Participants* section. As regards differences across organization groups, the PC invested the largest amount of resources to improve the quality of training documentation. With respect to differences across training method groups, one possible explanation of the differences in

satisfaction was that the online method was relatively new, so the corresponding documents may have undergone less correction and revision than in the traditional method.

At all events, further research is required, and in order to analyse the possible interaction between the aspect measured and the groups, we plan to perform a bias study via differential item functioning to test whether the same level of the studied concept entails differences in item 8 (Welkenhuysen-Gybels, 2004).

We would therefore like to invite any interested readers who are able and willing to measure satisfaction with training programmes to collaborate with this project.

References

Averns, H., Maraschiello, M., van Melle, E., & Day, A. (2009). Evaluation of a web-based teaching module on examination of the hand. *Journal of Rheumatology, 36*(3), 623-627.

Barron, T. (1997). Is there an ROI in ROI? *Technical and Skills Training, January,* 21-26.

Basarab, D. J., & Root, D. K. (1992). *The training evaluation process.* Boston: Kluwer.

Bejar, I. I., & Doyle, K. O. (1981). Factorial invariance in student ratings of instruction. *Applied Psychological Measurement, 5*(3), 307-312.

Birnbrauer, H. (1996). Improving evaluation forms to produce better course designs. *Performance and Instruction, 35*(1)*,* 14-17.

Bollen, K. (1989a). A new incremental fit index for general structural models. *Sociological Methods and Research, 17*(3)*,* 303-316.

Bollen, K. (1989b). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological Methods, 16*(3), 265-284.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Browne, M., & Du Toit, S. (1992). Automated fitting of non-standard models. *Multivariate Behavioral Research, 27*, 269-300.

Byrne, B. (1998). *Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming.* Hillsdale, NJ: Erlbaum.

Chacón, S., & Shadish, W. (2008). Validez en evaluación de programas [Validity in program evaluation]. In M. T. Anguera, S. Chacón, & A. Blanco (Coords.), *Evaluación de*

programas sociales y sanitarios. Un abordaje metodológico [Social and health program evaluation. A methodological approach] (pp. 69-102). Madrid: Síntesis.

Chen, F. F. (2007): Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471–492.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

Coenders, G., Saris, W., & Satorra, A. (1997). Alternative approaches to structural equation modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling, 4,* 261-282.

Del Barrio, M. V., Carrasco, M. A., & Holgado, F. P. (2006). Factor structure invariance in the Children's Big Five Questionnaire. *European Journal of Psychological Assessment, 22*(3), 158-167.

Díaz, J. F., & Sánchez-López, M. P. (2004). Composite and preferences scales of Morningness: reliability and factor invariance in adult and university samples. *The Spanish Journal of Psychology, 7*(2), 93-100.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327-346.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Flora, D. B., Finkel, E. J., & Foshee, V. A. (2003). Higher order factor structure of a self-control test: Evidence from confirmatory factor analysis with polychoric correlations. *Educational and Psychological Measurement, 63*(1), 112–127.

Gillespie, M. A., Denison, D. R., Haaland, S., Smerek, R., & Neale, W. S. (2008). Linking organizational culture and customer satisfaction: Results from two companies in different industries. *European Journal of Work and Organizational Psychology, 17*(1), 112-132.

Guilley, W., & Uhlig, G. (1993). Factor analysis and ordinal data. *Education, 114*, 258–264.

Hoe, S. L. (2008). Issues and procedures in adopting structural equation modeling technique. *Journal of Applied Quantitative Methods, 3*(1), 76-83.

Holgado, F. P., Chacón, S., Barbero, M. I., & Sanduvete, S. (2006). Training satisfaction rating scale. Development of a measurement model using polychoric correlations. *European Journal of Psychological Assessment, 22*(4), 268-279.

Holgado, F. P., Chacón, S., Barbero, M. I., & Vila, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*. *International Journal of Methodology, 44*, 153-166.

Hopper, K. B., & Johns, C. L. (2007). Educational technology integration and distance learning in respiratory care: practices and attitudes. *Respiratory Care, 52*(11), 1510-1524.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.

Jöreskog, K. G. (2001). Analysis of ordinal variables 2: Cross-Sectional Data. Text of the workshop *Structural Equation Modeling with LISREL 8.51.* Jena: Friedrich-Schiller-Universität Jena.

Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: User's reference guide.* Chicago: Scientific Software International.

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling, 12*(3), 368-390.

Kirkpatrick, D. L. (1999). *Evaluación de acciones formativas* [Training programs evaluation]. Barcelona: Epise.

Konradt, U., Andreßen, P., & Ellwart, T. (2009). Self-leadership in organizational teams: A multilevel analysis of moderators and mediators. *European Journal of Work and Organizational Psychology, 18*(3), 322-346.

Lewis, T. (1996). A model for thinking about the evaluation of training. *Performance Improvement Quarterly, 9,* 13-18.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: a multifaceted approach. *Structural Equation Modeling, 30*, 841-860.

Masters, G. N., & Hyde, N. H. (1984). Measuring attitude to school with a latent trait model. *Applied Psychological Measurement, 8*(1), 39-48.

Maydeu, A., & D'Zurilla, T. J. (1995). A factor analysis of the Social Problem-Solving Inventory using polychoric correlations. *European Journal of Psychological Assessment, 11*(2)*,* 98-107.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*(2), 322-345.

Menjares, P. C., Michael, W. B., & Rueda, R. (2000). The development and construct validation of a Spanish version of an academic self-concept scale for middle school Hispanic students from families of low socioeconomic levels. *The Spanish Journal of Psychology, 3*(1), 53-62.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10*(1), 1–9.

Millsap, R. E., & Tein, J.–Y. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. London: Kluwer Academic Publisher.

Phillips, J. J. (1990). *Handbook of training evaluation and measurement methods.* London: Kogan Page.

Phillips, J. J. (1996a). How much is the training worth? *Training and Development, April,* 20-24.

Phillips, J. J. (1996b). ROI: The search for best practice. *Training and Development, February,* 42-47.

Saris, W., Van Wijk, T., & Scherpenzeel, A. (1998). Validity and reliability of subjective social indicators. *Social Indicators Research, 45,* 173-199.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton-Mifflin.

Shelton, S., & Alliger, G. M. (1993). Who's afraid of level 4 evaluation? A practical approach. *Training and Development Journal,47,* 43-46.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity*. *International Journal of Methodology, 43*(4), 599-616.

Thayer, P. (1991). A historical perspective on training. In I. L. Goldstein & Associates (Eds.), *Training and development in organizations* (pp. 457-468). San Francisco: Jossey-Bass.

Vanderberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1)*,* 4-70.

Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality & Quantity*. *International Journal of Methodology, 38*(6), 681–702.

Ybema, J. F., Smulders, P. G. W., & Bongers, P. M. (2010). Antecedents and consequences of employee absenteeism: A longitudinal perspective on the role of job satisfaction and burnout. *European Journal of Work and Organizational Psychology, 19*(1), 102-124.

Ying, P., & Fan, X. (2003). Assessing the factor structure invariance of self-concept measurement across ethnic and gender groups: findings from a national sample. *Educational and Psychological Measurement, 63*(2), 296-318.

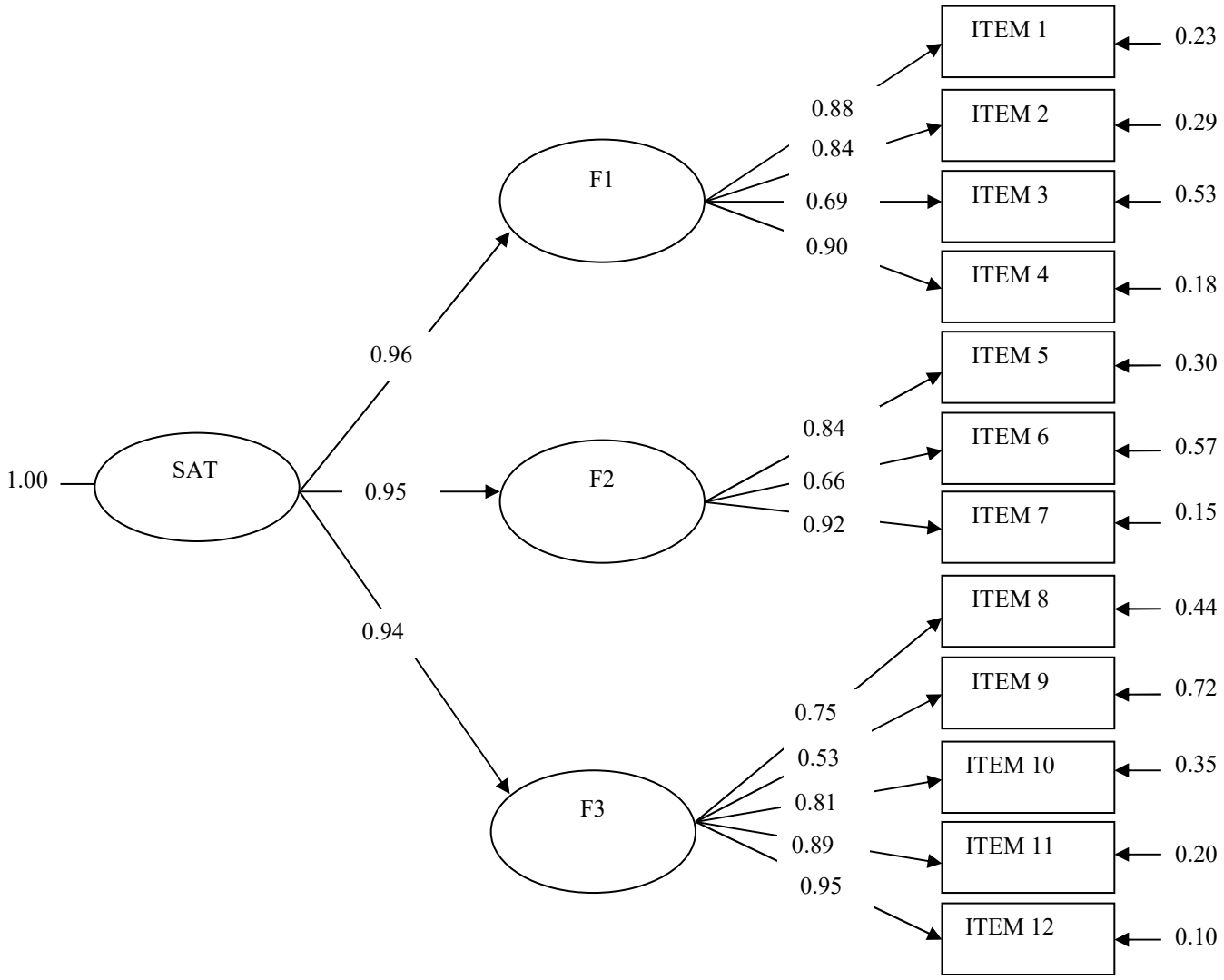

Figure 1. Second-order construct satisfaction measurement model obtained by Holgado et al. (2006).

Note: SAT: Satisfaction; F1: Objectives and content; F2: Method and training context; F3: Usefulness and overall rating.

Table 1

Fit indices in the different groups (a provincial council (PC), a university training centre for administrative and service staff (UT) and a regional sports institute (RS)) and for traditional and online methods.

	X^2	ρ	df	GFI	AGFI	CFI	RMSEA
PC	401.10	<.001	51	.99	.98	.98	.059
UT	419.64	<.001	51	.99	.98	.98	.067
RS	418.07	<.001	51	.99	.98	.96	.066
TRADITIONAL	798.06	<.001	51	.99	.98	.97	.059
ONLINE	367.55	<.001	51	.98	.97	.96	.075

Table 2

Standardized solutions for the three organization groups (a provincial council (PC), a university training centre for administrative and service staff (UT) and a regional sports institute (RS)).

Items	F1			F2			F3		
(λ)	PC	UT	RS	PC	UT	RS	PC	UT	RS
4. The method was well suited to the objectives and content	.96	.96	.93	---	---	---	---	---	---
1. In my opinion the planned objectives were met	.93	.93	.91	---	---	---	---	---	---
2. The issues were dealt with in as much depth as the length of the course allowed	.90	.92	.85	---	---	---	---	---	---
3. The length of the course was adequate for the objectives and content	.76	.71	.64	---	---	---	---	---	---
7. The training was realistic and practical	---	---	---	.96	.95	.90	---	---	---
5. The method used enabled us to take an active part in training	---	---	---	.90	.91	.81	---	---	---
6. The training enabled me to share professional experiences with colleagues	---	---	---	.75	.72	.75	---	---	---
12. The training merits a good overall rating	---	---	---	---	---	---	.98	.98	.96
11. The training received is useful for my personal development	---	---	---	---	---	---	.95	.90	.92
9. The training context was well suited to the training process	---	---	---	---	---	---	.88	.63	.66
8. The documentation given out was of good quality	---	---	---	---	---	---	.84	.77	.72
10. The training received is useful for my specific job	---	---	---	---	---	---	.83	.80	.81
SAT (γ)	.97	.99	.97	.96	.98	.94	.96	.97	.94

Table 3

Correlations between factors for the three different organizations studied (a provincial council (PC), a university training centre for administrative and service staff (UT) and a regional sports institute (RS)).

	F1			F2		
	PC	UT	RS	PC	UT	RS
F1	---	---	---	---	---	---
F2	.93	.97	.91	---	---	---
F3	.93	.95	.92	.93	.95	.88

Table 4

Standardized solutions for the two training method groups (Traditional (TRAD) and Online).

Items	F1		F2		F3	
(λ)	TRAD	ONLINE	TRAD	ONLINE	TRAD	ONLINE
4	.94	.95	---	---	---	---
1	.92	.92	---	---	---	---
2	.88	.88	---	---	---	---
3	.69	.77	---	---	---	---
7	---	---	.93	.95	---	---
5	---	---	.86	.89	---	---
6	---	---	.65	.69	---	---
12	---	---	---	---	.97	.97
11	---	---	---	---	.92	.90
10	---	---	---	---	.82	.85
8	---	---	---	---	.74	.82
9	---	---	---	---	.71	.73
SAT (γ)	.97	.99	.96	.97	.95	.97

Table 5

Correlations between factors for the two different training methods (Traditional (TRAD) and Online).

	F1		F2	
	TRAD	ONLINE	TRAD	ONLINE
F1	---	---	---	---
F2	.81	.87	---	---
F3	.84	.89	.82	.87

Table 6

Invariance of factor loadings

Model	*X^2 (ΔX^2)*	*df (Δ df)*	*ECVI*	*RMSEA*	*GFI*	*CFI*	*NFI*	*NNFI*	*RDR*
ORGANIZATION GROUPS (1)									
1Baseline	**1238.80**	**153**	**.27**	**.064**	**.99**	**.97**	**.97**	**.97**	**---**
1λF1,F2,F3	(210.17)*	(18)	.3	.065	.99	.97	.97	.96	.001
1λF1	(28.57)*	(6)	.27	.063	.99	.97	.97	.97	0
1λF2	(28.31)*	(4)	.27	.063	.99	.97	.97	.97	0
1λF3	(166.14)*	(8)	.29	.066	.99	.97	.97	.96	.003
1λIT1	(.64)	(2)	.27	.063	.99	.97	.97	.97	0
1λIT2	(10.54)*	(4)	.27	.063	.99	.97	.97	.97	0
1λIT3	(19.63)*	(4)	.27	.063	.99	.97	.97	.97	0
1λIT5	(7.92)	(4)	.27	.063	.99	.97	.97	.97	0
1λIT6	(28.69)*	(6)	.27	.063	.99	.97	.97	.97	0
1λIT8	(3055.75)*	(6)	.84	.120	.95	.90	.90	.88	.095
1λIT9	(160.38)*	(6)	.29	.067	.99	.97	.97	.96	.005
1λIT10	(12.59)	(6)	.27	.063	.99	.97	.97	.97	0
1λIT11	(21.93)*	(8)	.27	.062	.99	.97	.97	.97	.004
TRAINING METHOD GROUPS (2)									
2Baseline	**1165.61**	**102**	**.24**	**.063**	**.98**	**.97**	**.97**	**.96**	**---**
2λF1,F2,F3	(43.43)*	(9)	.25	.061	.98	.97	.97	.96	.001
2λF1	(14.24)*	(3)	.24	.062	.98	.97	.97	.96	0
2λF2	(12.42)*	(2)	.24	.063	.98	.97	.97	.96	0
2λF3	(20.58)*	(4)	.24	.062	.98	.97	.97	.96	0
2λIT1	(1.42)	(1)	.24	.063	.98	.97	.97	.96	0
2λIT2	(1.89)	(2)	.24	.062	.98	.97	.97	.96	0
2λIT3	(14.24)*	(3)	.24	.062	.98	.97	.97	.96	0
2λIT5	(2.27)	(3)	.24	.062	.98	.97	.97	.96	0
2λIT6	(14.25)*	(4)	.24	.062	.98	.97	.97	.96	0
2λIT8	(2865.45)*	(4)	.56	.099	.90	.92	.92	.91	.08

2λIT9	(2.69)	(4)	.24	.062	.98	.97	.97	.96	0
2λIT10	(5.29)	(5)	.24	.061	.98	.97	.97	.96	0
2λIT11	(9.99)	(6)	.24	.061	.98	.97	.97	.96	0

Note. * indicates significant increment in chi-squared ($\rho<.05$) compared to the baseline model

Table 7

Invariance of the structural model

Model	*X^2 (ΔX^2)*	*df (Δ df)*	*ECVI*	*RMSEA*	*GFI*	*CFI*	*NFI*	*NNFI*	*RDR*
ORGANIZATION GROUPS (1)									
1λF1,F2,F3	**1448.97**	**171**	**.3**	**.065**	**.99**	**.97**	**.97**	**.96**	**---**
1γF1,F2,F3	(20.02)*	4	.3	.068	.98	.97	.97	.97	.002
1φ	(24.94)*	2	.3	.065	.98	.97	.97	.95	.005
TRAINING METHOD GROUPS (2)									
2λF1,F2,F3	**1209.04**	**111**	**.25**	**.061**	**.98**	**.97**	**.97**	**.96**	**---**
2γF1,F2,F3	(1.19)	2	.25	.061	.98	.97	.97	.96	0
2φ	(1.57)	1	.25	.061	.98	.97	.97	.96	0

Note. * indicates significant increment in chi-squared ($\rho<.05$) compared to the baseline model