Title: Measurement Invariance of the Day Reconstruction Method: Results from the COURAGE in Europe Project.

Authors:

Blanca Mellor-Marsá, MS [a,b,c]; Marta Miret, PhD [b,a,c]; Francisco J. Abad, PhD [d]; Somnath Chatterji, MD [e]; Beatriz Olaya, PhD [f,b]; Beata Tobiasz-Adamczyk, PhD [g]; Seppo Koskinen, MD, PhD [h]; Matilde Leonardi, MD [i]; Josep Maria Haro, MD, PhD [f,b]; José Luis Ayuso-Mateos, MD, PhD [a,c,b], Francisco Félix Caballero, PhD [b,a,c] *

Affiliations:

a. Departamento de Psiquiatría, Universidad Autónoma de Madrid, Madrid, Spain.

b. Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental. CIBERSAM, Spain.

c. Instituto de Investigación de la Princesa (IIS-IP), Hospital Universitario de la Princesa, Madrid, Spain.

d. Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, Madrid, Spain.

e. Department of Health Statistics and Information Systems, World Health Organization, Geneva, Switzerland.

f. Parc Sanitari Sant Joan de Déu, Universitat de Barcelona, Spain.

g. Department of Medical Sociology, Jagiellonian University Medical College, Krakow, Poland.

h. National Institute for Health and Welfare, Helsinki, Finland.

i. Fondazione IRCCS, Neurological Institute Carlo Besta, Milano, Italy.

*Corresponding author: Francisco Félix Caballero, Departamento de Psiquiatría, Universidad Autónoma de Madrid. C/ Arzobispo Morcillo 4, 28029 Madrid, Spain. Telephone: (0034) 91 497 27 16; Fax: (0034) 91 497 43 89; e-mail: felix.caballero@uam.es

Disclosure of potential conflicts of interest: The authors declare that they have no conflict of interest. The views expressed in this paper are those of the authors, and do not necessarily represent the views or policies of the World Health Organization.

Funding: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 223071 (COURAGE in Europe), from the Instituto de Salud Carlos III-FIS research grants number PS09/00295 and PS09/01845, and from the Spanish Ministry of Science and Innovation ACI-Promociona (ACI2009-1010). The study was supported by the Instituto de Salud Carlos III, Centro de Investigación Biomédica Red de Salud Mental (CIBERSAM), and the AGES-CM Programme (AGES-S2010/BMD-2422), B.O. is grateful to the Sara Borrell postdoctoral programme (reference no. CD12/00429) supported by the Instituto de Salud Carlos III, Spain.

Introduction

To improve policies in the current global situation, it is necessary to produce data that reflect the situation of populations from a broad perspective; therefore, adding data on populations' health, disability, quality of life and well-being to economic information is becoming a growing need for more focused policy development strategies. In this sense, there is also a growing need for the development and validation of tools that can measure time use and emotional experience associated with day-to-day activities, based on self-reports that can be easily interpreted and which are, as much as possible, free from prejudice and the bias effects of memory and positivity (Hox et al. 2010).

The concept of subjective well-being includes an individual's satisfaction with various domains of life, their overall judgment of life satisfaction, and their current affective state, measured as a time-weighted metric of the amount of negative or positive emotions (Diener et al. 1999). By examining real-time experience, emotions can be measured and registered without assessment or retrieval biases, making it possible to produce a scientific measure of subjective well-being that is rooted in the present (Kahneman et al. 1999).

In order to capture this concept, Kahneman et al. (2004) developed the Day Reconstruction Method (DRM). Its scores have shown: 1) adequate test-retest correlations in the range of 0.50-0.70, which are sufficiently high for the studies conducted in the area of subjective well-being (Krueger and Schkade 2008); 2) reliable estimates, with between-person correlations ranging from 0.58 to 0.90, of the intensity of affect as compared with the ones generated by the gold standard tool in this area, the Experience Sampling Method, also known as Ecological Momentary Assessment (Dockray et al. 2010); 3) high internal consistency, with multilevel reliability estimates higher than 0.90 for negative affect and

positive affect (Bylsma et al. 2011); 4) the absence of memory and judgmental biases (Diener et al. 2010).

However, the DRM is a time-consuming tool (Diener et al. 2010; Miret et al. 2012)—taking from 45 to 75 minutes to be completed (Kahneman et al. 2004)—which is also expensive, like other instruments of its class, such as the Princeton Affect and Time Survey (Krueger and Stone 2008).

For this reason, an hetero-administered abbreviated version of the DRM was created within the World Health Organization Study on Global Ageing and Adult Health (SAGE) (Kowal 2012) comprising different forms (A, B, C) in order to assess the morning, afternoon, or evening, respectively, of the previous day. It has been employed in several countries (including India, Russia, Ghana, China, South Africa and Mexico) and it has shown adequate psychometric properties regarding reliability (with composite reliability coefficients ranging from 0.77 to 0.91 for negative affect, and from 0.70 to 0.89 for positive affect) and the goodness-of-fit indices for the two-correlated-factors model proposed (CFI and TLI values higher than 0.98. RMSEA values from 0.026 to 0.074). Most of the countries showed a similar diurnal variation of affect, with a tendency to improve throughout the day (Ayuso-Mateos et al. 2013). This brief version has the advantage that it can be used in large studies with representative samples of the population worldwide: evaluating the affective state using the abbreviated version of the DRM provides a profile of the population similar to that obtained with the full version, while the different scores in the measures which can be obtained from the DRM have shown moderate test-retest reliability, with Intraclass Correlation Coefficient (ICC) values ranging from 0.60 to 0.80 and indicating a good test-retest reliability according to the standards (Landis and Koch 1977). Positive affect showed

higher test-retest reliability [ICC = 0.49; 95% CI = (0.46, 0.52)] than negative affect [ICC = 0.28; 95% CI = (0.24, 0.33)]; and the affect scores (positive and negative affect) of sets A, B, and C, taken together, had a moderate predictive ability (measured by means of Area Under the ROC Curves; AUC = 0.67 for positive affect and AUC = 0.61 for negative affect) considering the affect scores obtained using the full version of the DRM as gold standard (Miret et al. 2012). Other previous work analyzed the psychometric evidence of the abbreviated Spanish version of the DRM, applied in Mexico and Spain (Caballero et al. 2014). In that study, reliability and validity of the scores were found to be adequate, and the two samples showed opposite patterns in the diurnal variation of affect. Nevertheless, no time frame comparisons were done to study the influence of this dimension on the affect measures, and all the analyses were conducted separately in both countries, since different rating scales were employed in the samples.

In spite of all these studies conducted using the abbreviated version of the DRM, there has hitherto been no work on possible differential test performances across countries, different application time frames, or days of the week. Moreover, dimensionality studies on the brief version of the DRM have, to date, focused on factorial structure and reliability (Ayuso-Mateos et al. 2013; Miret et al. 2012); while positive affect, negative affect or net affect scores (Kahneman et al. 2004), have been used to report differences between samples, but a deeper assessment of measurement invariance among the study groups is missing.

In order to show that a test produces measures of the same construct in different groups, the study of bias and differential test functioning is indispensable, so that the outcome scores can be compared with the same measurement scale (Deaton 2008). In substantive terms, by applying a test to different samples in the absence of bias, it can be said

that there is equity between the scores assigned to groups. In this sense, the measurement invariance of the day reconstruction process has to be shown in terms of critical dimensions: firstly, the cross-country invariance of the measures; secondly, given the nature of the different application forms A, B, and C (assessing the previous morning, afternoon, or evening) in this abbreviated version of the instrument, the invariance across different time frames; finally, the possible influence of the "weekend effect" (Ryan et al. 2010) on the scores (a pattern found in the literature, whereby mood is more positive and less negative on weekends than during the rest of the week), by checking whether measurement invariance holds when the scores of the abbreviated version of the DRM refer to weekdays vs. weekends.

Considering the above-mentioned evidence available in the literature on DRM studies, a number of unsolved questions have arisen. The knowledge gaps that this research paper meant to address were: 1) whether the abbreviated version of the DRM produces valid and reliable scores for the comparison of well-being measures across countries, and other application design categories, as the different assessed parts of the day or the distinct days of the week; 2) whether there are differences in positive and negative affect among countries, across the different DRM application sets, and between weekdays and weekends, once the measurement invariance has been tested.

The main objectives of the present study were to assess the measurement invariance of the brief version of the DRM across nationally representative samples from three European countries (Finland, Poland, and Spain) within the Collaborative Research on Ageing in Europe (COURAGE in Europe) project (Leonardi et al. 2013), to test measurement invariance across several characteristics of the questionnaire (time frames and

day of the week) and to examine differences across samples in the scores obtained from the abbreviated version of the DRM. These objectives were addressed through the following strategies: 1) a dimensionality study of the brief version of the DRM considering a two-correlated-factors model in each country; 2) an assessment of the reliability at the factor level and the reliability of the net affect (a composite score); 3) a sequential invariance analysis across different groups (country, time frame and day of the week) by means of multiple group confirmatory factor analysis; and 4) a comparison of latent scores across countries, time frames and days of the week.

In light of these scientific uncertainties and given the –elsewhere exposed (Ayuso-Mateos et al. 2013; Miret et al. 2012; Caballero et al. 2014) – evidence on the DRM psychometric properties, the hypotheses that guided this work were: 1) it is expected that scores produced by the abbreviated version of the DRM allow to confirm there is measurement invariance across three European countries under study: Poland, Finland and Spain; 2) differences in positive and negative affect scores will be found across samples, reflecting real differences among the three compared countries; 3) the affect structure measured by the abbreviated DRM questionnaire will remain invariant in each country when comparing different time frames, and when considering different days of the week.

Methods

Design

The COURAGE in Europe project (http://www.courageineurope.eu/) is a cross-sectional survey that interviewed nationally representative samples of the general non-institutionalized adult population from three European countries: Finland, Poland, and Spain.

These countries were selected to give a broad representation spanning different European regions, taking into consideration their population and health characteristics.

Sample and procedure

Interviews were carried out face-to-face at respondents' homes, with Computer-Assisted Personal Interviewing (CAPI), and were conducted between April 8, 2011 and May 8, 2012. For the administration of the COURAGE in Europe survey, all the interviewers participated in a training course. The questions of the interview were translated from English into Finnish, Polish, and Spanish, following the World Health Organization's translation guidelines for assessment instruments: a forward translation, a targeted back-translation, a review by a bilingual expert group, and a detailed translation report (The World Health Organization 2013).Quality assurance procedures were implemented during fieldwork (Üstün et al. 2005).

In Poland and Spain, a stratified multistage random sampling method was used, and strata were created according to the geographical administrative regions and number of people living in the habitat. Age strata were used to select households according to the age structure of the population. The respondents were randomly selected among inhabitants of a household from a certain age group. In Finland, the design was a stratified two-stage cluster sampling design, and strata were created based on the largest towns and university hospital regions. A systematic sampling of people was conducted so that the sample size in each stratum was proportional to the corresponding population base. A probability proportion-to-size design was used to select clusters, and people 50+ years old were overrepresented. A total of 10800 individuals were interviewed: 1976 from Finland, 4071 from Poland, and

4753 from Spain. The countries' response rates were calculated as the total number of interviews divided by the number of persons eligible for inclusion (a selected household that turns out to be a vacant dwelling, for example, is not eligible). The resulting response rates were 53.4% for Finland, 66.5% for Poland, and 69.9% for Spain. If a participant was cognitively impaired and unable to respond to the interview, a proxy was asked for some questions about the participant's health. For the purposes of the present analyses, these participants were not included.

Ethical approvals were obtained from the Bioethical Committee, Jagiellonian University, Krakow, Poland; Ethics Review Committee, Parc Sanitari Sant Joan de Déu, Barcelona, Spain; Ethics Review Committee, La Princesa University Hospital, Madrid, Spain; and the Ethics Review Committee, National Public Health Institute, Helsinki, Finland. Informed consent from each participant was also obtained.

Measures

The brief version of the DRM (Kahneman et al. 2004) (http://www.who.int/healthinfo/systems/sage/en/index.html) was used to gather information about participants' daily activities and their subjective well-being. This version comprised three sets (A, B, and C), which only differed on the time frame in which they were applied, and was limited to a maximum of 15 minutes of interview. Respondents were asked to reconstruct a portion (morning, afternoon, or evening) of their previous day's activities, depending on the set (A, B, and C, respectively) to which they were randomly assigned. Participants responded to questions about each activity on a 7-point response scale (0 = not at all, 6 = very much, with the remaining points unlabeled), including the nature of the

activity (e.g. working, doing housework, reading) and the extent to which they experienced various feelings during the time that the activity lasted: *calm/relaxed, enjoying, worried, rushed, irritated/angry, depressed,* and *tense/stressed*. An example of the DRM questions' structure is shown in Table 1.

In each of the aforesaid seven items, aggregated scores were created by computing a duration-weighted average for each respondent. For each participant, the scores were calculated based on the evaluation of the seven emotions for each activity separately. The participants reported between one and ten activities. The duration of each activity -reported in hours and minutes- was divided by the total length of all the activities described by the subject. The following formula was employed in order to calculate the aggregated score in each item:

$$X_i = \sum_{t=1}^{k} w_t * X_I\,(t)$$

where X_i is the aggregated score in each item, with *i* ranging between 1 and 7; t is the activity, ranging from 1 to *K*, the total number of activities reported by the subject; w_t is the proportion of time employed in that activity; and $X_i\ (t)$ is the affect score in the item X_i associated to the activity *t*. The items *calm/relaxed* and *enjoying* were noted as X_1 and X_2, respectively; the items *worried, rushed, irritated/angry, depressed* and *tense/stressed* were noted as X_3, X_4, X_5, X_6 and X_7, respectively.

The construct comprising the five negative items (*worried, rushed, irritated/angry, depressed,* and *tense/stressed*) is called negative affect, while the construct comprising the positive ones (*calm/relaxed* and *enjoying*) is called positive affect. Kahneman et al. (2004)

defined positive affect as the average of positive feelings collected in the DRM, and negative affect as the average of the negative ones. A global measure combining positive and negative affect has also been defined in the literature: the net affect, which is defined as the positive affect minus the negative affect (Kahneman and Krueger 2006). The aggregated scores on each item were calculated for each subject and then the scores on positive and negative affect were estimated.

Respondents were also asked to provide socio-demographic information: age, sex, highest level of education achieved (low, corresponding to up to primary school; middle, comprising secondary school and high school; and high, for college/pre-university/university and post-graduate degree), marital status (never married, currently married/cohabiting, separated/divorced, widowed), and residential setting (rural, urban).

Data analysis

Participants who did not complete the interview and those who did not respond to one or more of the DRM's affect questions, were excluded from the analysis. Frequency analysis and descriptive statistics were used to analyze the demographic characteristics of the samples corresponding to each country, after excluding missing values. Differences in proportions and scores were analyzed using chi-square and ANOVA/unpaired *t* tests.

In each country, sets A, B, and C were pooled, and the factorial structure of the seven adjectives associated with the different activities coded on the DRM was tested through Confirmatory Factor Analysis (CFA). Aggregated scores were created by computing a duration-weighted average for each respondent, and CFA was conducted over these aggregated scores on the seven items of the DRM. A factorial structure comprising two

correlated factors was hypothesized based on previous findings (Ayuso-Mateos et al. 2013; Caballero et al. 2014), with the items *calm/relaxed* and *enjoying* as part of the positive affect factor, and the items *worried*, *rushed*, *irritated/angry*, *depressed*, and *tense/stressed* as part of the negative affect factor. Standardized loadings associated with both factors were obtained.

MLM estimation, based on maximum likelihood parameter estimates with standard errors and a mean-adjusted chi-square test statistic robust to non-normality, was employed. Robust chi-squares and standard errors are generally more accurate than the asymptotic tests when data are non-normal and when the model is miss-specified, and they are assumed to offer some protection against unmodeled heterogeneity (Hox et al. 2010).

The goodness-of-fit of the two-correlated-factors model was assessed according to the standard recommendations (Hu and Bentler 1999; Reise et al. 1993): values of the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) above 0.90 were considered to represent an adequate fit; values of Root Mean Square Error of Approximation (RMSEA) lower than 0.08 indicated a good fit (Steiger 2007). Lack of significance of chi-square is one of the most famous criteria proposed in the literature to assess goodness-of-fit; however, since the chi-square statistic is sensitive to sample size (Schreiber et al. 2006), the chi-square values might be inflated, and statistically significant, when the sample size is large, which might erroneously imply a poor data-to-model fit (Schumacker and Lomax 2004). For this reason, the chi-square statistic was reported but not considered a definitive indicator for decisions made in the present work.

Reliability of the global score on the DRM (net affect, defined as positive affect minus negative affect) in each country was assessed by means of the reliability of a

composite score (Wainer and Thissen 2001). This coefficient provides an estimation of the correlation that would be obtained between these scores and those obtained on a parallel form. Reliability for the positive affect and the negative affect in each country was assessed by means of the omega coefficient, which provides the proportion of scale variance associated with each of the common factors in the model (McDonald 1999). Omega values can range from 0 (no reliability) to 1 (perfect reliability) and ideally and ideally values between 0.70 and 0.90 are expected when the items measure the same latent construct (Campo-Arias and Oviedo 2008).

Measurement invariance across countries was tested through sequential constraint imposition over the models' parameters in order to assess whether the same constructs have been measured across the Finnish, Polish, and Spanish samples—and, therefore, whether these groups are comparable. A multiple group CFA comprising two correlated factors was used, considering the country as the variable defining the group. A similar multiple group analysis was conducted, by country, based on different characteristics: time frame (morning, afternoon or evening, according to the set assigned) and day of the week (weekday or weekend, according to the day reconstructed). In all these cases, the equivalence of the factor structure itself (configural invariance), followed by the equivalence of the factor loadings (metric invariance), intercepts (strong invariance), and residual variances (strict invariance), were tested. A complete description of each of these types of measurement invariance can be found elsewhere (Gregorich 2006; van de Schoot 2012). When full measurement invariance is not achieved, an intermediate point between full measurement invariance and complete lack of invariance can be achieved: the partial measurement invariance. Byrne et al. (1989) introduced this concept, where only a subset of parameters in a model is constrained to be

invariant while another subset of parameters is allowed to vary across groups. Hence, partial measurement invariance may allow appropriate cross-group comparisons even if full measurement invariance is not obtained.

If the model fit corresponding to metric/strong/strict measurement invariance is significantly worse than the fit of the previous tested model, then the partial invariance can be established at the corresponding level, freeing some of the fixed parameters in each of the groups (factor loadings, intercepts, and residual variances, respectively) to test the partial metric/strong/strict measurement invariance progressively. In the present study, the evidence for parameters' invariance between the successive nested models was based on a change in CFI lower than 0.01 (Cheung and Rensvold 2002).

Several Wald tests were used to compare the latent scores on positive affect and negative affect across the different samples considered when the level of strong invariance was achieved; the latent mean and Standard Deviation (SD) in positive and negative affect of one of the groups were set to zero and one respectively, considering this group as the reference category. A correction for multiple comparisons was made; Cohen's *d* was used to estimate the effect size associated to the differences, since the significance could be due to large samples sizes.

Confidence Intervals (CI) were constructed at the 95% confidence level. When significant differences were found, an effect size measure was reported: Cramer's *V* for chi-square test, Cohen's *d* for pairwise comparisons, and Cohen's *f* for overall differences found in ANOVA. Cohen's guidelines were used as a reference (Cohen 1988): Cohen's *f* values of 0.10, 0.25, and 0.40 represent small, medium, and large effect sizes; Cohen's *d* values of 0.20, 0.50, and 0.80 constitute small, medium, and large effect sizes, respectively; for

Cramer's *V* these values would be 0.10, 0.30, and 0.50, respectively. Mplus version 6 (Muthén and Muthén 2010) was employed for structural equation modeling and reliability analysis. The remaining analyses were performed using Stata SE version 11 (StataCorp 2010).

Results

Approximately 4% of the 10800 interviewed subjects were excluded from the analysis since they did not complete the interview or they did not respond to one or more items of the DRM, and the final sample comprised 10350 subjects. The respondents excluded from the survey did not differ from the final sample regarding any major demographic characteristics (the differences were not significant, or had associated a small effect size), except age. No significant differences were found in residential setting (29.8% people living in rural areas in the excluded sample vs. 26.4% in the final sample; χ^2 (1) = 2.50; p = 0.114). Significant differences were found in age (75.8 ± 15.8 vs. 58.4 ± 16.8; t (493.71) = 22.63; p<0.001; Cohen's d = 1.01); sex (63.1% women vs. 57.1%; χ^2 (1) = 6.45; p = 0.011; Cramer's V = 0.02); level of education (67.3% low, 23.4% middle, 9.3% high vs. 36.1% low, 44.0% middle, and 19.9% high; χ^2 (2) = 163.75; p<0.001; Cramer's V = 0.12); and marital status (10.1% never married, 43.5% currently married/cohabiting, 4.4% separated/divorced, 42.0% widowed vs. 15.4% never married, 59.3% currently married/cohabiting, 8.0% separated/divorced, and 17.3% widowed: χ^2 (3) = 163.63; p<0.001; Cramer's V= 0.12).

Table 2 presents the main socio-demographic characteristics of the 10350 respondents by country. Significant differences were found across countries, with effect sizes

ranging from small to moderate. Higher differences were found for residential setting, with Poland presenting the highest proportion of people living in rural areas. Moderate differences were also found among the three countries with respect to level of education.

The two-correlated-factors model proposed showed a good fit in each country, according to the goodness-of-fit statistics (Table 3). RMSEA values ranged from 0.048 to 0.062 across countries, while values for CFI and TLI were higher than 0.90 in all cases. The factor loadings ranged from 0.71 to 0.98 on the positive affect factor, and from 0.53 to 0.91 on the negative affect factor, indicating that the two factors were well defined. The total proportion of variance explained by the two-correlated-factors model in each country was 55% for Finland, 73% for Poland and 76% for Spain. Regarding the scores obtained on the DRM, adequate omega coefficient values were found for positive affect (0.83 in Finland, 0.89 in Poland, and 0.88 in Spain) and negative affect (0.82, 0.92, and 0.94, respectively for the three countries). The values found for the reliability of the composite score (i.e., the net affect, the global score on the DRM) were also high: 0.88 for the Finnish sample, 0.92 for the Polish sample, and 0.93 for the Spanish sample.

A multiple group CFA was conducted to study invariance across countries. The goodness-of-fit indices of the models considered to test measurement invariance are shown in Table 4. All the models presented RMSEA values lower than 0.080, and CFI and TLI values higher than 0.90. In terms of ΔCFI, the model proposed to check for metric invariance showed a decrease of 0.02 in the CFI value with regard to the model used to check for configural invariance. Partial metric invariance was confirmed after freeing the loading of the item *depressed* on the negative affect factor. The strong invariance model showed a good fit (RMSEA = 0.059, CFI = 0.955, TLI = 0.949), but the decrease in the CFI value for the

subsequent model (strict invariance) was 0.037, and therefore higher than the cut-off point of 0.01. Finally, the partial strict invariance was concluded only when the residual variances associated with the items *enjoying*, *rushed*, and *irritated/angry* were allowed to be free across groups. In this last model, the goodness-of-fit indices were: RMSEA = 0.059, CFI = 0.948 and TLI = 0.950. The standardized loadings corresponding to the configural invariance model and the partial strict invariance model in each country are shown in Fig. 1, where it can be observed that in both cases the loadings of item *depressed* on the negative factor were smaller in the Finnish group while they showed similar values in the other two samples.

Invariance analysis based on time frame and day of the week were conducted in each country and the corresponding model fit indices are shown in Table 5. The level of strict invariance was achieved in both dimensions in Poland and Spain; while in the Finnish sample, the scores reached a level of strong invariance between groups only when the intercept parameter associated with item *rushed* was allowed to be free in both cases: across time frame and across day of the week. Residual variances associated to the items *worried, rushed, and irritated/angry,* respectively, were also set free in the Finnish sample to reach partial strict invariance for the time frame comparison analyses. Finally, in the same sample, the partial strict invariance by day of the week was achieved when item *rushed* was allowed to be free.

Significant differences in latent scores (positive affect and negative affect) were found across countries (Table 6). On the positive affect, the Finnish sample presented the highest score (followed by Spain and, lastly, Poland), while the Spanish sample showed the highest score on the negative affect (followed by Poland and, lastly, Finland). Furthermore,

all the pairwise comparisons were found to be significant (p<0.001), although the effect sizes were moderate or low.

The mean latent scores based on the remaining characteristics are shown in Table 7. Some significant differences in positive and negative affect were found across time frame and day of the week, although the effect sizes associated were low or moderate. In the three countries, the positive affect was higher in set C (corresponding to activities conducted in the evening) than in set A (corresponding to activities conducted in the morning), with effect sizes ranging between 0.11 and 0.23. Negative affect was also lower in set C for the Finnish sample. Regarding the day of the week, positive affect and negative affect were higher and lower, respectively, in weekend; the highest differences in affect between weekend and weekdays were found in Finland (d = 0.32 for positive affect and d = 0.28 for negative affect). In Poland, significant differences were also found, although with a low effect size associated (d = 0.16 for positive affect and d = 0.09 for negative affect).

Discussion

As part of the validation process of this tool for the assessment of subjective well-being, the main objective of the present study was to test the measurement invariance of the abbreviated version of the DRM across different dimensions: the country of the sample; the application time frame; and the day of the week to which the affect scores referred.

As to the essential scientific requirement of the possibility of comparing measurements obtained in different contexts (Dorans and Schmitt 1991), little has been said about measurement invariance in multiple item affect measures. One exception can be found in Devins et al. (1997), where the measurement invariance of the Affect Balance Scale

(Bradburn 1969) was assessed among different languages. However, these scale's indicators have no linkage to specific episodes or activities and therefore, experienced well-being is poorly conceptualized. Some studies have validated well-being multiple item measures in different countries (Oishi 2010; Vázquez and Hervás 2013), but no invariance analyses were performed and the reliability and validity of the scores were assessed separately for each country.

In the present study, adequate goodness-of-fit indices were found for a two-correlated-factors model (positive affect and negative affect) in all three countries: Finland, Poland, and Spain. Omega coefficient values ranged from 0.82 to 0.94 in all factors analysed, indicating that at least 82% of the variance in the scale observed scores (positive affect, negative affect and net affect scores) is attributable to the positive affect, negative affect and net affect target constructs respectively; therefore it may be said that the scores obtained by the DRM show satisfactory measurement precision since, in all cases, omega values were above 0.80, so above 0.70 as recommended by Campo-Arias & Oviedo (2008) and were even higher when analyzing the Polish and Spanish sample with regard to negative and net affect. These outputs are in line with previous studies conducted in several countries (Ayuso-Mateos et al. 2013; Caballero et al. 2014; Miret et al. 2012).

For the three dimensions analyzed (country, time frame and day of the week), measurement invariance was tested through a multiple group CFA framework. According to the cross-country invariance results, it can be said that in Finland, Poland, and Spain the same constructs are measured by the scores obtained with the application of the brief version of the DRM. Moreover, the results obtained provide evidence for metric invariance in all the countries, which evidences the absence of non-uniform Differential Item Functioning (DIF)

for every group studied on all but one item -*depressed*- which did not remain invariant in the Finnish sample. This effect can be assumed to be small since only one item was problematic in one sole country and its loadings on the negative affect factor were smaller in the Finnish sample. Finally, four of the seven item scores showed strict invariance in the cross-country analysis. This requirement makes it possible to state that the differences in the item variances that are not explained by the common factors could imply differences in measurement accuracy between the individuals from the three national samples. According to the results obtained in the cross-country invariance analysis, positive and negative emotions were measured with the same precision on the items *calm/relaxed*, *worried*, *depressed* and *tense/stressed* in the three countries.

Beyond the cross-country comparison, the DRM measures have shown to hold across different time frames and days of the week. In this sense, the present study contributes new relevant information regarding whether the items assessed achieve measurement invariance in a range of different substantive measure units and providing evidence about the comparability not only of the country samples, but also of the scores of the subjects reporting their affect states during different time frames and days of the week. Evidences for metric invariance were found in all the samples. Poland and Spain showed a strict level of invariance. However in the Finnish sample, one intercept—corresponding to the item *rushed*—did not achieve the level of strong invariance. As before, this effect can be assumed to be small considering that only one item is involved and only in one sample. As a result, latent means can be compared for all items but this one for the Finnish sample. At the strict level, four of the items' residual variances—*calm/relaxed, enjoying, depressed, tense/stressed*—remained invariant. This implies that positive and negative affect were

measured with the same precision in these items. Accordingly, these relevant findings provide additional value to this abbreviated version of the DRM, since the scores have shown to hold along different time frame and day of the week conditions. Also, as it has been shown, these dimensions should be considered in the application design of any study when applying this instrument to measure affect and subjective well-being.

Additionally, since the strong invariance hypothesis was supported by the results obtained in the measurement invariance analysis, the mean scores on the latent factors can be compared. It can be assumed that the variability found across groups is due to real differences in positive affect and negative affect, in the considered countries and conditions. This property suggests an absence of a uniform DIF for all items. After comparing mean latent scores, cross-country differences were found in positive affect and negative affect, with the Finnish sample showing the highest positive affect and the lowest negative affect; these results are consistent with related findings (Miret et al. 2014).

Previous studies using the DRM found significant differences in net affect (positive affect minus negative affect) across seven country samples, with moderately high effect size associated to the differences found. The highest mean net affect score was found in South Africa and the largest difference in net affect were found between South Africa and India (Ayuso-Mateos et al. 2013). In other study, comparing the scores in affect between a Mexican and a Spanish sample, net affect and positive affect were higher in the later one than in the former. Income and gender as well as age were found to be predictors of positive affect (Caballero et al. 2014). In the present study, cross-country differences were found in positive affect and negative affect. Finland's higher per-capita gross domestic product (GDP), followed by Spain and Poland (The World Bank 2011), could be an important

factor explaining the highest positive affect of the Finnish sample and the lowest negative affect as life satisfaction has been found to be higher in countries with higher per-capita GDP (Diener et al. 2010; Deaton 2008). Furthermore, the Gini coefficient (35.8 in Spain, 33.2 in Poland and 27.8 in Finland) (The World Bank 2010), that accounts for inequalities in the societies, has shown to correlate with factors that might lead to lower satisfaction and lower well-being (Wilkinson and Pickett 2007, 2009).

These results might lead to interpret the values of positive and negative affect attending to the socio-economic status of the countries for which representative samples were compared. In this sense, countries with stronger economies and societies with less social inequalities show higher levels of subjective well-being. Future studies could focus on the influences that socio-demographic variables have on the DRM scores, especially in the case of level of education or residential setting, for which moderate significant differences were found among samples here. The outcomes found in this work are consistent with the ones exposed by other authors (Miret et al. 2014), which considered positive and negative affect as average scores instead of latent scores.

Analyses regarding time frames and days of the week presented in this article also produced some significant differences in positive and negative affect, although with low to moderate effect sizes associated. The outcome of a higher positive affect and a lower negative affect in the evenings is similar to other previous findings (Ayuso-Mateos et al. 2013). On the other hand, in the present study, positive affect was related with weekends, whereas weekdays were associated with worse mood outcomes (higher negative affect scores), as shown in former studies (Ryan et al. 2010).

Nevertheless, the results of the present study should be interpreted taking into account some limitations. Regarding the fact that in the COURAGE in Europe study only two indicators were considered to define the positive affect factor, this relates to a decision reached by the consensus of experts responsible for the development of the instrument in the pilot study of the World Health Organization Study on Global Ageing and Adult Health (SAGE) (Kowal 2012). In that study, the adjectives *calm/relaxed* and *enjoying* were considered the only ones capable of representing the construct of positive affect without adding redundant information.

An additional limitation of the present study could be the different sampling strategy conducted in Finland. Due to the field work specific circumstances across the three countries, the sampling methodology had to be strategically adapted in each of them. However, a probability sampling method was employed in all cases and the three samples were nationally representative.

Also, the influence of residential setting and level of education on the affect scores has not been studied, and this could have been a source of variability across groups since there is a high heterogeneity of the respondents regarding these variables. Future studies could analyze how socio-demographic information may account for some of the differences found in positive and negative affect measured with the DRM.

Although evidences of adequate model fit were found for the three samples, the total proportion of variance explained by the model in each case was not excessively high -especially in Finland-. The remaining amount of variance could be partly explained by unknown, lurking variables or inherent variability. However, further studies could test and develop alternative models with a different factorial structure.

On the other hand, one of the strengths of the present study is that it uses data from the COURAGE in Europe project, which comprises large nationally representative samples from three European countries (Finland, Poland, and Spain) offering a broad representation across different European regions. Moreover, the analysis of the abbreviated DRM factorial invariance across the three European countries, the three time frames, and the days of the week for which activities were reported, was conducted by means of sequential constraint imposition, using the multiple group CFA framework. This method improves the traditional calculation of the factorial congruence coefficient (Mulaik 1972), which is inadequate because it does not take into account all the parameters that define the invariance, and because if rejected, it provides no information on the origin of this lack of invariance (Elosua and López-Jáurega 2002). Furthermore, the time frame and day of the week invariance studies make it possible to consider the scores produced by the abbreviated version of the DRM as valid and equivalent across a wide range of important dimensions that should help to plan the design study when researching subjective well-being.

A final observation when interpreting the cross-country affect results entails to take into account that ideas about well-being differ substantially between societies (Napier, et al., 2014). Cultural values and norms are frequently implicit and using standardized survey items as well as closed-ended questions might impose meanings in some cultural contexts (Watkins, 2010). When assessing well-being, and after controlling for income and other life circumstances, a certain amount of residual variance between regions remains unexplained, and culture has been put forth as one possible explanatory factor (Napier et al. 2014). While, no cognitive debriefing was conducted for the items in the present work, a robust

standardized translation protocol (The World Health Organization 2013) was performed in order to prevent any potential bias in their interpretation.

Additionally, the items used to elicit the subject's emotional evaluation of the reported activities followed no formal process of selection from an appropriate list of adjectives; yet, Kahneman and Krueger (2006) recommend this list should vary depending on the purpose of the study. Furthermore, according to Russell's theory of core affect (2003), describing these emotions items along two dimensions (from pleasure to displeasure, and from highly activated to deactivated) provides a useful framework for considering descriptors to include in a study of well-being. In this line, originally, the DRM authors tried to select emotions that represented this spectrum, and although the response scales can be understood unevenly, scores obtained through the application of DRM can be used to generate more complex and solid measures - such as the amount of time the subject goes into a negative mood or U-index (Kahneman and Krueger 2006) to compare the affect across cultures. This said, despite the important strengths that the quantitative methodology employed has, its ability to aid understanding on subjective well-being could be enforced by incorporating anchoring vignettes, or other qualitative methodologies in further works using the DRM to help to take cultural bias evidence into account when comparing subjective well-being across cultures, instead of simply correcting for it (Napier et al. 2014).

A further question would be whether it is necessary to treat the items that presented differential performance across samples. In the present study, the possibility of eliminating them was not considered, as it does not make sense to develop a different version of the questionnaire for each sub-population under study given the considered invariance analysis (Cheung and Rensvold 2002). This procedure might also involve an excessive uniformity of

the group of items, therefore decreasing their validity and predictive accuracy (Drasgow et al. 1987). In this case, the differences found could indicate differences in the theoretical conception of the items. For example, for the present data, the most frequently non-invariant item was *rushed*, for which it could be worthwhile to perform a cognitive debriefing to ensure that it is both culturally acceptable and contextually relevant to target populations. A suggestion would be to heed the opinion of experts who could recommend, judge and decide on the maintenance or revision of the variables, thus complementing the statistical analysis outcomes with a more practical or substantive analysis (Zieky 1993).

In conclusion, the scores on the abbreviated version of the DRM have shown a reasonable invariance across three European countries, three different time frames and different days of the week. Therefore it may be assumed that the differences found among them are due to differences in affect across groups in the absence of biases associated with the country, the time frame considered, and the day of the week. These results provide evidences that enable the authors to recommend the use of the abbreviated version of the DRM as a valid method to measure affect in large-scale surveys, with a solid psychometric basis.

Compliance with Ethical Standards

Disclosure of potential conflicts of interest: The authors declare that they have no conflict of interest. The views expressed in this paper are those of the authors, and do not necessarily represent the views or policies of the World Health Organization.

Research involving Human Participants: Ethical approvals were obtained from all participant institutions. Informed consent from each individual was also obtained.

References

Ayuso-Mateos, J. L., Miret, M., Caballero, F. F., Olaya, B., Haro, J. M., Kowal, P., et al. (2013). Multi-country evaluation of affective experience: validation of an abbreviated version of the day reconstruction method in seven countries. *PLoS One, 8*(4), e61534, doi:10.1371/journal.pone.0061534.

Bradburn, N. M. (1969). *The structure of psychological well-being*. Chicago: Aldine.

Bylsma, L. M., Taylor-Clift, A., & Rottenberg, J. (2011). Emotional reactivity to daily events in major and minor depression. *J Abnorm Psychol, 120*(1), 155-167, doi:10.1037/a0021662.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 3*(105), 456-466.

Caballero, F. F., Miret, M., Olaya, B., Perales, J., López-Ridaura, R., Haro, J. M., et al. (2014). Evaluation of Affect in Mexico and Spain: Psychometric Properties and Usefulness of an Abbreviated Version of the Day Reconstruction Method *Journal of Happiness Studies*(15), 915–935.

Campo-Arias, A., & Oviedo, H. C. (2008). Propiedades Psicométricas de una Escala: la Consistencia Interna. *Revista de Salud Pública, 10*(5), 831-839.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Deaton, A. (2008). Income, health, and well-being around the world: evidence from the Gallup World Poll. *J Econ.Perspect.*(22), 53-72.

Devins, G. M., Beiser, M., Dion, R., Pelletier, L. G., & Edwards, R. G. (1997). Cross-cultural measurement of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the Affect Balance Scale. *American Journal of Public Health*(87), 794–799.

Diener, E., Ng, W., Harter, J., & Arora, R. (2010). Wealth and happiness across the world: material prosperity predicts life evaluation, whereas psychosocial prosperity predicts positive feeling. *Journal of Personality and Social Psychology*(99), 52–61.

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective Well-Being: Three Decades of Progress *Psychological Bulletin*(125), 276–302.

Dockray, S., Grant, N., Stone, A. A., Kahneman, D., Wardle, J., & Steptoe, A. (2010). A Comparison of Affect Ratings Obtained with Ecological Momentary Assessment and the Day Reconstruction Method. *Soc Indic Res, 99*(2), 269-283, doi:10.1007/s11205-010-9578-7.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach. Princeton, NJ: Educational Testing Service.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*(11), 59-79.

Elosua, P., & López-Jáurega, A. (2002). Indicadores de dimensionalidad para ítems binarios. *Metodología de las Ciencias del Comportamiento*(4), 121-137.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. [Research Support, N.I.H., Extramural Review]. *Med Care, 44* (11 Suppl 3), S78-94, doi:10.1097/01.mlr.0000245454.12228.8f.

Hox, J. J., Mass, C. J. M., & Brinkhuis, J. S. (2010). The effect of estimation mehod and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64*, 157-170.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Kahneman, D., Diener, E., & Schwarz, N. (1999). *Well-being: The foundations of hedonic psychology*

New York: Russell Sage Foundation.

Kahneman, D., & Krueger, A. (2006). Developments in the Measurement of Subjective Well-Being. *Journal of Economic Perspectives*(20), 3–24.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: the day reconstruction method. *Science, 306* (5702), 1776-1780, doi:10.1126/science.1103572

Kowal, P. C., S.; Naidoo, N.; Biritwum, R.; Fan, W.; Lopez Ridaura, R.; Maximova, T.; Arokiasamy, P.; Phaswana-Mafuya, N.; Williams, S.; Snodgrass, J. J.; Minicuci, N.; D'Este, C.; Peltzer, K.; Boerma, J. T. (2012). Data resource profile: the World Health Organization Study on global AGEing and adult health (SAGE). *Int J Epidemiol, 41*(6), 1639-1649, doi:10.1093/ije/dys210.

Krueger, A. B., & Schkade, D. A. (2008). The Reliability of Subjective Well-Being Measures. *J Public Econ, 92*(8-9), 1833-1845, doi:10.1016/j.jpubeco.2007.12.015.

Krueger, A. B., & Stone, A. A. (2008). Assessment of pain: a community-based diary survey in the USA. *Lancet, 371*(9623), 1519-1525, doi:10.1016/S0140-6736(08)60656-X.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*(33), 159–174.

Leonardi, M., Chatterji, S., Koskinen, S., Ayuso-Mateos, J. L., Haro, J. M., Frisoni, G., et al. (2013). Determinants of Health and Disability in Ageing Population: The COURAGE in Europe Project(Collaborative Research on Ageing in Europe). *Clinical Psychology and Psychotherapy*, doi:10.1002/cpp.1856.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah, NJ: Lawrance Erlbaum Associates.

Miret, M., Caballero, F. F., Chatterji, S., Olaya, B., Tobiasz-Adamczyk, B., Koskinen, S., et al. (2014). Health and happiness: Results from the Collaborative Research on Ageing in Europe (COURAGE in Europe) project. *Bulletin of the World Health Organization, 92*(10), 716–725.

Miret, M., Caballero, F. F., Mathur, A., Naidoo, N., Kowal, P., Ayuso-Mateos, J. L., et al. (2012). Validation of a measure of subjective well-being: an abbreviated version of the day reconstruction method. *PLoS One, 7*(8), doi:10.1371/journal.pone.0043887.

Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. New York: McGraw Hill.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.
Napier, A. D., Ancarno, C., & Butler, B. (2014). Culture and Health. *Lancet*(384), 1607-1639.
Oishi, S. (2010). Culture and well-being: conceptual and methodological issues. In J. F. H. E. Diener, & D. Kahneman (Ed.), *International differences in wellbeing* (pp. 34-69). New York: Oxford University Press.
Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.
Russell, J. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review, 110*(1), 145–172.
Ryan, R. M., Bernstein, J. H., & Brown, K. W. (2010). Weekends, work, and well-being: psychological need satisfactions and day of the week effects on mood, vitality, and physical symptoms. *Journal of Social and Clinical Psychology, 29*(1), 95-122.
Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Educational Research;, 99*(6), 323-338.
Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.
StataCorp (2010). *Stata Statistical Software. Release 11*. College Station, TX: Stata Corporation.
Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modelling. *Personality and Individual Differences, 42*, 893-898.
Üstün, T. B., Chatterji, S., Mechbal, A., & Murray, C. J. L., & WHS Collaborating groups (2005). Quality assurance in surveys: standards, guidelines and procedures. In *Household Surveys in Developing and Transition Countries*. New York: United Nations Statistics Division.
van de Schoot, R. L., P.; Hox, J. (2012). A checklist for testing measurement invariance. European Journal of Developmental Psychology. *European Journal of Developmental Psychology, 9*, 486-492.
Vázquez, C., & Hervás, G. (2013). Addressing current challenges in cross-cultural measurement of well-being: The Pemberton Happiness Index. In A. D. Fave, & H. H. Koop (Eds.), *Well-being and cultures. Perspectives from Positive Psychology* (pp. 31-49). NY: Springer-Verlag.
Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen, & H. Wainer (Eds.), *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Wilkinson, R. G., & Pickett, K. E. (2007). The problems of relative deprivation: why some societies do better than others. *Soc.Sci.Med*(65), 1965-1978.
Wilkinson, R. G., & Pickett, K. E. (2009). *The spirit level: why more equal societies almost always do better*. London: Penguin Books.
The World Bank (2010). http://data.worldbank.org/indicator/SI.POV.GINI/countries?display=default.
The World Bank (2011). http://data.worldbank.org/country.
The World Health Organization (2013). http://www.who.int/substance_abuse/research_tools/translation/en/.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). New Jersey: Lawrence Erlbaum Associates.

Table 1. Example of the Day Reconstruction Method questions' structure as they appear in the questionnaire.

Now, please think about how you felt yesterday during that time of the day. Please rate between 0 and 6 where 0 means you did not feel like that at all and 6 means you felt very much like that.								
		Not at all						Very much
X_1	How calm or relaxed were you feeling?	0	1	2	3	4	5	6
X_2	How much were you enjoying what you were doing?	0	1	2	3	4	5	6
X_3	How worried were you feeling?	0	1	2	3	4	5	6
X_4	How rushed were you feeling?	0	1	2	3	4	5	6
X_5	How irritated or angry were you feeling?	0	1	2	3	4	5	6
X_6	How depressed were you feeling?	0	1	2	3	4	5	6
X_7	How tense or stressed were you feeling?	0	1	2	3	4	5	6

NOTE: Respondents were asked to reconstruct a portion (morning, afternoon, or evening) of their previous day's activities, depending on the set (A, B, and C, respectively) to which they were randomly assigned. Then they assessed their associated affect by answering these questions for each one of the (1-10) activities that they reported.

Table 2. Socio-demographic characteristics by country.

	Finland (*n*= 1840)	**Poland (*n* = 3929)**	**Spain (*n* = 4581)**	**Effect size**
Female: n (%)	1041 (56.6%)	2360 (60.1%)	2505 (54.7%)	0.05
Age, years: mean (SD)	58.3 (16.0)	57.0 (17.9)	59.7 (15.9)	0.07
Highest education level completed: n (%)				0.27
Low	218 (11.9%)	984 (25.0%)	2532 (55.3%)	
Middle	981 (53.3%)	2148 (54.7%)	1423 (31.1%)	
High	640 (34.8%)	797 (20.3%)	625 (13.7%)	
Marital status: n (%)				0.07
Never married	269 (14.6%)	660 (16.8%)	667 (14.6%)	
Currently married/cohabiting	1166 (63.4%)	2194 (55.8%)	2776 (60.6%)	
Separated/divorced	195 (10.6%)	296 (7.5%)	342 (7.5%)	
Widowed	210 (11.4%)	779 (19.8%)	796 (17.4%)	
Rural setting: n (%)	441 (22.4%)	1698 (43.2%)	625 (13.6%)	0.31

NOTE: All the differences were significant at a 99% confidence level. Effect size: Cramer's *V* for χ^2 test (categorical variables) and Cohen's *f* for ANOVA test (quantitative variables).

Table 3. Goodness-of-fit indices and standardized factor loadings for the two-correlated-factors model, by country.

	Finland (n = 1840)		**Poland (n = 3929)**		**Spain (n = 4581)**	
χ^2(d.f.)	3.915* (13)		3.459* (13)		3.423* (13)	
CFI	0.940		0.980		0.974	
TLI	0.904		0.968		0.957	
RMSEA	0.058		0.048		0.062	
(90% CI)	(0.047, 0.069)		(0.040, 0.055)		(0.055, 0.069)	
ITEMS	**PA**	**NA**	**PA**	**NA**	**PA**	**NA**
Calm/relaxed	0.93		0.93		0.98	
Enjoying	0.71		0.85		0.77	
Worried		0.62		0.86		0.86
Rushed		0.64		0.72		0.86
Irritated/angry		0.77		0.86		0.91
Depressed		0.53		0.90		0.81
Tense/stressed		0.88		0.84		0.91

NOTE: d.f. = degrees of freedom; *$p < 0.001$. PA: positive affect; NA: negative affect.

Table 4. Goodness-of-fit indices for the different cross-country invariance models.

	RMSEA (90% CI)	CFI	ΔCFI	TLI
Configural	0.056 (0.051, 0.060)	0.972	-	0.955
Metric	0.065 (0.061, 0.069)	0.952	0.020*	0.938
Partial metric [λ_6]	0.056 (0.052, 0.061)	0.965	0.007	0.954
Strong	0.059 (0.055, 0.063)	0.955	0.010	0.949
Strict	0.071 (0.067, 0.074)	0.918	0.037*	0.928
Partial strict [Θ_4]	0.065 (0.061, 0.068)	0.934	0.021*	0.939
Partial strict [Θ_4,Θ_5]	0.061 (0.058, 0.065)	0.942	0.013*	0.945
Partial strict [Θ_4,Θ_5,Θ_2]	0.059 (0.055, 0.062)	0.948	0.007	0.950

NOTE: In square brackets: parameters which were considered as free to assess partial invariance; λ_6 = Loading of *depressed* on the negative affect factor; Θ_2,Θ_4,Θ_5 = residual variances associated to the items *enjoying*, *rushed*, and *irritated*/*angry*, respectively; *ΔCFI > 0.01 regarding the previous invariant model (configural/partial metric/strong).

Table 5. Goodness-of-fit indices for the invariance models across time frame and day of the week.

	Finland		Poland		Spain	
Time Frame (Morning, Afternoon, Evening)	**RMSEA**	**CFI**	**RMSEA**	**CFI**	**RMSEA**	**CFI**
Configural	0.056	0.944	0.054	0.976	0.064	0.973
Metric	0.052	0.939	0.047	0.975	0.059	0.970
Strong	0.054	0.920*	0.047	0.972	0.059	0.964
Partial Strong [μ_4]	0.052	0.930				
Strict	0.063	0.873*	0.037	0.979	0.050	0.968
Partial Strict [Θ_4]	0.056	0.902*				
Partial Strict [Θ_4, Θ_5]	0.051	0.919*				
Partial Strict [$\Theta_4, \Theta_5, \Theta_3$]	0.046	0.937				
Day of the Week (Weekday, Weekend)	**RMSEA**	**CFI**	**RMSEA**	**CFI**	**RMSEA**	**CFI**
Configural	0.061	0.934	0.046	0.980	0.066	0.973
Metric	0.058	0.929	0.042	0.981	0.063	0.970
Strong	0.059	0.914*	0.041	0.978	0.062	0.966
Partial Strong [μ_4]	0.057	0.922				
Strict	0.057	0.909*	0.033	0.983	0.053	0.970
Partial Strict [Θ_4]	0.053	0.921				

NOTE: In square brackets, parameters which were considered as free to assess partial invariance; μ_4 = Intercept of *rushed*; Θ_3, Θ_4, Θ_5 = residual variances associated to the items *worried, rushed, and irritated/angry*, respectively; *ΔCFI > 0.01 regarding the previous invariant model (configural/ metric/partial strong).

Table 6. Latent means on positive and negative affect in each country.

	Finland [(1)]		Poland		Spain		Effect size for each pairwise comparison		
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Finland-Poland**	**Finland-Spain**	**Poland-Spain**
Positive affect	0.00	1.00	-0.59*	1.42	-0.10*	0.96	0.45	0.10	0.41
Negative affect	0.00	1.00	0.50*	1.45	0.68*	1.36	0.38	0.54	0.13

NOTE: [(1)] Reference group; * All the differences were significant at 99% confidence level regarding the reference group. Effect size: Cohen's *d*.

Table 7. Latent means (SD) on positive and negative affect by set and day of the week in each country.

	Positive affect			Negative affect		
	Finland	**Poland**	**Spain**	**Finland**	**Poland**	**Spain**
Time Frame						
(Ref. Morning)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)
Afternoon	-0.09 (0.98)	-0.05 (0.99)	0.06 (0.96)	0.16* (1.03)	0.02 (0.95)	0.01 (1.01)
Evening	0.21*** (0.81)	0.11** (0.92)	0.13*** (0.94)	-0.14** (0.67)	-0.06 (0.90)	-0.05 (1.01)
Day of the Week						
(Ref. Weekday)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)
Weekend	0.29*** (0.79)	0.16*** (0.96)	0.03 (1.11)	-0.24*** (0.66)	-0.09* (0.94)	-0.07 (1.00)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, regarding the reference group in each case.

Fig. 1 Standardized loadings (λ) and factor correlations for both, the configural model (**a**) and the partial strict invariance model (**b**), in each country.

NOTE: Standardized loadings are not the same despite the fact that metric invariance has been achieved, due to variation in error variances and latent factor variances