Con formato: Fuente de párrafo predeter., Fuente: (Predeterminada) Times New Roman, 12 pto, Sin Negrita, Color de fuente: Negro, Diseño: Claro

Con formato: Fuente: (Predeterminada) Arial, 8,5 pto, Sin Negrita, Color de fuente: Negro, Diseño: Claro (Color personalizado(RGB(245;245;245)))

TRAINING SATISFACTION RATING SCALE: DEVELOPMENT OF A MEASUREMENT MODEL USING POLYCHORIC CORRELATIONS [1]

Authors

Francisco Pablo Holgado Tello
Dpto. Metodología de las Ciencias del Comportamiento
Facultad de Psicología. UNED
c/ Juan del Rosal, nº 10.
28040 Madrid
Spain
Tel. +34 91 3988648
e-mail: pfholgado@psi.uned.es

Salvador Chacón Moscoso
Metodología de las Ciencias del Comportamiento
Facultad de Psicología. Universidad de Sevilla
c/ Camilo José Cela
41018 Sevilla
Spain
Tel. +34-95-4557812; Fax: +34-95-4551784
e-mail: schacon@.us.es

Isabel Barbero García
Dpto. Metodología de las Ciencias del Comportamiento
Facultad de Psicología. UNED
c/ Juan del Rosal, nº 10.
28040 Madrid
Spain
Tel. +34 91 3987900
e-mail: mbarbero@psi.uned.es

Susana Sanduvete Chaves
Metodología de las Ciencias del Comportamiento
Facultad de Psicología. Universidad de Sevilla
c/ Camilo José Cela
41018 Sevilla
Spain
Tel. +34954554331; Fax: +34-95-4551784
e-mail: sussancha@.us.es

[1] The present study forms part of the results obtained in research project BSO2000-1462, funded by Spain's *Ministerio de Ciencia y Tecnología.*

TRAINING SATISFACTION RATING SCALE: DEVELOPMENT OF A MEASUREMENT MODEL USING POLYCHORIC CORRELATIONS

Abstract

We describe the process of testing a measurement model of a satisfaction rating scale for use with training programs. The scale was developed by the Evaluation Unit of the University of Seville's Training Centre, having as general framework of reference Kirkpatrick's (1999) training evaluation model. Following an initial content validity study in which we reviewed how training evaluation is modelled, a 12-item rating scale was developed and administered to a sample of 2,746 subjects. The measurement model was examined through an exploratory factor analysis using polychoric correlations, the results of which were consistent with the previous theory. In order to refine the model under study a confirmatory factor analysis (also based on polychoric correlations) was then performed. The results showed factor validity to be adequate and would thus seem to support use of the scale for measuring satisfaction with training.

Key words: satisfaction, training, polychoric correlations, factor analysis.

1. Introduction

The aim of the present study was to test the validity of a rating scale developed to measure satisfaction among people attending various training programs. It was conducted in the University of Seville's Training Centre for Administrative and Service Personnel. The main objective of the Centre's Evaluation Unit is to plan and systematically evaluate the whole training process, the philosophy being that evaluation is the best way of ensuring that training becomes a tool for improving institutional functioning. Given this approach, and in accordance with current developments in intervention programs, the Centre emphasises the continuous interaction between intervention and evaluation.

Starting from the four-level model of training evaluation (Kirkpatrick, 1999) we focus mainly on one of the aspects related to the evaluation of training outcomes: the measurement of participants' reactions to training. In practice this type of analysis involves analysing levels of satisfaction among participants with respect to various aspects of the training process (Phillips, 1990; Basarab & Root, 1992; Ventosa, 1998; Kirkpatrick, 1999). Semi-standard instruments, such as rating scales, are frequently developed in order to evaluate satisfaction, and their design is based mainly on the concrete characteristics of the organisation and the specific training context. This practice has implied that specialized literature has not systematized main theoretical dimensions to take into account in satisfaction evaluation in relation to Kirkpatrick or other possible models. Nonetheless we consider that developing a feasible 'standard' scale to evaluate satisfaction in training programs in different organizations is useful.

Another criticism of this type of evaluation is illustrated by the term that is often used to describe it: "happy sheets". This refers to the fact that such measures and their traditional data analysis present a low degree of sensitivity to detect specific differences between obtained records. Subjects are assigned to the same assessment categories despite of being in different points of the assessment continuum, which implies a relevant decrease of data variability. Then, are unable to detect concrete aspects in need of improvement, unless these are particularly significant (either

positive or negative) elements of the intervention program (Thayer, 1991). However, analytic advances have been made which consider the effects of categorising supposedly continuous variables, and these enable a more appropriate study of construct validity with regard to, for example, the dimensionality of the instrument.

We understand validity as a unitary concept that integrates the adequacy, meaning and utility of the inferences derived from scores obtained on measurement instruments. All this is based on the central aspect of construct validity, where scientific criteria (representativeness and utility) and social values (consequences of applying the measure) converge (Messick, 1994; APA, AERA, NCME, 1999; Muñiz, 2004). In a complementary way to study validity, an approach to criterion reference measurement could be interested in order to define performance standards of persons with a given score in the satisfaction scale (Linn, 1994).

In relation to the two main weaknesses mentioned above (lack of available validated satisfaction scales directly related to theoretical models and the consequences of considering the effects of categorizing supposedly continuous variables), the aim of the present study was to test a measurement model of a satisfaction rating scale developed by the Evaluation Unit of the University of Seville's Training Centre. To this end the study is divided into two parts. Firstly, at the conceptual level, we outline Kirkpatrick's (1999) evaluation model in training programs as it is the most cited one (68 citations in the last 25 years, 'cited reference index' in the ISI web of Knowledge database) and we consider that it can provide a general framework for the measurement of satisfaction, and then move on to discuss some of the consequences of treating ordinal variables as continuous variables in factor analysis. In the second, empirical part we describe the method and results obtained from the process of validating the scale, as well as the conclusions drawn. The specific process used was cross validation: after conducting exploratory factor analysis (EFA) using a matrix of polychoric correlations we tested the structure based on theoretical coherence and results obtained in EFA by means of a confirmatory factor analysis (CFA), also based on polychoric correlations.

2. The four levels of training evaluation

The framework in which the satisfaction of participants in continuous training programs is usually measured is a final evaluation model known as *Kirkpatrick's four-level evaluation model* (Kirkpatrick, 1999). This model enables development of a working approach that provides information regarding the following issues: a) the satisfaction of participants; b) the knowledge, skills and attitudes learnt through training; c) whether or not participants have changed their behaviour in the workplace as a result of the training received; and d) whether these changes have had a positive effect on the organization.

The present study focuses specifically on the first issue (evaluation of satisfaction) which involves recording and evaluating the reactions of participants to the training. This is usually done through satisfaction questionnaires that cover various aspects of the training, such as contents, trainers, objectives, methodology or utility (Salanova & Grau, 1999; Ventosa, 1998).

Although these evaluation issues have been widely described in many specialised reports (Shelton & George, 1993; Lewis, 1996; Birnbrauer, 1996; Phillips, 1996a, 1996b; Willyerd, 1996; Barron, 1997; Kirkpatrick, 1999; Fernández, 2000; Pineda, 2000) the same degree of systematisation and specialised literature is not available for already validated specific instruments or with respect to the analytic techniques to be used. We haven't found any available validated satisfaction scale directly related to Kirkpatrick or other possible models. Furthermore, a questionable feature of most models is that they are excessively summative, especially given that current trends in evaluation consider the mutual dependence between intervention and evaluation from a non-linear perspective (Shadish, Cook & Leviton, 1991; Chelimsky, 1997; Chacón, Anguera, Perez & Holgado, 2002).

We believe that satisfaction levels can be the starting point for both the evaluation of results and the introduction of improvements into the training process. In this field, satisfaction is usually measured with rating scales and these have become widely used as they are easy to develop and administer. It is assumed that, through a system of categories, these scales represent supposedly

continuous latent variables whose range has been restricted. As will be discussed below, this has a number of analytic repercussions when it comes to testing construct validity.

3. Effects of categorisation in factor analysis

When using rating scales, in which, as pointed out above, subjects' responses to each item are restricted in accordance with the categories established, the only thing that is assumed is that the person choosing a given category possesses this characteristic to a greater degree than if he or she had chosen a lower category; however, nothing more is known. As the measured variables are ordinal they cannot be treated as if they were continuous, although this often occurs in practice. 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. In order to study the association between variables of this kind the only useful piece of 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 validity of the scale, and quantify the degree of association between ordinal variables lacking a metric scale, the values obtained will be lower as all subjects situated at different points of the interval will be assigned the same score; however, if the subjects were able to be situated along the latent continuum, without the category restrictions, the scores obtained may be different. As pointed out above, this is because Pearson correlations reduce the magnitude of the coefficients obtained among observed variables due to the fact that the categorisation reduces variability; therefore, problems of estimation may arise (Guilley & Uhlig, 1993). Consequently, 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 (Saris, Van Wijk & Scherpenzeel, 1998; DiStefano, 2002). It is therefore important, when using factor analysis to test a measurement model, to take into account the type of scale used for measuring the observable variables (Maydeu & D'Zurilla, 1995; Jöreskog, 2001; Flora, Finkel & Foshee, 2003).

Jöreskog and Sörbom (1996b), 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, found polychoric correlations to be the most consistent and robust estimator. The PRELIS and LISREL programs enable 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, 1996a, 1996b). Given these findings the factor analyses in the present study were carried out using matrices of polychoric correlations.

Of course, categorical estimators may not be a viable alternative if the models have a large number of observable variables or sample sizes are small (Bollen, 1989).

The basis of polychoric correlations is as follows. Let us suppose that Z_1 and Z_2 are two ordinal items with m_1 and m_2 categories. Their distribution in the sample is given by the contingency table. If we suppose that underlying these items are variables Z^*_1 and Z^*_2, which are normally distributed, it can be assumed that their combined distribution is a normal bivariate distribution with a correlation ρ. The polychoric correlation is the correlation ρ in the bivariate normal distribution of the latent variables Z^*_1 and Z^*_2. If $m_1 = m_2 = 2$ then the correlation is tetrachoric.

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). Furthermore, given the sensitivity of chi-square tests, particularly in large samples, it is necessary to find alternative statistics for evaluating the assumption of normality. Jöreskog (2001) proposes using the root mean square error of approximation (RMSEA) as a fit index, as when its values are no greater than 0.1 parameter estimation is not significantly affected, even when the variables do not show bivariate normality.

It can be seen from the above that the metric of the data influences how they should be analysed. Thus, 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, 1989). However, if the variables are ordinal, the relationships between them should be analysed using polychoric correlations, using 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 generalised 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, 1996b, p. 171). In fact, a simulation study carried out by DiStefano (2002) found that WLS showed a small bias in estimating parameters and that this bias was reduced as sample size increased. On the other hand, in another simulation conducted by Bollen (1989) using WLS, GLS, unweight least squares (ULS) and ML for polychoric correlations found that the factor loading from WLS and ML are the closest. However, the standard errors are the smallest for the estimated factor loading from WLS.

In sum, when using factor analysis to test a measurement model, the scale used to measure the observable variables must be taken into account (Maydeu & D'Zurilla, 1995, Jöreskog, 2001; Flora, Finkel & Foshee, 2003).

4. Method

4.1. Subjects

The sample was purposive and comprised 2,746 subjects chosen from among administrative and service personnel who had participated in training programs run by the University of Seville's Training Centre. We try to guarantee the anonymity in order to obtain the high sample size, for that reason we do not record any demographical variable.

The data were obtained along 78 training actions with various standard editions for each one. The average duration of the training actions was 20 hours, the shortest one had 9 hours and the longest 40. The contents of the training actions were quite different (law, sport services, quality management, libraries, training of trainers, economic services, …-all available training action for each year can be found in 'www.forpas.us.es'). The participants were able to attend only once to

each training actions related to their work. Then, each case of the sample represents another independent subject.

4.2. Instruments

The instrument used for the study was a rating scale in which participants were asked to evaluate different aspects of the training: objectives and content, method and usefulness.

The questionnaire was developed from a review of other tests and measures of satisfaction used by different training units from Spanish Universities, as well as on the basis of contributions by the Centre's own managers. We performed a content validity study, through an expert judge – Training Centres managers and trainers- (Osterlind, 1998). From an initial version containing 72 items a final 12-item, five-point (1 = totally disagree; 5 = totally agree) scale was obtained, the items being grouped into three dimensions (Holgado, 2002): a) Objectives and content (items 1-3); b) Method and training context (items 4-9), and; c) Usefulness and overall rating (items 10-12). The specific content of each item is shown in Table 2. In appendix one, we present the list of the original 72 items and the main bibliographical references of the reviewed questionnaires.

For the sample used the scale had a Cronbach α coefficient of 0.888, and a mean discrimination index of 0.674.

The software used for data analysis and processing was the PRELIS 2.30 and LISREL 8.54 programs (Jöreskog & Sörbom, 2003).

4.3. Procedure

First, the sample was randomly divided into two sub-samples (sub-samples A and B) of equal size (n = 1373). Sub-sample A was used for the exploratory factor analysis and sub-sample B for the confirmatory analysis. Both procedures made use of a matrix of polychoric correlations (Flora, Finkel & Foshee, 2003).

Once the matrix of polychoric correlations had been estimated, the assumption of bivariate normality was tested. This was done by 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 for the complete set of correlations. In addition, and following Jöreskog (2001), the percentage of correlations whose RMSEA was less than 0.1 was reported.

EFAs were then carried out with sub-sample A, we used the ML method of estimation along with a Promax oblique rotation and with a Varimax orthogonal rotation to test different competing models (1, 2 and 3 factor solutions).

Then, a CFA was carried out to test one of those models obtained in the previous EFA that was the most coherent with the theoretical criteria used in developing the scale (we present an exhaustive description of the model in section 5.2. CFA). This resulting model was tested with sub-sample B, using the asymptotic variance-covariance matrix as the weighting element in the WLS procedure.

Due to the ordinal character of the data the assumption of multivariate normality probably was violated, and we used both ML and WLS in order to check the robustness of the results.

5. Results

As the scale comprised 12 items a total of 66 correlations (12 * 11/2) were obtained. In 41 of these the assumption of bivariate normality was rejected at a significance level of 0.00075 (α= 0.05/66), which corresponds to a chi-square value of 38.52 with 15 degrees of freedom. Despite this high number of correlations the RMSEA value was significantly lower than 0.1 in all cases. These results support the use of the matrix of polychoric correlations as the basis for the factor analyses. Table 1 shows both the polychoric (below the diagonal) and the Pearson correlations (above the diagonal). It can be seen that all the polychoric correlations obtained are higher than the Pearson values.

INSERT TABLE 1: Matrix of polychoric (below the diagonal) and Pearson correlations (above the diagonal)

5.1. Exploratory factor analysis

As mentioned before, we carried out different EFAs, nonetheless, here we only present the latent three-factor structure because its loading pattern structure is the most coherent with the conducted content validity study. The first factor explained 51% of the variance, the second 5%, the third 3.6%, and 2.9% the fourth one. Although the scale could be considered to be unidimensional due to the small percentage of variance explained by the rest of factors, they were included on theoretical grounds. Table 2 shows the factor loadings of the rotated solution for each of the items with respect to the three factors. The highest loadings are shown in bold type.

INSERT TABLE 2. Loadings pattern and factor correlations (cursive)

It can be seen that the factor structure is similar to that proposed in the satisfaction survey that was obtained according with three theoretical dimensions of training satisfaction (Phillips, 1990, Basarab & Root, 1992, Salanova & Grau, 1999; Ventosa, 1998; Kirkpatrick, 1999) based on the content validity study. In the first factor the items with the highest loadings refer to objectives and content, item 1 (*In my opinion the planned objectives were met)* having the highest loading (0.751). The second factor mainly consisted of those items referring to method and training context, and here it was item 5 (*The method used enabled us to take an active part in training)* which had the highest loading (0.842). Finally, the third factor included those items that concerned the usefulness of training and the overall rating awarded it, item 11 (*The training received is useful for my personal development)* being the one with the highest factor loading here (0.855).

The results obtained suggest, with certain qualifications, that the structure underlying subjects' responses to the survey items fits the theoretical approach used in developing it. As pointed out above this approach involved an exhaustive review of other questionnaires used in various institutions, see appendix one, (Holgado, 2002), as well as taking into account both the contributions of the Training Centre managers and the analysis of theoretical aspects related to the dimensions involved in training processes (Phillips, 1990; Basarab & Root, 1992; Salanova & Grau, 1999; Ventosa, 1998; Kirkpatrick, 1999).

The matrix of correlations between factors is shown in Table 2; it can be seen that the correlations are high, indicating the possible existence of a second order factor.

5.2. Confirmatory factor analysis

The criteria used to define the structure to be validated were theoretical coherence to content validity study and the loadings obtained in the prior tri-factor EFA (Table 2). Factor 1 is related to objectives and content; it consisted of items 1, 2, 3 and 4. Items 1, 2 and 3 matched the initial factor assignment obtained in content validity study; and item 4 was incorporated into factor 1 as it referred to objectives and loaded on this first factor in the EFA. Factor 2 is related to method and training context; it comprised items 5, 6 (matched the initial factor assignment) and 7 (referred to practical dimension of the method and loaded on this second factor in EFA). Factor 3 is related to usefulness and overall rating; it consisted of items 8, 9, 10, 11, 12. Items 10, 11 and 12 (matched the initial factor assignment); items 8 and 9 (present low loadings on factor 2 in EFA and were also conceptually consistent with factor 3; related to overall rating). We defined Model 1 based on this tri-factorial structured.

The value obtained in model 1 with 51 degrees of freedom was $\chi^2 = 358.78$ with $p = 0.0001$. In light of these results other fit indices were used: the goodness-of-fit index (GFI), the adjusted goodness-of-fit index (AGFI), and the abovementioned root mean square error approximation (RMSEA) (see Table 3).

Due to the correlation found between the three factors a second-order factor analysis was conducted (model 2).

The fit indices obtained for both model 1 (three-factor) and model 2 (second-order factor) are shown in Table 3.

INSERT TABLE 3: Fit indices of models 1 and 2 *(p = 0.048)

Figure 1 shows the representation obtained for model 2. This model represents the data adequately.

INSERT FIGURE 1: Model 2

It can be concluded from these results that, in general, the fit indicators are adequate as for both indices (GFI and AGFI) a value of 0.90 or above is considered indicative of a good fit (Bollen & Long, 1993; Hopko, 2003; Byrne, 2001; Jöreskog, 2001). The RMSEA value is regarded as adequate when it is below 0.05, although values of up to 0.08 are considered to represent reasonable errors of approximation (Browne & Cudeck, 1993). In addition to the above indicators the expected cross validation index (ECVI) and the consistent Akaike information criterion (CAIC) were also calculated. Both these indices are used to measure the comparative fit between two or more models, and the smaller the obtained values the better the fit (Bandalos, 1993).

It can be seen that the GFI, AGFI and ECVI indices are the same for the two models. The significant increase (p=0.048) in Chi-square test between both models implies a decrease in the goodness of fit. On the other hand, both the RMSEA and the CAIC present lower values for model 2. In sum, although both models are very close to each other, model 2 is also more coherent with the theoretical model used in developing the scale in the sense that it considers a construct, defined as satisfaction with training, which explains the proposed dimensions.

6. Discussion and Conclusions

Although Kirkpatrick's model represents the most cited general frame work to evaluate training programs, we didn't find any specific available validated satisfaction rating scale designed in accordance with Kirkpatrick's (1999) four-level evaluation model to measure the reaction of participants to training in specialized literature (Phillips, 1990; Basarab & Root, 1992; Ventosa, 1998; Kirkpatrick, 1999) nor in usual practice. The first stage of this study was to develop a satisfaction scale based on a theoretical review of the dimensions involved in training (aspects such as objectives, utility, or methodology are usually considered). Then we carried out a content validity study, through an expert judge –Training Centres managers and trainers-, the aim being to ensure the validity of the scale's content and consider its consequences and utility.

Focusing on the main objective of this paper we have studied factor structure of the proposed scale. Testing a measurement model is a key aspect in this regard, as it lends support to

the correspondence between data and theory, the theoretical coherence, and the usefulness and consequences of the scale. The analysis of a measurement model must be consistent with the measurement procedure used as this can affect the results obtained (Guilley & Uhlig, 1993; Flora, Finkel & Foshee, 2003). Thus, on the basis of the EFA with polychoric correlations a given measurement model was outlined from the items used to measure satisfaction. This model coincided, with certain qualifications, with the approach used by the Training Centre. Given the content of the items, the dimensions were defined in terms of objectives/content, method and usefulness/overall rating.

In order to specify the measurement model obtained in the above stage a CFA was then carried out; polychoric rather than Pearson correlations were also used because, as Table 1 shows, the latter underestimate the relationship between the supposedly continuous variables which have been categorised. The model was specified according mainly to theoretical criteria used in the developed scale in accordance to the tri-factor EFA results. In addition to the three-factor structure a second-order structure (model 2) was investigated due to the correlations obtained between factors. This revealed that although the three-factor model (model 1) may represent the data regarding satisfaction with training adequately, in terms of coherence with the initial theoretical development of the scale a second-order factor structure (where satisfaction with training is the second-order factor) provides better evidence of construct validity, given the relationship between theory and data. In sum, model 1 and model 2 fit the data. However, even when significant, the difference between the models is small ($p=0.048$), so that the hierarchical model (2) is preferred, mainly because of theoretical reasons, as it is a basis for a total scale of training satisfaction.

On a related matter, we would like to highlight the fact that measurement in the social sciences often implies important degrees of error, both random and systematic, which may bias the estimates of the relationships between the variables measured and, therefore, produce bias in the substantive conclusions. In this regard, this paper is focused on one of the key problems relates to the consequences that use of the different types of correlations have on substantive and

methodological research, as they may increase error and contribute to the misinterpretation of evidence of construct validity, the cornerstone of basic and applied research (Cronbach, 1984, Messick, 1994, APA, AERA, NCME, 1999, Shadish, Cook & Campbell, 2002).

Finally, we want to mention some weaknesses of this study and some of the subsequent present and future researches that we are performing. We believe that further research is required into the invariance of the scale's dimensionality, this being an important issue in the field of measurement (Vandenberg & Lance, 2000). One of the tasks ahead is analysing whether the scale measures the same concept both in subjects from different administrative departments of the University of Seville and in those from other institutions. In order to investigate the adequateness of the measurement model in other samples, we are using the same scale to collect data from different organizations (Provincial Council of Seville, Regional Sport Institute in Andalusia). At the same time, in order to obtain more evidences about the validity of the scale, in relation to criterion validity we are recording new variables of interest for the Training Center (participants' outcome measures in workplaces). On the other hand, the hierarchical nature of the data should be considered in future researches. In this sense, the Training Centre tries to standardize all the editions of each training action. Then, we should take into account the hierarchical nature of the data and search second order variables that could be causing a significant variability between courses (i.e. participants could be considered to be at level 1; nested in training actions, considered as level 2 in a hierarchical structure of data).

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

Acknowledgement

The authors highly appreciate all comments received from journal reviewers. We consider that quality and content of the paper have been substantially improved because of this.

References

APA, AERA, NCME (1999). *Standards for Educational and Psychological Testing*. Washington: Author.

Bandalos, D.L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research 28,* 351-374.

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 Academic Publishers.

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

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley.

Bollen, K.A. & Long, J.S. (1993). *Testing Structural Equation Models*. CA: Sage.

Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen and J.S. Long (1993). *Testing Structural Equation Models* (pp. 445-455). CA: Sage.

Byrne, B.M. (2001). *Structural Equation Modeling with AMOS: Basic Concepts, Applications and Programing.* New Jersey: Lawrence Erlbaum Associates.

Chacón, S., Anguera, Mª T., Pérez, J.A. & Holgado, F.P. (2002). A Mutual Catalytic Model of Formative Evaluation: The Interdependent Roles of Evaluators and Local Practitioners. *Evaluation. The International Journal of Theory, Research and Practice 8(4),* 413-432

Chelimsky, E. (1997). The coming transformations in evaluation. In E. Chelimsky and W.R. Shadish (Eds.). *Evaluation for the 21st Century: A Handbook* (pp. 1-26). London: Sage Publications.

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

Cronbach, L. (1984). *Essentials of Psychological Testing*. Harper and Row: New York.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling 9 (3)*, 327-346.

Fernández, P. (2000). *Experiencia de un proceso de evaluación en el BSCH.* Paper presented at the *I Jornadas Universitarias de Formación Continua*.

Flora, B.F., 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.

Guilley, W. & Uhlig, G. (1993). Factor analysis and ordinal data. *Education 114 (2)*, 258-264.

Holgado, F.P. (2002). *Evaluación institucional: alternativas metodológicas en la delimitación y mejora de la calidad universitaria*. Unpublished doctoral dissertation: University of Seville [Institutional evaluation: methodological alternatives to measure and improve university quality].

Hopko, D.R (2003). Confirmatory factor analysis of the math anxiety rating scale-revised. *Educational and Psychological Measurement 63 (2),* 336-351.

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

Jöreskog, K. & Sörbom, D. (1996a). *LISREL 8: User´s reference guide.* Chicago: Scientific Software International.

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

Jöreskog, K. & Sörbom, D. (2003). *LISREL 8.54*. Chicago: Scientific Software International.

Kirkpatrick, D. (1999). *Evaluación de acciones formativas: Los cuatro niveles.* Barcelona: Training Club y Epise. [Evaluating Training Programs: The Four Levels].

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

Linn, R.L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues & Practice, 13*, 12-14.

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.

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

Muñiz, J. (2004, July). New Trends in Psychological Measurement. Keynote Lecture presented at the *24th Biennial Conference of the Society for Multivariate Analysis in the Behavioral Sciences* (SMABS 2004). Jena: University of Jena.

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.*

Pineda, P. (2000). *Rentabilidad de la formación.* Paper presented at the *I Jornadas Universitarias de Formación Continua*.

Salanova, M. & Grau. R. (1999). Análisis de necesidades formativas y evaluación de la formación en contextos de cambio tecnológico. *Revista de Psicología General y Aplicada*, *52*, 329-350.

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.

Shadish, W., Cook, T. & Leviton, L. (1991). *Foundations of Program Evaluation. Theories of Practice.* New York: Sage Publications.

Shelton, S. & George (1993). Who's afraid of level 4 evaluation? *Training and Development, June,* 43-46.

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

Vandenberg, R. & Lance, Ch. (2000). A review and synthesis of the measurement invariance literature; suggestions, practices, and recommendations for Organizational Research. *Organizational Research Methods, 3 (1)*, 4-69.

Ventosa, P. (1998). *Desde la evaluación de la formación al rendimiento de la inversión*. Barcelona: Episе. [From training evaluation to investment outcomes]

Willyerd, K. (1997). Balancing your evaluation act. *Training, March,* 52-58.

Appendix 1. Main bibliographical references of the reviewed questionnaires and list of the original 72 items.

Medina, M. (1996). Evaluation of the quality of assistance in social services. *Intervención Psicosocial, 14,* 23-42.
Meliá, J.L., Peiró, J.Mª. & Calatayud, C. (1986). El cuestionario general de satisfacción en organizaciones laborales: Estudios factorials, fiabilidad y validez. *Millars, 11*, 43-77.
Meliá, J.L.& Peiró, J.Mª. (1989). El cuestionario de satisfacción S10/12: estructura factorial, fiabilidad y validez. *Psicología del Trabajo y de las Organizaciones, 11*, 179-187.
Meliá, J.L.& Peiró, J.Mª. (1988). La medida de la satisfacción laboral en los contextos organizacionales: El cuestionario de satisfacción S20/23. *Psicologemas*, 59-74.
Fernández, J.A. & Ovejero, A. (1994). Satisfacción laboral en un centro hospitalario: Un análisis del cuestionario de Porter. *Psicología del Trabajo y de las Organizaciones, 10*, 39-61.

The following list of 72 items were extracted from these references and different available questionaries used in various Spanish institutions, but no published in any journal (12 items from used questionnaire are specified):

Objectives and contents objetivos y contenidos

1. The objectives show clearly what they pretend to reach (los objetivos del curso son claros en cuanto a sus pretensiones).
2. The objectives were widely published and spreaded (los objetivos han sido ampliamente publicados y difundidos).
3. The information I received about the training is suitable (la información que he recibido acerca del curso es adecuada).
4. The issues were dealt with in as much in depth as the length of the course allowed. *(Item2)*
5. The length of the course was adequate for the objectives and content. *(Item3)*
6. At the beginning of the training, the general objectives were explained (al inicio del curso se explicó a los asistentes cuáles eran los objetivos generales).
7. The objectives of the training responded to my needs and interests (los objetivos del curso han respondido a mis necesidades e intereses).
8. In general, I am satisfied with the goals and developed objectives (en general estoy satisfecho con las metas y objetivos desarrollados).
9. In my opinion the planned objectives were met. *(Item 1)*
10. The contents of the training have a practical applicability (creo que los contenidos abordados son de aplicabilidad práctica).
11. The contents are clearly specified (los contenidos están claramente explicitados).
12. The contents are rightly structured (en mi opinión los contenidos están estructurados de forma adecuada).
13. In general, I am satisfied with the treated contents in the training (en general estoy satisfecho con los contenidos tratados en la acción formativa).
14. The main concepts and ideas were clear (los conceptos e ideas fundamentales han sido claros).
15. Continuing the pace of work turned out to be easy (creo que seguir el ritmo de trabajo ha resultado fácil)
16. The level of participation of the group was high (el grado de participación del grupo ha sido alto).

Method and Training context (Metodo y contexto de formación).

17. I'm looking forward to beginning of the training (espero con interés que empiece el curso).
18. I know exactly what must I do in this training (conozco con bastante exactitud qué se debe hacer en este curso).

19. This training is boring (este curso es aburrido).
20. The training was realistic and practical (este curso ha sido realista y práctico). *(Item 7)*
21. The teacher is the only one who decides what to be done in this training (el formador es el único que decide qué se hace en este curso).
22. The teacher fostered to work in groups (el formador ha creado un espíritu de trabajo en equipo).
23. The teacher is receptive to the offers or initiatives that are formulated (el formador es receptivo a las propuestas o iniciativas que se le formulan)
24. I established good professional relationships with colleagues (he entablado buenas relaciones con los demás compañeros).
25. The training enabled me to share professional experiences with colleagues. *(Item 6)*
26. The training was developed in a collaboration context (el curso se ha desarrollado con un ambiente de cooperación).
27. The interest and motivation of the teacher were high (el interés y motivación del formador ha sido importante).
28. The teacher fomented the reflection (el formador favorece la reflexión).
29. There was a good relationship between teachers and participants (ha existido una buena relación entre profesores y participantes).
30. The teacher has the capacity of hearing and fomenting relationships between participants (el formador tiene la capacidad de saber escuchar y de propiciar relaciones interpersonales).
31. The teacher helped in situations of conflict and resolved them adequately (el formador medió en situaciones de conflicto y las resolvió de forma adecuada).
32. This training started on time rarely (esta acción formativa rara vez ha empezado a su hora).
33. The participants are not satisfied with proponed training activities (los alumnos no están contentos con lo que se hace en esta acción formativa).
34. The training context was well suited to the training process (en general estoy satisfecho con el ambiente en que se ha desarrollado esta acción formativa).
35. The timetable was suitable (el horario del curso ha sido adecuado).
36. The training enviroment is well suited to the training process (el espacio docente reúne las condiciones adecuadas).
37. The training context is well suited to the training process (el espacio docente reúne las condiciones adecuadas). *(Item 9)*
38. Documentation has a clear scientific and comprehensive language (los materiales escritos entregados tienen un lenguaje científico claro y comprensible).
39. The documentation given out is current and relevant for us (los materiales escritos entregados son actuales y relevantes para nosotros).
40. The proposed exercises are practical and pertinent for the objectives that the training tried to reach (los ejercicios propuestos son prácticos y pertinentes en cuanto al entrenamiento de las destrezas que pretenden mejorar).
41. The documentation given out have a presentation of good quality about format, size, etc. (los materiales escritos tienen un buen aspecto y calidad en cuanto a su formato, presentación, tamaño, etc.)
42. The documentation given out was of good quality. *(Item 8)*
43. The documentation given out is systematically arranged. They present a logical schedule. (los materiales escritos están ordenados de forma sistemática. Siguen un orden lógico).
44. The language of the documentation given is clear: the technician vocabulary is gradually introduced; the difficult concepts are clarified with examples. (el lenguaje empleado en los materiales escritos es comprensible: el vocabulario técnico se introduce gradualmente, los conceptos difíciles se clarifican mediante ejemplos).
45. The documentation given out was valuable and useful (el material documental utilizado ha resultado valioso y útil).

46. In general, I am satisfied with the documentation of this training (en general estoy satisfecho con los materiales de este curso).
47. My problems/hesitations were efficaciously resolved (se me han resuelto las dudas de manera eficaz).
48. The training is adequately organized (la acción formativa está adecuadamente organizada.
49. The ideas and concepts were adequately developed (las ideas y conceptos han sido desarrollados adecuadamente).
50. I like the training method (me gusta el método de enseñanza).
51. The method was well suited to the objectives and content *(Item 4)*
52. The method used enabled us to take an active part in training *(Item 5)*
53. The teacher knows the subject (el formador conoce la material).
54. The teacher does not know how to explain the contents to the participants (el professor no sabe explicar la materia a los alumnus).
55. The instructions to resolve the proposed exercises are clear and facilitate the work (las instrucciones para resolver las tareas propuestas son claras y facilitan el trabajo).
56. In general, I am satisfied with the work developed by the teacher (en general estoy satisfecho con el trabajo que desarrolla el formador).
57. The reasons for which I took part in the training were satisfied (los motivos que me llevaron a participar en la acción formativa han sido satisfechos).
58. The participants' opinion to modify aspects referred to the training was considered (la opinión de los asistentes para modificar aspectos relacionados con la acción formativa se han tenido en cuenta).
59. The acoustic conditions of the classroom are suitable (las condiciones acústicas del aula son adecuadas).
60. The luminous conditions of the classroom are suitable (las condiciones lumínicas del aula son adecuadas).
61. The used logistic resources (slideprojector, videos, computers, etc.) are suitable (los recursos técnicos empleados (retropoyectores, videos, ordenadores,...) son adecuados.
62. The cleanliness, hygiene and healthiness of the training center are suitable (la limpieza, higiene y salubridad del centro de formación son adecuadas).
63. The furniture could be considered suitable and comfortable enough (el mobiliario se puede considerar aceptable y posee la suficiente comodidad).
64. In general, I am satisfied with the habitability conditions of the training class (en general estoy satisfecho con las condiciones de habitabilidad del espacio docente).

Usefulness and overall rating.

65. The training received is useful for my specific job (este curso fue útil para mi puesto de trabajo).*(item10)*
66. The received training is relevant for my workplace (este curso es relevante para para mi puesto de trabajo
67. The training received is useful for my personal development. *(Item11)*
68. The training merits a good overall rating. *(item 12)*
69. The practical exercises were a nearby reflex of the work reality (los ejercicios prácticos realizados han sido un reflejo cercano de la realidad laboral).
70. In general, the training gave me a very good impression and I was satisfied (en general la acción formativa me ha causado muy buena impresión y me ha dejado satisfecho).
71. This training was good organized (este curso ha estado bien organizado).
72. The training satisfied my previous expectation (el curso ha satisfecho mis expectativas previas).

Table 1. Matrix of polychoric (below diagonal) and Pearson (above diagonal) correlations

	Obj1	**Obj2**	**Obj3**	**Met4**	**Met5**	**Met6**	**Met7**	**Met8**	**Met9**	**Use10**	**Use11**	**Use12**
Obj1	---	.616	.524	.624	.451	.326	.590	.406	.023	.449	.442	.613
Obj2	.674	---	.485	.563	.418	.246	.476	.365	.180	.369	.443	.566
Obj3	.598	.546	---	.483	.348	.289	.380	.277	.153	.279	.274	.400
Met4	.700	.633	.522	---	.553	.320	.598	.442	.292	.398	.456	.614
Met5	.547	.519	.415	.668	---	.434	.528	.287	.246	.307	.374	.484
Met6	.367	.283	.342	.391	.539	---	.433	.199	.124	.305	.357	.378
Met7	.645	.545	.447	.684	.662	.515	---	.485	.283	.477	.502	.631
Met8	.489	.471	.326	.550	.430	.301	.545	---	.366	.312	.347	.467
Met9	.292	.245	.174	.366	.350	.181	.357	.437	---	.233	.250	.311
Use10	.508	.408	.337	.492	.418	.364	.530	.410	.341	---	.565	.542
Use11	.523	.471	.323	.534	.474	.418	.585	.442	.367	.687	---	.619
Use12	.691	.626	.476	.708	.601	.462	.714	.599	.385	.611	.703	---

Table 2. Loadings pattern and factor correlations (*cursive*)

Item	*F 1*	*F 2*	*F 3*
Objectives and content (F1)		*.623*	*.623*
1. In my opinion the planned objectives were met (OBJ1)	**.751**	.016	.142
2. The issues were dealt with in as much in depth as the length of the course allowed (OBJ2)	**.724**	-.029	.125
3. The length of the course was adequate for the objectives and content (OBJ3)	**.691**	.062	-.088
Method and training context (F2)			*.681*
4. The method was well suited to the objectives and content (MET4)	**.555**	**.345**	.043
5. The method used enabled us to take an active part in training (MET5)	.142	**.842**	-.123
6. The training enabled me to share professional experiences with colleagues (MET6)	-.077	**.535**	.181
7. The training was realistic and practical (MET7)	.266	**.347**	**.324**
8. The documentation given out was of good quality (MET8)	.289	.099	**.322**
9. The training context was well suited to the training process (MET9)	-.019	.194	**.315**
Usefulness and overall rating (F3)			
10. The training received is useful for my specific job (USE10)	.042	-.058	**.793**
11. The training received is useful for my personal development (USE10)	.040	-.036	**.855**
12. The training merits a good overall rating (USE12)	**.385**	.088	**.524**

Table 3. Fit indices of models 1 and 2 *(p = .048)

Model	*RMSEA*	*GFI*	*AGFI*	*ECVI*	*CAIC*	χ^2	*d.f.*	$\Delta\chi^2$	Δ *d.f.*
1	.071	.98	.97	.34	577.64	358.78	51	---	---
2	.069	.98	.97	.34	567.49	364.83	53	6.05*	2

Figure 1. Model 2 (completely standardized solution)