Workplace Stress in Real Time: Three Parsimonious Scales for the Experience Sampling Measurement of Stressors and Strain at Work

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

* Corresponding Author

Luca Menghini Ph.D.

Department of Psychology, University of Bologna,

Viale Carlo Berti Pichat, 5 – 40127, Bologna, Italy

E-mail: luca.menghini3@unibo.it; luca.menghini.job@gmail.com

Tel: +39-340-7524899

Abstract

Experience sampling methods are increasingly used in workplace stress assessment, yet rarely developed and validated following the available best practices. Here, we developed and evaluated parsimonious measures of momentary stressors (Task Demand and Task Control), and the Italian adaptation of the Multidimensional Mood Questionnaire as an indicator of momentary strain (Negative Valence, Tense Arousal, and Fatigue). Data from 139 full-time office workers that received seven experience sampling questionnaires per day over three workdays suggested satisfactory validity (including weak invariance cross-level isomorphism), level-specific reliability, and sensitivity to change. The scales also showed substantial correlations with retrospective measures of the corresponding or similar constructs, and a degree of sensitivity to work sampling categories (type and mean of job task, people involved). Opportunities and recommendations for the investigation and the routine assessment of workplace stress are discussed.

Keywords: experience sampling methods; workplace stress assessment; scale development; psychometric properties; multilevel confirmatory factor analysis

Experience sampling methods (ESM) are increasingly used as promising alternatives to retrospective reports in organizational research (Fisher & To, 2012; Gabriel et al., 2019) and other fields of psychological assessment. Consisting of the repeated sampling of current psychological states, experiences, and activities, ESM focus on within-individual occasions at weekly, daily, or momentary level to quantify individual differences (stable levels as indexed by averaged ratings) and intraindividual fluctuations (transient deviations from stable levels). Moreover, in-context and real-time experience sampling allows linking subjective ratings to contextual episodes/conditions while minimizing recall biases (Beal, 2015).

Workplace stress research is particularly at the forefront of ESM application, with an increasing number of studies (e.g., Pindek et al., 2019) evaluating the dynamic co-occurrence of job stressors, defined as the "work-related environmental conditions thought to impact on the health and well-being of the worker", and job strain, the "worker's psychological and physiological reactions to such exposures" (Hurrell et al., 1998, p. 368). The possibility to capture key concepts of risk management (e.g., frequency of exposure) while controlling for individual-level confounders (e.g., negative affectivity) and contextual factors such as task design features (Robinson, 2009) are among the obvious advantages of ESM for both researchers and practitioners.

Yet, ESM development and validation are rarely conducted following the available best practices (Fisher & To, 2012; Gabriel et al., 2019). A Scopus search of the terms "job" and "experience sampling" or "daily diary" associated with "stress", "stressor" or "strain", covering the period 2011-2021, resulted in 57 job-related empirical articles, of which only eight used previously validated measures, or provided validity and reliability indicators at both levels (see Supplementary Materials). In 51 studies, measures were adapted from retrospective scales, but only 13 provided a rationale for item selection. Only a minority of scales was accompanied by

level-1 reliability (24.6%) or validity indices (29.8%), whereas none of them was tested for cross-level isomorphism, the invariance of the factor structure and loadings across levels – a critical condition needed when level-2 constructs (e.g., individual stress level) are conceptualized as aggregates of level-1 constructs (e.g., momentary stress levels) (Stapleton et al., 2016).

Such an increasing use of ESM without an increasing availability of valid ESM measures is particularly worrying since lack of transparency and ignorance of psychometric indicators can threaten the construct and statistical validity of a study and, ultimately, the credibility of its conclusions (Flake & Fried, 2020). To avoid wasting the potentialities of ESM for workplace stress assessment, there is a clear need for studies developing and validating ESM scales.

The present study

Here, we aimed at developing and validating a set of ESM measures to assess job stressors and strain at both momentary and individual level. Instead of focusing on a single instrument, we accounted for the lack of validated scales and the multifaceted nature of workplace stress by developing a battery of indicators of stressor and strain constructs among those that received more consolidated theoretical and empirical support.

These included Job Demand and Job Control, two key factors of widely supported models (e.g., Karasek et al., 1998) consistently associated with several strain indicators at both inter- and intraindividual level (Bowling et al., 2015; Pindek et al., 2019). Specifically, we focused on those subdimensions more connectable to the ongoing job task: workload, reflecting “the amount or difficulty of one’s work” (Bowling et al., 2015, p. 96), and decision authority, “the organizationally mediated possibilities for workers to make decisions about their work” (Karasek et al., 1998, p. 323).

Momentary strain was operationalized in terms of negative mood, in light of recent meta-analyses identifying affective strain as the most direct and immediate response to job stressors (e.g., Pindek et al., 2019), possibly creating "cognitive, motivational, and/or physical pathways to distal outcomes" (p. 6). Due to higher availability of ESM mood measures, we adapted an existing scale instead of developing a new one. We focused on the Multidimensional Mood Questionnaire (MDMQ) by Wilhelm and Schoebi (2007), also due to its compatibility with influential models of job-related affective wellbeing (e.g., Warr, 1994). The scale measures moods, defined as diffused, time-varying, and consciously available affective states distributed over three correlated but distinct dimensions, which we conceptually reversed for better matching the concept of strain: Negative Valence, Tense Arousal, and Fatigue.

Then, using multilevel data from a sample of office workers, we evaluated whether the proposed measures show the expected factor structure (H1.1): a single-factor structure for Job Demand and Job Control, and a three-factor structure for the MDMQ. In addition, we expected (H1.2) weak cross-level isomorphism for each factor, (H2) sufficient reliability at both levels, and (H3) substantial individual-level correlations with existing retrospective tools measuring the same or similar constructs (convergent validity). Since retrospective reports are the current standard to quantify job stressors and strain at the individual level (see Tabanelli et al., 2008), they represent the best available criterion for convergent validity at level 2.

Finally, we characterized the scales at the momentary level by inspecting temporal patterns within and across weekdays, and scale sensitivity to task design features (type and mean of job task, people involved). That is, we explored the possibility to differentiate objective task categories by the associated momentary appraisals, since a degree of sensitivity to working

conditions is theoretically expected, and it is critical for planning task-level interventions (e.g., job redesign).

Materials and Methods

Participants

A convenience sample of 215 Italian-speaker full-time office workers was recruited via e-mail within the university staff and the private network of the authors and their collaborators. Recruitment focused on white-collars mainly involved in back-office activities. Participation was voluntary, anonymous, and preceded by an informed consent. The study was approved by the Ethics Committee of the Departments of Psychology (University of Padova, protocol 2760). A 'reasonable' sample size of 100 or more participants with five or more observations each was estimated via a priori power analysis, based on the expected loadings from the models described below (see Supplementary Materials). 49 participants were excluded due to missing response to the preliminary and/or any ESM questionnaire, eight due to incompatible jobs (e.g., nurses), and 19 due to less than five ESM entries in total. The results reported below were obtained from 139 respondents (70 females) aged 35.04 ± 9.65 years, mainly employed in the private sector (69.1%), and mainly working as office employees (31.6%), research staff (18%), and managers (14.4%).

Procedure

The study took place between 2018 and 2019, consisting of an online preliminary questionnaire followed by an ESM protocol. The former was linked in the recruitment e-mail, also including the instructions to install and use the open-source Sensus Mobile application (Xiong et al., 2016) over three non-consecutive workdays (Monday, Wednesday, Friday). Each day, participants received seven notifications on their smartphone, scheduled each 80 to 100

min randomly determined from 9:15 to 18:15, and expiring after 20 min. The only exception was the first morning questionnaire, which was scheduled at 9.15, and expired after 60 min. Whereas strain was measured on all occasions, stressor and work sampling measures were not included in the first morning questionnaire. Filling ESM questionnaires required 4 ± 3.6 min.

Measures

ESM were developed based on a review of the existing tools in workplace stress (see Tabanelli et al., 2008) and affective research. Following recommended practices (Ohly et al., 2010; Shrout & Lane, 2012), we identified ideal compromises between the parsimony and the redundancy needed for minimizing response burden while covering multifaceted constructs and reporting on their reliability (scales with three or more items were prioritized). The final battery consisted of 16 items (see Supplementary Materials). Strain items were presented at the beginning, followed by work sampling, and stressor items. Both stressors and strain were rated on a slider scale from 1 ("Not at all") to 7 ("Very much").

Momentary stressor assessment

Task Demand Scale (TDS): three items ("work fast", "work hard", "do too much") were selected from the Quantitative Workload Inventory (Spector & Jex, 1998), validated in Italian by Barbaranelli et al. (2013), based on to their face validity, simplicity, and shared content with Job Demand items from Karasek et al. (1998). A fourth item ("doing multiple things at once") was also included to account for the multi-tasking component of Task Demand, whose manipulation has been associated with mental demand and physiological activation (e.g., Wetherell & Carter, 2014). TDS items were introduced by the instruction "In relation to the main job task performed in the last 10 minutes…".

Task Control Scale (TCS): two items from the Diary for Ambulatory Behavioral States (Kamarck et al., 2002) ("could change task if I chose to", "could schedule the time of the task"), and one item from the Instrument for Stress-oriented Task Analysis (Semmer et al., 1995) ("could decide how to perform the task") were selected due to their previous use in ESM studies, the simplicity and specificity of item wording, and the content match with the decision authority dimension. Measures of timing and method control were preferred over more general indicators of decision latitude (e.g., "a lot of say"), less indicative of modifiable task features.

Momentary strain assessment

The six MDMQ items (Wilhelm & Schoebi, 2007) were back-translated to Italian with the help of two bilinguals, and integrated with three additional items (i.e., Negative Valence: "in a positive-negative state", Tense Arousal: "nervous-placid", Fatigue: "fatigued-rested") following Peter Wilhelm's suggestion, and based on a pilot study. Items were presented consistently with the original scale in terms of response format (i.e., bipolar, with endpoints associated with the label "very") and order, with consecutive items switching both dimension and polarity (e.g., item 1: "unwell-well", item 2: "relaxed-tense"). Positively worded items were recoded so that higher scores indicated negative mood. MDMQ items were introduced by the instruction "How do you feel right now?".

Work sampling measures

Task-related contextual features were measured by adapting Robinson (2009)'s measures, including the type of work task ("what" categories were selected among knowledge work activities, e.g., "information acquisition", "networking"), the mean of work ("how", e.g., "face-to-face", "on the computer"), and the persons involved in the task ("who", e.g., "anyone",

"co-workers", "supervisor"). Items were introduced by the instruction "Think about the main job task performed in the last 10 minutes".

Retrospective reports

The preliminary questionnaire included sociodemographic indicators, and the retrospective scales measuring individual-level job stressors and strain, rated using five-point Likert scales from "Never or almost never" to "Always/Very often".

Job stressors: Job Demand was measured with the five-item Quantitative Workload Inventory (Barbaranelli et al., 2013; Spector & Jex, 1998) (Cronbach's α = .88, 95% CI [.86, .90]). Job Control was measured with three Decision Authority items from Karasek et al. (1998), also included in the Italian adaption of the Stress Indicator Tool (Toderi et al., 2013), integrated with two Influence at Work items from Thorsen and Bjorner (2010) ("influence on what you do" "influence on how quickly you work") to better match the TCS content (timing and method control) while improving reliability (α = .78 [.73, .82]).

Job strain: affective strain was operationalized in terms of Job-related Affective Wellbeing (JAW) and Burnout. JAW was measured with the 12-item measure by Van Katwyk et al. (2000), adapted and widely used in the Italian context (e.g., Balducci et al., 2010). The scale uses three items for measuring each of the four dimensions emerging from the valence and arousal axes (e.g., high-pleasure/high-arousal: "enthusiastic"), referred to the job context over the last 30 days (subscales' α [95% CI] ranging from .68 [.62, .74] to .84 [.81, .87]). Work-related Burnout was measured using the seven-item subscale of the Copenhagen Burnout Inventory (Kristensen et al., 2005) (α = .84 [.81, .87]), validated in Italian by Avanzi et al. (2013).

Data analysis

Data were analyzed with R 4.0.3. (R Core Team, 2018). First, multilevel confirmatory factor analyses (MCFAs) were conducted separately for each scale, following Kim et al. (2016). All latent variables were conceptualized as configural cluster constructs (Stapleton et al., 2016), and cross-level isomorphism was evaluated following Jak and Jorgensen (2017). Model comparison was based on the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean squared residual (SRMR), in addition to the Akaike weight (Aw), quantifying the strength of evidence (likelihood and parsimony) of multiple models, and interpretable as the probability that a model is the most evident, given the data and the set of alternative models (Wagenmakers & Farrell, 2004). RMSEA ≤ .06, CFI ≥ .95, and SRMR ≤ .08 were considered as indicative of satisfactory fit (Hu & Bentler, 1999).

Second, we evaluated the reliability of each scale by computing level-specific indices of composite reliability (ω) from MCFA models, following Geldhof et al. (2014). Moreover, following Shrout and Lane (2012), we partitioned the item scores variance by participant, time point, item, and their interactions to compute indices of between-person reliability considering either one fixed occasion (R_{1F}) or the entire set of 21 occasions (R_{KF}), in addition to the sensitivity-to-change index (R_C), reflecting the ability to detect systematic intraindividual changes over time.

Third, we analyzed the aggregated scores for each scale (i.e., occasion-specific arithmetic means of item scores) to evaluate convergent validity based on zero-order Pearson correlations at both level 1 (mean-centered scores) and level 2 (individual mean scores), and those between level-2 ESM aggregates and the corresponding retrospective scales. Based on Cohen (1988), we considered medium (.30 ≤ r < .50) and strong correlations (r ≥ .50) as

substantial. Finally, we used multilevel modeling to explore ESM scales sensitivity to contextual factors. Following an assessment of their temporal trajectories, we evaluated the size of the differences between work sampling categories. Each model was compared with the corresponding null model (either intercept-only, or intercept and time) based on the Aw. Parameters and 95% profile-likelihood confidence intervals were only inspected for models showing higher Aw (Aw > .50) than the corresponding null model.

Results

The following results were obtained from a total of 1,774 ESM data entries out of 2,919 scheduled questionnaires (response rate = 60.8 ± 15.2%), of which 86% also included stressor measures. On average, included participants responded to 12.8 ± 3.2 out of 21 questionnaires.

Momentary stressors

MCFAs indicated satisfactory fit for the single-factor weak invariance models of both TDS ($\chi^2(8) = 32.33$, RMSEA = .045, CFI = .991, SRMR-within = .016, SRMR-between = .061, Aw = .68) and TCS item scores ($\chi^2(3) = 5.38$, RMSEA = .023, CFI = .998, SRMR-within = .008, SRMR-between = .035, Aw = .68), with overall better fit indices than the respective configural and strong invariance models (see Supplementary Materials), and standardized loadings from .60 to .99 (see Figure 1)[1]. Both scales showed satisfactory reliability indices, and adequate sensitivity to change (see Table 1).

[1] The results reported for the TCS and the MDMQ were obtained by excluding, respectively, four and five influential participants associated with Heywood cases. Similar results were obtained with the full sample (see Supplementary Materials).

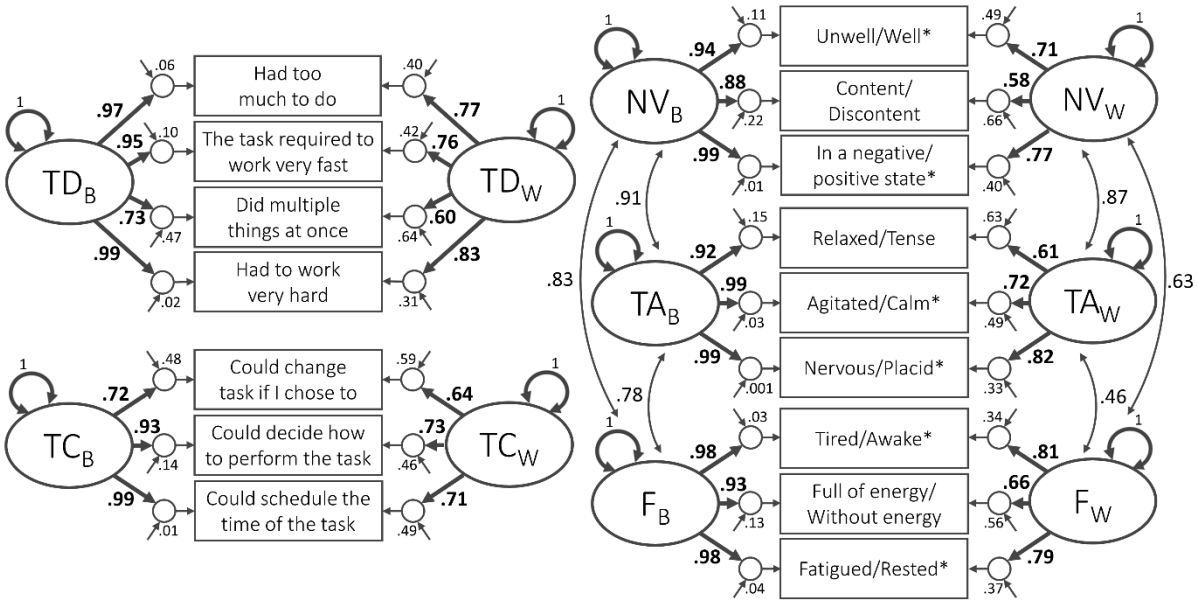


Figure 1. Completely standardized solutions at the between (B) level and within (W) level from the weak cross-level invariance models selected for Task Demand (**TD**), Task Control (**TC**) and mood, respectively. **NV**, Negative Valence; **TA**, Tense Arousal; **F**, Fatigue; *****, Mood items that were reversed prior to analyze the data.

Table 1. Reliability indices of the experience sampling scales.

Measure	ω_{within}	$\omega_{between}$	R_{1F}	R_{KF}	R_C
Task Demand	.83	.95	.79	.99	.83
Task Control	.74	.92	.74	.98	.74
Negative Valence	.73	.96	.70	.98	.65
Tense Arousal	.76	.98	.73	.98	.69
Fatigue	.80	.98	.64	.97	.68

Notes. **ω**, level-specific composite reliability index computed from the selected weak invariance models; R_{1F}, between-person reliability index considering one fixed occasion; R_{KF}, between-person reliability considering the entire set of occasions (i.e., up to 18 for Task Demand and Task Control, up to 21 for Mood); R_C, reliability for detecting differences in systematic changes within-individual over time.

Descriptive statistics and correlations are reported in Table 2. At both levels, TDS and TCS scores were not substantially correlated. At level 2, convergent validity was supported by medium-to-strong correlations between averaged ESM stressor measures and retrospective measures of the corresponding constructs. Level-2 stressor aggregates also showed weak

correlations with retrospective strain in the expected directions, and a medium correlation between TDS and the low-pleasure/high-arousal JAW dimension.

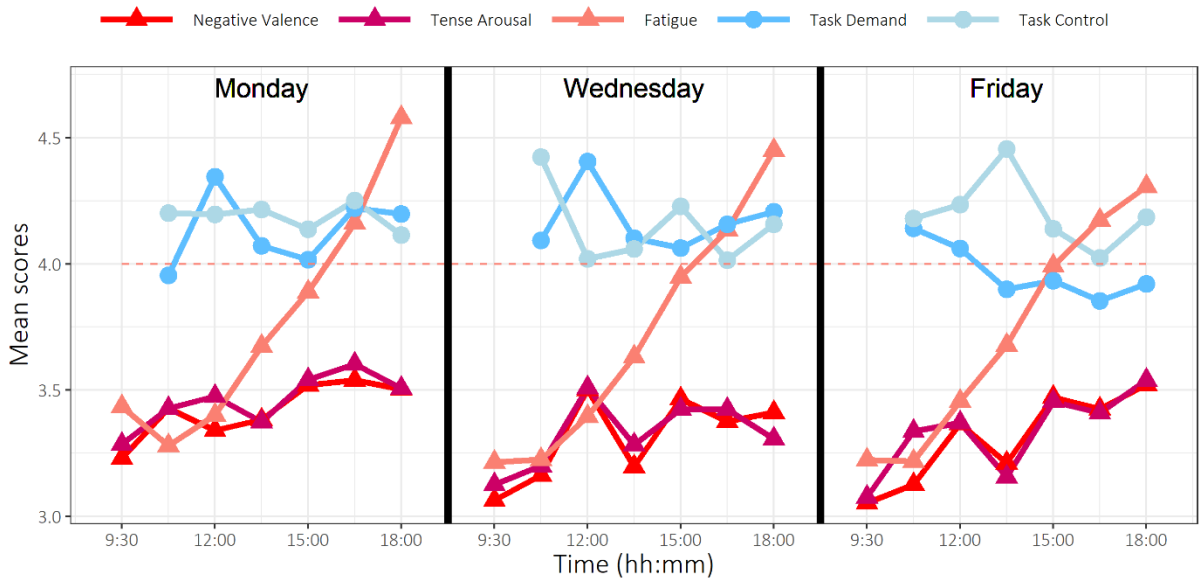
Table 2. Descriptive statistics of, and zero-order Pearson correlations between experience sampling and retrospective measures.

	Measure	No. obs.	Mean (SD)	ICC	Correlations				
					1	2	3	4	5
Experience Sampling (1-7)	1. Task Demand	1,523	4.09 (1.30)	.38		-.12	.12	.27	.01
	2. Task Control	1,491[a]	4.18 (1.49)	.42	-.06		-.22	-.20	-.12
	3. Negative Valence	1,774	3.35 (1.10)	.42	.16	-.36		.64	.49
	4. Tense Arousal	1,774	3.37 (1.19)	.44	.26	-.27	.86		.36
	5. Fatigue	1,774	3.73 (1.22)	.31	.07	-.20	.76	.70	
Retrospective (1-5)	Job Demand	139	3.52 (.83)		.42	-.08	.06	.17	.05
	Job Control	139	3.61 (.71)		.06	.38	-.30	-.21	-.26
	High-pleasure/High-arousal	139	3.05 (.87)		-.02	.23	-.41	-.26	-.27
	High-pleasure/Low-arousal	139	3.18 (.72)		-.14	.14	-.38	-.36	-.28
	Low-pleasure/High-arousal	139	2.47 (.86)		.31	-.20	.37	.42	.28
	Low-pleasure/Low-arousal	139	2.53 (.83)		.20	-.09	.29	.31	.27
	Work-related burnout	139	2.59 (.71)		.20	-.15	.33	.34	.29

In the upper side of the table, level-2 correlations (i.e., between individual mean scores, N = 139) are shown below the diagonal, whereas level-1 correlations (i.e., between mean-centered scores, N = N. obs.) are shown above the diagonal. Shaded cells highlight correlations ≥ .30. **ICC**, intraclass correlation coefficient; **SD**, standard deviation. [a] Discrepancies between the No. of observations between Task Demand and Task control were due to technical problems resulting in 1.31-3.14% of missing responses.

The inspection of the sensitivity to temporal (see Figure 2) and contextual factors did not reveal any linear trend of momentary stressors within and between weekdays (Aw < .15), whereas substantial differences were found across work sampling categories. The type of task predicted substantial differences in TDS and TCS scores (Aw = .99), with "social" tasks (i.e., networking and dissemination, 15.7%) showing lower TDS (*b* = -0.28 (standard error = 0.09) [95% CI = -0.45, -0.11]) and TCS scores (*b* = -0.56 (0.10) [-0.76, -0.35]) than other categories ("information acquisition" was used as reference). Momentary stressors were also sensitive to the mean of work, with "computer" tasks (62.5%) showing higher TDS and TCS scores (Aw = .99) compared to "face-to-face" (TDS: *b* = 0.19 (0.07) [0.05, 0.33]; TCS: *b* = 0.68 (0.08) [0.52,

0.84]) and "others" (TDS: *b* = 0.55 (0.09) [0.38, 0.72]; TCS: *b* = 0.33 (0.10) [0.13, 0.52]), whereas the involvement of other people (44.36%) predicted lower Task Control compared to tasks performed "alone" (Aw = .99; *b* = -0.79 (0.07) [-0.92, -0.65]).



Momentary strains

The three-factor model with weak cross-level invariance was selected based on overall better fit ($\chi^2(57) = 334.91$, RMSEA = .054, CFI = .958, SRMR-within = .033, SRMR-between = .039) than the corresponding configural model (which, however, showed higher Aw = .99 and CFI = .963), and all alternative models (all rejected, including the strong invariance model)[1]. As shown in Figure 1, the selected model showed standardized loadings between .58 and .99, with strong correlations among MDMQ dimensions from .46 (Tense Arousal and Fatigue at level 1) to .91 (Negative Valence and Tense Arousal at level 2). Composite reliability indices suggested

satisfactory reliability at both levels, coherently with variance partitioning, also indicating acceptable sensitivity to change (Table 1).

At level 1, MDMQ scores were only weakly correlated with momentary stressors, while showing substantial intercorrelations at both levels (see Table 2). At level 2, convergent validity was supported by mostly substantial correlations in the expected directions between mood tone and both JAW and Burnout, ranging from |.27| to |.42|. MDMQ subscales were also weakly-to-moderately correlated with retrospective indicators and level-2 ESM aggregates of stressors, with the strongest relationships observed between Negative Valence and both Job Control and level-2 TCS aggregates.

No temporal trends were found across weekdays (Aw < .22), although some differences emerged across days of participation (see Supplementary Materials). As shown in Figure 2, Fatigue increased linearly throughout the workday (Aw = .99; *b* = 0.10 (0.01) [0.08, 0.12]), whereas such a trend was not observed in Negative Valence and Tense Arousal (Aw < .15). Finally, we found higher Negative Valence in "data analysis/authoring" (24.1%) compared to "information acquisition" tasks (28.75%) (Aw = .93; *b* = 0.24 (0.07) [0.10 0.38]), whereas no substantial differences were observed in terms of means of work and persons involved (Aw < .43).

Overall, we obtained similar results considering two alternative subsamples by using more (i.e., 90 participants with at least three ESM responses per day) or less restrictive criteria (i.e., 175 participants with at least one response in total) (see Supplementary Materials).

Discussion

This study aimed at developing and validating a set of ESM measures of workplace stress to be used in both research and routine assessment. The described set of 16 items was

identified as an ideal compromise between the need of parsimony (requiring less than five minutes to respond) and that of reliably quantifying theoretically and practically relevant variables (Beal, 2015), including widely investigated task characteristics (Task Demand and Task Control), and core dimensions of affective strain (Negative Valence, Tense Arousal, and Fatigue).

Our results suggested satisfactory construct validity (hypothesis H1) and reliability (H2) at both momentary and individual level, with MCFAs corroborating the hypothesized multilevel models (H1.1). Importantly, the satisfactory fit showed by weak invariance models (H1.2) provides initial support to their ability of reflecting configural cluster constructs (Stapleton et al., 2016), also implying weak measurement invariance across respondents (Jak & Jorgensen, 2017). Moreover, the proposed scales showed satisfactory sensitivity to systematic changes within participants over time (see Shrout & Lane, 2012).

Our study confirmed the pattern of results reported for the original MDMQ (Wilhelm & Schoebi, 2007), with Negative Valence and Tense Arousal being highly intercorrelated, and almost indistinguishable at level 2. Whereas the correlations estimated among mood dimensions were very high in general, the strong relationship between Negative Valence and Tense Arousal questions their conceptualization as different constructs. Nevertheless, alternative models with the corresponding items being saturated in the same dimension showed poor fit, and we found differentiated patterns of level-2 correlations and sensitivity to contextual factors. Possible explanations of low discriminant validity might rely, for instance, on a magnification of the common method variance due to the MDMQ items order (i.e., each Tense Arousal item was preceded by a Negative Valence item), in addition to overlaps in the item content between the two scales. More studies are needed to clarify the conceptual distinction between MDMQ dimensions, and the potential reasons for the overall higher

correlations found in our study compared to Wilhelm & Schoebi (2007), such as the different sampling protocol, the introduction of three additional items, the potential changes in the latent variables due to item translation, and the homogeneity of the response setting (workplace).

Convergent validity (H3) was also supported, with substantial correlations in the expected directions at both levels. Fatigue showed the lowest correlations with JAW, possibly due to the different dimensionality of the retrospective scale (i.e., "fatigued": low-pleasure/low-arousal, "energetic": high-pleasure/high-arousal) (see Van Katwyk et al., 2000), and the lack of specific criterion variables for Fatigue. Whereas some evidence of criterion validity for this variable is provided by its increasing linear trend observed throughout the workday (see also Wilhelm & Schoebi, 2007). Overall, in terms of stressor-strain relationships, our results were coherent with previous studies showing weak-to-moderate correlations at both levels (Pindek et al., 2019).

Finally, some of the scales (i.e., TDS, TCS, and Negative Valence) showed sensitivity to contextual factors, including the type and mean of job task, and the people involved. The availability of scales sensitive to meaningful task categories would be useful for both organizational scholars (e.g., stress-based task taxonomies) and practitioners (e.g., tailor-made job redesign accounting for context-specific work sampling).

The main limitations of this study include the lack of objective (e.g., psychophysiological) criterion variables and the limited number of days, which were not considered as a separate level (as done by Wilhelm & Schoebi, 2007). Moreover, response rate was relatively low (61%), possibly due to the lack of face-to-face interactions with participants (data collection was entirely automatized), technical problems, and lack of monetary incentives (see Gabriel et al.,

2019). Although results were consistent across three subsamples with different response rates, such a loss of information might have affected our conclusions.

Notwithstanding the above limitations, our study provides a parsimonious set of psychometrically sounding measures to be used for the investigation and the routine assessment of workplace stress, accompanying them with an exhaustive range of information for future users. Given the increasing acknowledgment of ESM as preferential tools to assess dynamic phenomena such as workplace stress, it is hoped that this article and the attached materials will contribute to the advancement of workplace stress assessment.

References

Avanzi, L., Balducci, C., & Fraccaroli, F. (2013). Contributo alla validazione italiana del Copenhagen Burnout Inventory (CBI). *Psicologia Della Salute*, *2*, 120–135. https://doi.org/10.3280/PDS2013-002008

Balducci, C., Fraccaroli, F., & Schaufeli, W. B. (2010). Psychometric Properties of the Italian Version of the Utrecht Work Engagement Scale (UWES-9). *European Journal of Psychological Assessment*, *26*(2), 143–149. https://doi.org/10.1027/1015-5759/a000020

Barbaranelli, C., Fida, R., & Gualandri, M. (2013). Assessing counterproductive work behavior: A study on the dimensionality of cwb-checklist. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *20*(3), 235–248. https://doi.org/10.4473/TPM20.3.3

Beal, D. J. (2015). ESM 2.0: State of the Art and Future Potential of Experience Sampling Methods in Organizational Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Bowling, N. A., Alarcon, G. M., Bragg, C. B., & Hartman, M. J. (2015). A meta-analytic

examination of the potential correlates and consequences of workload. *Work and Stress*, *29*(2), 95–113. https://doi.org/10.1080/02678373.2015.1033037

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior*, *33*(7), 865–877. https://doi.org/10.1002/job.1803

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hurrell, J. J., Nelson, D. L., & Simmons, B. L. (1998). Measuring job stressors and strains: Where we have been, where we are, and where we need to go. *Journal of Occupational Health Psychology*, *3*(4), 368–389. https://doi.org/10.1037/1076-8998.3.4.368

Jak, S., & Jorgensen, T. D. (2017). Relating Measurement Invariance, Cross-Level Invariance,

and Multilevel Reliability. *Frontiers in Psychology*, *8*(OCT), 1–9. https://doi.org/10.3389/fpsyg.2017.01640

Kamarck, T., Janicki, D., Shiggman, S., Polk, D., Muldon, M., Libenauer, L., & Schwartz, J. (2002). Psychosocial demands and ambulatory blood pressure: a field assessment approach. *Physiology & Behavior*, *77*(4–5), 699–704. https://doi.org/10.1016/S0031-9384(02)00921-6

Karasek, R., Brisson, C., Kawakami, N., Houtman, I., Bongers, P., & Amick, B. (1998). The Job Content Questionnaire (JCQ): An instrument for internationally comparative assessments of psychosocial job characteristics. *Journal of Occupational Health Psychology*, *3*(4), 322–355. https://doi.org/10.1037/1076-8998.3.4.322

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel Factor Analysis: Reporting Guidelines and a Review of Reporting Practices. *Multivariate Behavioral Research*, *51*(6), 0–0. https://doi.org/10.1080/00273171.2016.1228042

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary Studies in Organizational Research. *Journal of Personnel Psychology*, *9*(2), 79–93. https://doi.org/10.1027/1866-5888/a000009

Pindek, S., Arvan, M. L., & Spector, P. E. (2019). The stressor–strain relationship in diary studies: A meta-analysis of the within and between levels. *Work and Stress*, *33*(1), 1–21. https://doi.org/10.1080/02678373.2018.1445672

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/

Robinson, M. A. (2009). Work sampling: Methodological advances and new applications. *Human Factors and Ergonomics in Manufacturing*, *20*(1), n/a-n/a.

https://doi.org/10.1002/hfm.20186

Semmer, N. K., Zapf, D., & Dunckel, H. (1995). Assessing stress at work: A framework and an instrument. In O. Svane & C. Johansen (Eds.), *Work and Health - Scientific basis of progress in the working environment* (pp. 105–113). Office for Official Publications of the European Communities.

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. S. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Sudying Daily Life* (pp. 302–320). The Guilford Press.

Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job stressors and strain: Interpersonal Conflict at Work Scale, Organizational Constraints Scale, Quantitative Workload Inventory, and Physical Symptoms Inventory. *Journal of Occupational Health Psychology*, *3*(4), 356–367. https://doi.org/10.1037/1076-8998.3.4.356

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct Meaning in Multilevel Settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://doi.org/10.3102/1076998616646200

Tabanelli, M. C., Depolo, M., Cooke, R. M. T., Sarchielli, G., Bonfiglioli, R., Mattioli, S., & Violante, F. S. (2008). Available instruments for measurement of psychosocial factors in the work environment. *International Archives of Occupational and Environmental Health*, *82*(1), 1–12. https://doi.org/10.1007/s00420-008-0312-6

Thorsen, S. V., & Bjorner, J. B. (2010). Reliability of the Copenhagen Psychosocial Questionnaire. *Scandinavian Journal of Public Health*, *38*(3_suppl), 25–32. https://doi.org/10.1177/1403494809349859

Toderi, S., Balducci, C., Edwards, J. A., Sarchielli, G., Broccoli, M., & Mancini, G. (2013). Psychometric Properties of the UK and Italian Versions of the HSE Stress Indicator Tool.

European Journal of Psychological Assessment, *29*(1), 72–79. https://doi.org/10.1027/1015-5759/a000122

Van Katwyk, P. T., Fox, S., Spector, P. E., & Kelloway, E. K. (2000). Using the Job-Related Affective Well-Being Scale (JAWS) to investigate affective responses to work stressors. *Journal of Occupational Health Psychology*, *5*(2), 219–230. https://doi.org/10.1037/1076-8998.5.2.219

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

Warr, P. B. (1994). A conceptual framework for the study of work and mental health. *Work & Stress*, *8*(2), 84–97. https://doi.org/10.1080/02678379408259982

Wetherell, M. A., & Carter, K. (2014). The Multitasking Framework: The Effects of Increasing Workload on Acute Psychobiological Stress Reactivity. *Stress and Health*, *30*(2), 103–109. https://doi.org/10.1002/smi.2496

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

Xiong, H., Huang, Y., Barnes, L. E., & Gerber, M. S. (2016). Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 415–426. https://doi.org/10.1145/2971648.2971711

Open data

The raw data, the data analysis pipeline and code, including the procedures used for a priori power analysis, and all the supplementary materials relevant for the present article are available from the following public repository: https://doi.org/10.5281/zenodo.6489666

We report how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. When using inferential tests, we reported exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including the codebook.

Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology.

Preregistration of Studies and Analysis Plans: This study was not preregistered.