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 S3.2: Supplementary analysis of time trends, measurement reactivity, and sensitivity to contextual factors

The Supplementary Material S3.2 includes a deeper analysis of the time trends in each evaluated experience sampling methods (ESM) scale, also accounting for measurement reactivity. Indeed, in addition to the "natural" patterns of construct change over time (e.g., the daily trend in Fatigue ratings highlighted in the main manuscript, or the "blue Mondays" effect sometimes highlighted for mood ratings across weekdays, e.g., Pindek et al., 2020), ESM ratings can biased by the repeated-measure, sometimes annoying, nature of the data collection. For instance, Shrout et al. (2018) highlighted that subjective reports may show an "initial elevation bias", with more negative ratings in the first days/occasions compared to the following ones. Such a measurement reactivity (see Barta et al., 2012) is particularly critical in workplace stress research, where stressor-related strain responses should be clearly disambiguated from measurement-related strain. To account for measurement reactivity (i.e., initial elevation bias), we re-analyzed ESM differences between and within weekdays, and across work sampling categories, by also considering the temporal order of days and individual responses within-day.

To evaluate initial elevation bias between days, we allowed participants to freely choose when starting the protocol (i.e., either on Monday, on Wednesday, or Friday). This choice resulted in a relatively balanced number of participants starting in each weekday (i.e., 36.7% of participants started on Monday, 33.8% on Wednesday, 29.5% on Friday).

1. Analytical strategy

For each ESM measure, we specified a set of linear mixed-effects regression (LMER) models by progressively including the predictors described below to a null model (M0) (i.e., only intercept and random deviations between respondents). At each step, the corresponding model was compared with the null model, and with all previous models based on the Akaike weights (Aw) (Wagenmakers & Farrell, 2004). At each step, the corresponding predictor was retained in the following models only when its inclusion was associated with higher Aw than the null model and all previously selected models.

First, we included the linear effect of time, operationalized as the number of hours since the first response of each day, and included in the first model (M1) in addition to the autoregressive term (i.e., controlling for the corresponding stressor/strain rating entered on the previous occasion). As shown in the main manuscript, Fatigue was the only variable showing a substantial linear trend over time. Second, we included a categorical (dummy) predictor indexing the weekday (M2) (i.e., Monday, Wednesday, or Friday) and, in a following step, a second categorical predictor indexing the temporal positioning of each day for each participant (M3) (i.e., 1st, 2nd, or 3rd day of participation). Models M3 were also re-specified by retaining the differences between weekdays from models M2, even when not substantial, in order to disentangle weekday effects from sequence effects due to initial elevation bias. Third, we evaluated initial elevation bias within the 'average' workday, by including a further categorical predictor indexing the order of data entry within each day (M4) (i.e., a set of dummy variables was created to reflect response order from the 1st to the 7th, independently from the measurement occasion).

Finally, we re-analyzed the differences in each ESM measure across categorical features of the job tasks (i.e., type and mean of work, people involved; see the main manuscript) by

including work sampling categories as additional predictors (M5) to those models selected in the previous steps. This was done both as a robustness check of the analysis reported in the main manuscript, and as a way to evaluate the potential biasing effects that measurement reactivity might exert on the conclusion of an ESM study using the proposed scales.

This supplemental analysis and the figures showed below were implemented using R (Version 4.0.3; R Core Team, 2018), and specifically the R-packages *ggplot2* (Version 3.3.3; Wickham, 2016), *gridExtra* (Version 2.3; Auguie & Antonov, 2017), *lme4* (Version 1.1.26; Bates et al., 2014), and *MuMIn* (Kamil, 2020) (see details in the full supplementary report S3).

2. Results

As reported in the main manuscript, a substantial linear trend over time (M1) was only highlighted for Fatigue, and the inspection of ESM measures across the three weekdays (M2) did not reveal any substantial difference (all Aw < .22 against the null models). In contrast, the inclusion of the temporal order of participation days (M3) resulted in substantial trends in MDMQ ratings, with lower Negative Valence (Aw = .90) and Tense Arousal (Aw = .97) reported in the first day compared to both the second (Negative Valence: b = 0.21 (standard error = 0.05), t = 4.39; Tense Arousal: b = 0.22 (0.05), t = 4.36) and the third day of participation (Negative Valence: b = 0.19 (0.05), t = 3.94; Tense Arousal: b = 0.23 (0.05), t = 4.36), independently from the weekday effect included in the previous step (M2).

As shown in Figure S3.2A, a substantial difference across days of participation (M3) was also found for the Task Control Scale (Aw = .84 against the null model), with higher Task Control reported in the first day compared to both the second (b = -0.26 (0.07), t = -3.56) and the third day of participation (b = -0.21 (0.07), t = -2.87). However, when the sequence effect was included in the model already including the differences between weekdays (M2), the former

effect was not substantial according to the Aw (i.e., null model's Aw = .97, weekday model's Aw = .005, weekday and day order model's Aw = .03). No substantial differences were found across days of participation in Fatigue (Aw = .02), Task Demand (Aw = .05), and Situational Constraints (Aw = .22).

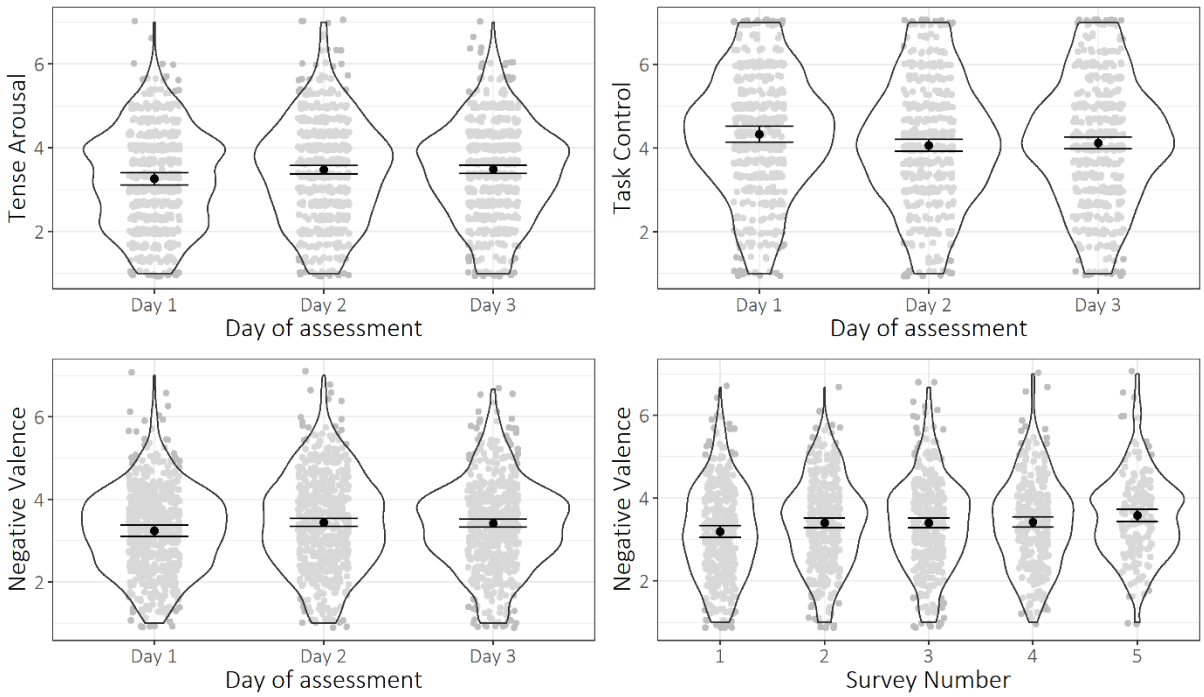

Figure S3.2A. Examples of differences highlighted in ESM scores across temporal categories (i.e., day of participation and order of data entry). Gray dots represent single observations; violin plots represent distribution densities; error bars indicate 95% confidence intervals estimated with linear mixed-effects regression models (see full supplementary report S3).

Figure S3.2A also shows the differences highlighted for Negative Valence ratings over the order of data entry (M4), whose inclusion was associated with stronger evidence (Aw = .665) compared to the null model (M0 Aw = .03), the model including the weekday effect (M1 Aw = .001), and that including both the weekday and the day sequence effect (M2 Aw = .30). Model M4 indicated higher Negative Valence in the second (b = 0.21 (0.06), t = 3.69), the third (b = 0.21 (0.06), t = 3.56), the fourth (b = 0.23 (0.06), t = 3.68), and the fifth (b = 0.40 (0.08), t = 5.26), but not the sixth (b = 0.13 (0.10), t = 1.32) or the seventh (b = 0.14 (0.17), t = 0.79)

compared to the first data entry within an 'average' workday, and controlling for both the weekday and the day sequence effect. None of the remaining MDMQ dimensions or momentary stressors showed any substantial difference over the order of data entry (all Aw against null models < .32).

Figure S3.2B shows examples of ESM differences across work sampling categories, as reported in the main manuscript (i.e., without controlling for initial elevation bias). The type of job task was recoded into five categories: "information acquisition" (N = 439, 28.75%), "data analysis" and "authoring" (N = 368, 24.10%), "administrative activities" (N = 238, 15.59%), social activities (including both "networking" and "dissemination"; N = 240, 15.72%), and "break" (N = 49, 3.21%). Job tasks marked as "other" (12.64%) were not considered. The mean of work was recoded into three categories: "on the computer" (N = 955, 62.54%), "face-to-face" (N = 367, 24.03%), and others (including "phone", "videocall", "tablet", "paper-and-pencil", and "other"; N = 205, 13.43%). The involvement of other people was binary recoded as being "alone" (N = 853, 55.35%) or "with others" (N = 680, 44.36%). Data entries in which participants indicated more than one option (27.05% for the type of activity, 25.87% for the mean of work, 12.90% for the involvement of other people) were recoded by considering only the first entered option, with the exception of "break" and "other" (i.e., only considered when selected as the unique option).

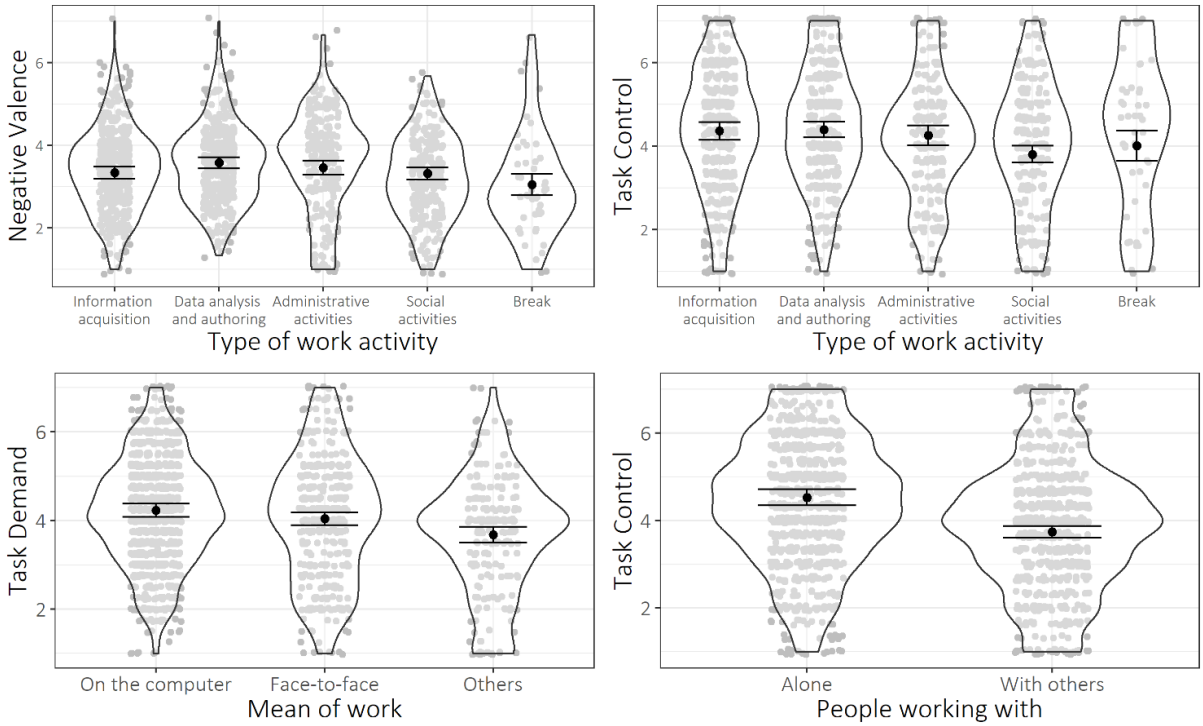

Figure S3.2B. Examples of differences in ESM scores across work sampling categories. Gray dots represent single observations; violin plots represent distribution densities; error bars indicate 95% confidence intervals estimated with linear mixed-effects regression models (see full supplementary report S3).

The inclusion in these models of the initial elevation bias effects highlighted above (M5) resulted in substantial changes for Negative Valence, whose differences across the type of job task (Aw = .93 against the null model) were no longer substantial (Aw < .001) when the latter was included in the model with the data entry and day order terms (M4). In contrast, the effect of the type of job task was still substantial when included in the model predicting Task Control by the order of participation days (Aw = .99), with social activities (i.e., networking and dissemination) being associated with lower Task Control (*b* = -0.56 (0.10), *t* = -5.417) than "information acquisition" and other job task categories. Similarly, the inclusion of the term for initial elevation bias across days of participation did not substantially reduce the differences in Task Control between job tasks performed "one the computer" and those performed "face-to-face" (*b* = -0.66 (0.08), *t* = -8.06) or by using "other" means (*b* = -0.33 (0.10), *t* = -3.37; Aw =

.99), neither it reduced the substantial Task Control differences highlighted between job tasks performed “alone” and those performed with other people (*b* = -0.78 (0.07), *t* = -11.44).

3. Short discussion

The additional analysis reported above highlights the potential biasing effects of measurement reactivity on the proposed ESM scales, as well as on their relationships with other variables such as work sampling categories. Three of the proposed ESM measures (Negative Valence, Tense Arousal, and Task Control) showed potential signs of measurement reactivity throughout the three days of participation, whereas only Negative Valence showed a substantial trend over the order of data entry within the ‘average’ workday. Critically, the differences highlighted for Negative Valence across the considered types of job tasks were no longer substantial after controlling for these initial elevation biases.

Coherently with Shrout et al. (2018), we observed sings of these biases for variables referred to internal and negative states (i.e., Negative Valence and Tense Arousal), with small but substantial differences between the first and the following days/occasions, comparable with those found across work sampling categories, and thus, of practical significance. Moreover, the 3×3 factorial design accounting for both weekday and day of participation allowed to exclude further interpretations such as the “blue Mondays” effect.

Whereas alternative interpretations of this bias were proposed (e.g., sometimes referred as “attenuation effect”, suggesting that repeated ratings are likely to show a decline throughout the data collection, and thus that later reports are biased), Shrout et al. (2018) argued and provided evidence in favor of an initial bias effect (i.e., initial reports are biased), which seems compatible with our findings (i.e., only the first day/occasion was different from the last ones). However, we found an opposite trend than that highlighted by the authors, with

more positive reports during initial compared to later measurements. Potential explanations might involve the use of bipolar response scales in our study, or the potentially stronger effect of social desirability in job-related self-reports, but further studies are needed to better investigate these differences.

In any case, these results highlight how scholars and practitioners should be aware of the sources of bias potentially characterizing ESM, such as the initial elevation bias, and take actions accordingly (e.g., excluding the first days/occasions). This is particularly critical in workplace stress assessment, where stressor reactivity (i.e., strain) needs to be clearly distinguished from measurement reactivity. Letting participants to freely chose when starting the assessment is useful both for improving compliance, and for disambiguating the effect of weekdays by initial elevation biases and other forms of measurement reactivity (Gabriel et al., 2019; Shrout et al., 2018).

4. References

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

Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement Reactivity in Diary Research. In M. S. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Sudying Daily Life* (pp. 108–123). The Guilford Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv Preprint ArXiv:1406.5823*. http://arxiv.org/abs/1406.5823

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience Sampling Methods: A Discussion of Critical Trends and Considerations for Scholarly Advancement. *Organizational Research Methods*, *22*(4), 969–1006. https://doi.org/10.1177/1094428118802626

Kamil, B. (2020). *MuMIn: Multi-Model Inference*. https://cran.r-project.org/package=MuMIn

Pindek, S., Zhou, Z. E., Kessler, S. R., Krajcevska, A., & Spector, P. E. (2020). Workdays are not

created equal: Job satisfaction and job stressors across the workweek. *Human Relations*, 001872672092444. https://doi.org/10.1177/0018726720924444

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/

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

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