Workplace stress in real time: Three parsimonious scales for the experience sampling measurement of stressors and strain at work

Luca Menghini[1], Massimiliano Pastore[2], Cristian Balducci[1]

1. Department of Psychology, University of Bologna, Italy

2. Department of Developmental and Social Psychology, University of Padua, Italy

SUPPLEMENTARY MATERIAL S1: A priori power analysis results

The supplemental material S1 includes a brief description of the procedure used to establish a 'reasonable' sample size to evaluate the factor structure of the proposed set of ESM scales, and the obtained results (more details and the R code are reported in the Supplemental Material S2). To be brief, here we focus on the Multidimensional Mood Questionnaire (MDMQ) (Wilhelm & Schoebi, 2007), implying the most complex measurement model among those considered in our study (i.e., nine items measuring three dimensions vs. three/four items measuring one dimension for the Task Demand Scale and the Task Control Scale).

Power analysis was implemented using R (Version 4.0.3; R Core Team, 2018) and the R-packages *ggplot2* (Version 3.3.3; Wickham, 2016), *gridExtra* (Version 2.3; Auguie & Antonov, 2017), *lavaan* (Version 0.6.7; Rosseel, 2012), *MASS* (Version 7.3.53; Venables & Ripley, 2002), *reshape2* (Version 1.4.4; Hadley Wickham, 2007), and *stringr* (Version 1.4.0; Hadley Wickham, 2017).

1. Procedure

The Monte Carlo approach was used to randomly simulate 10,000 samples for each of 44 combinations of sample sizes at level 2 (i.e., participants; possible N2s = 50, 100, 150, 200, 250, 300, 350, 400, 500, 800, 1000) and level 1 (i.e., occasions per participant; possible N1s = 5, 10, 15, 21), and standardized loadings on both levels (possible values = .40, .60, .80). In each

simulation, a sample of mood item scores was generated by aggregating a level-2 dataset (with N2 rows) with N2 level-1 datasets (with N1 rows). Each sample was generated from the same between-participants and within-participant covariance matrices, both defined with the following model written in R code:

```
      # measurement model
m <- 'NegativeValence =~ LOAD*x1 + LOAD*x2 + LOAD*x3
      TenseArousal =~ LOAD*x4 + LOAD*x5 + LOAD*x6
      Fatigue =~ LOAD*x7 + LOAD*x8 + LOAD*x9

      # correlations between latent factors
      NegativeValence ~~ .80 * TenseArousal
      TenseArousal ~~ .59 * Fatigue
      Fatigue ~~ .70 * NegativeValence'
```

Where *LOAD* is the pre-set loading value (i.e., .40, .60, .80; using the same value for all items), *x1* … *x9* are the scores to the nine items of the MDMQ (observed variables), and the =~ and ~~ symbols stand for "is manifested by" and "correlates with", respectively. Correlations between latent factors were set based on those reported by Wilhelm & Schoebi (2007) at the between level. Only in the case of the correlation between Negative Valence and Tense Arousal, we used the parameter reported at the within level, since the authors were unable to distinguish the two dimensions at level 2 (the reported correlation was .99), whereas we hypothesized a three-factor model at both levels for our version of the MDMQ.

Then, the following R code was used to fit a multilevel model on each simulated sample, and parameter estimates were stored into a dataset of parameters:

```
mcfa <- 'level: 1
         NegativeValence_within =~ x1 + x2 + x3
         TenseArousal_within =~ x4 + x5 + x6
         Fatigue_within =~ x7 + x8 + x9
         level: 2
         NegativeValence_between =~ x1 + x2 + x3
         TenseArousal_between =~ x4 + x5 + x6
         Fatigue_between =~ x7 + x8 + x9'
```

Parameters were stored considering only models that reached convergence with no Heywood cases. Then, parameter estimates were used to visualize power variation over sample sizes and loading values assumed for the population (see below).

2. Simulation diagnostics and goodness of fit results

For each of the 1,320,000 simulated samples, we fitted the hypothesized three-factor MCFA configural model, with the same three mood dimensions at both levels, but unequal factor loadings across levels (m3x3, see main article). As shown in Figure S1B and S1C, our simulations showed satisfactory convergence rates, and satisfactorily low rates of negative variance estimates, with fit indices confirming the goodness of the simulation procedure. Samples generated from a population with factor loadings = .40, and N2 < 100 were associated with higher rates of nonconvergence and improper solutions.

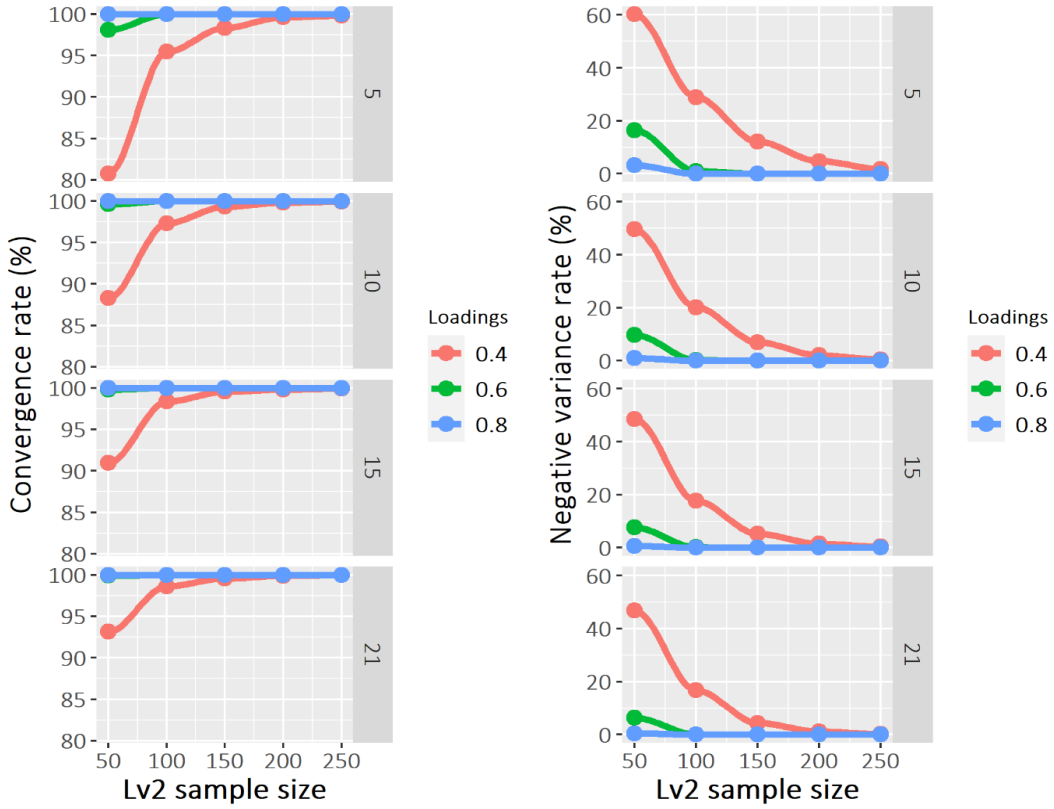

Figure S1B. Percentage of samples in which the MCFA model reached convergence (left panel), and showed one or more negative residual variance estimates (right panel) by the value pre-set to the factor loadings generating the sample (colors), and the sample size at level 1 (vertical panels) and 2 (x-axis).

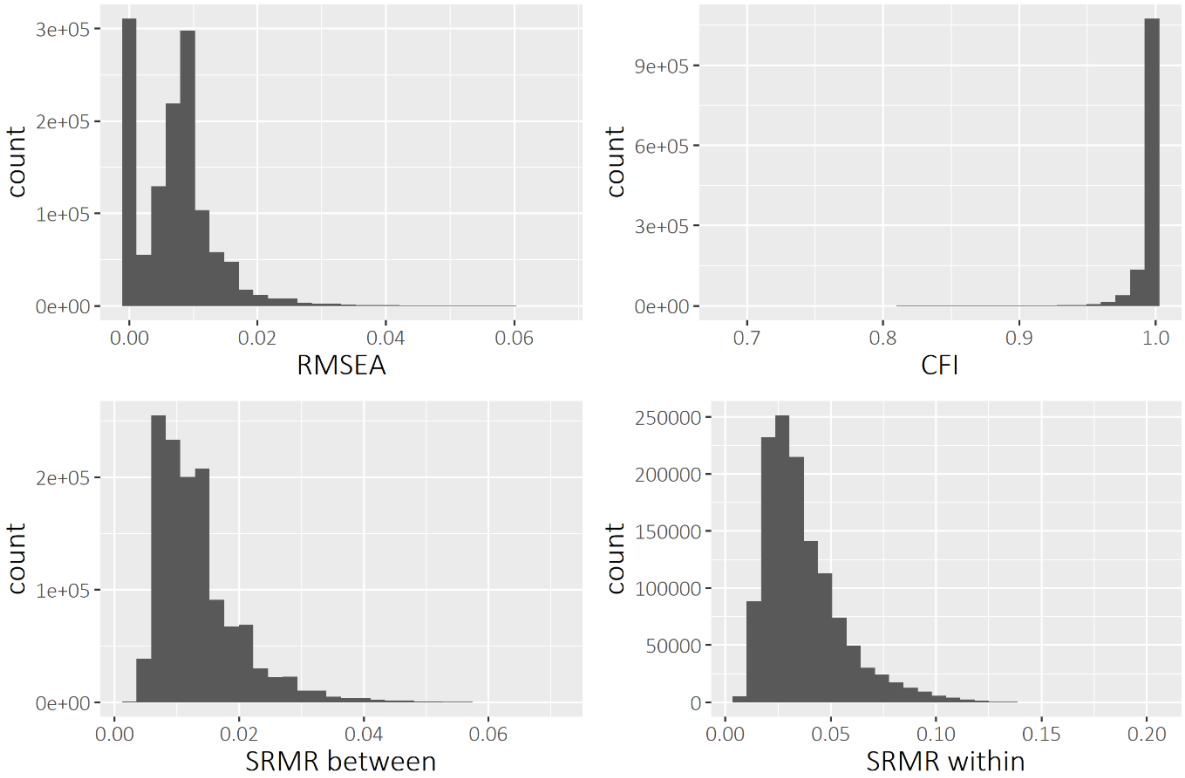

Figure 61C. Distribution of fit indices obtained from the multilevel models fitted on the simulated data (i.e., considering only those models that reached convergence with no Heywood cases when fitted on each combination of loadings and level-1 and level-2 sample sizes). **RMSEA** = root mean square error of approximation; **CFI** = comparative fit index; **SRMR-B** = SRMR between subjects; **SRMR-W** = standardized root mean squared residual within subject.

3. Power analysis of the configural model m3x3

Then, we performed a power analysis by defining statistical power as the *percentage of MCFA models in which all standardized loadings were significantly higher ($p < .05$) than the arbitrary chosen cut-off value of .15, at both levels*. We considered a statistical power of 80% or above as satisfactory.

Power plots are shown in Figure S1D. Whereas at level 1 the power was > 80% in all cases with N1 ≥ 10 or N2 ≥ 100, only a samples with 500 or more participants would show a satisfactory level-2 power when population loadings are close to 0.4. In contrast, a satisfactory statistical power was showed by all scenarios in which the population parameters were ≥ 0.6

and the sample size at level 2 was ≥ 100. Assuming a population with factor loadings ≥ 0.60 for our Italian version of the MDMQ, we concluded that a sample size of 100 or more participants with five or more responses each was adequate for evaluating its construct validity.

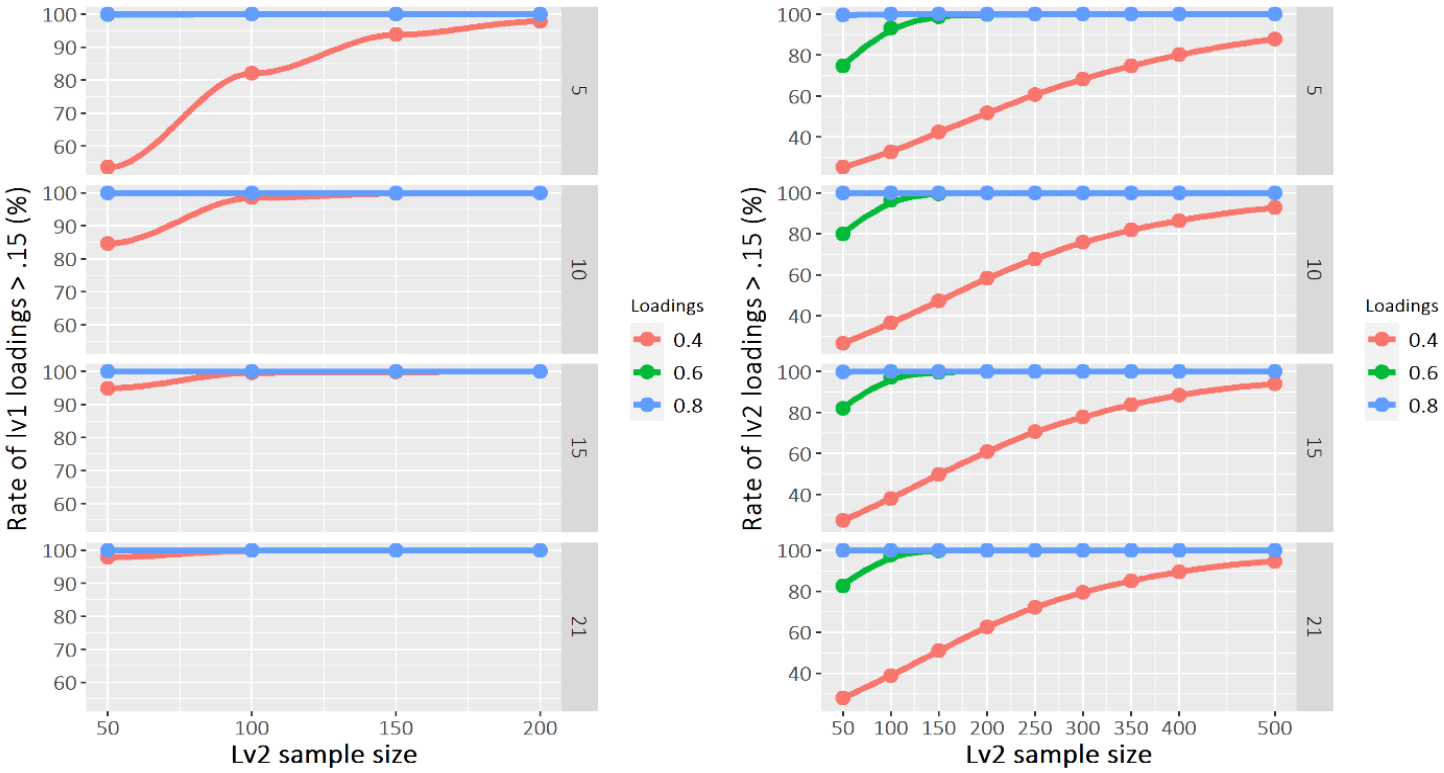

Figure 61D. Percentage of simulated samples from which all estimated factor loadings at level 2 (between, left panel) and level 1 (within, right panel) were significantly higher (p < .05) than .15, depending on the pre-set factor loadings generating the sample (colors), and the sample size at level 1 (vertical panels) and level 2 (x-axis).

4. References

Auguie, B., & Antonov, A. (2017). *gridExtra: miscellaneous functions for "grid" graphics*. https://cran.r-project.org/package=gridExtra

R Development Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, http://www.r-project.org/. http://www.r-project.org/

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics (Fourth S., editor) New York*. Springer. http://www.stats.ox.ac.uk/pub/MASS4/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis 2016 Springer-Verlag, New York*. https://ggplot2.tidyverse.org

Wickham, Hadley. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wickham, Hadley. (2017). *Stringr: Simple, consistent wrappers for common string operations*. https://cran.r-project.org/package=stringr

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment*, *23*(4), 258–267. https://doi.org/10.1027/1015-5759.23.4.258