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 S4.1: Details on MCFA models

The supplemental material S4.1 includes details on the analytical strategy used for, and the results obtained from, the Multilevel Confirmatory Factor Analyses (MCFAs) conducted on the Task Demand Scale (TDS), the Task Control Scale (TCS), and the Italian adaptation of the Multidimensional Mood Questionnaire (MDMQ).

1. Construct conceptualization and measurement models

All latent variables were conceptualized as configural cluster constructs (see Stapleton et al., 2016). At level 1 (within-individual), Task Demand and Task Control were defined, respectively, as the amount of work/difficulty in, and the organizationally mediated possibilities to make decisions about, a specific job task. At level 2 (between individuals), Job Demand and Job Control were conceptualized as the individual-level aggregates of level-1 constructs, that is "the amount or difficulty of one's work" (i.e., the Workload component of Job Demand; see Bowling et al., 2015) and "the organizationally mediated possibilities for workers to make decisions about their work" (Karasek et al., 1998).

Similarly, the MDMQ dimensions Negative Valence, Tense Arousal, and Fatigue (i.e., labeled to express negative mood states), were conceptualized as multilevel configural constructs whose level-2 (e.g., individual Fatigue levels) components should simply reflect an aggregated form of the level-1 (e.g., momentary Fatigue levels within-individual), as proposed by Wilhelm & Schoebi (2007), and supported by following studies conducted with working populations (e.g., Dettmers et al., 2016). Coherently, a single-factor configural structure was hypothesized for both TDS and TCS items, whereas a three-factor configural structure was expected for the MDMQ, as represented in Figure S4.1A.

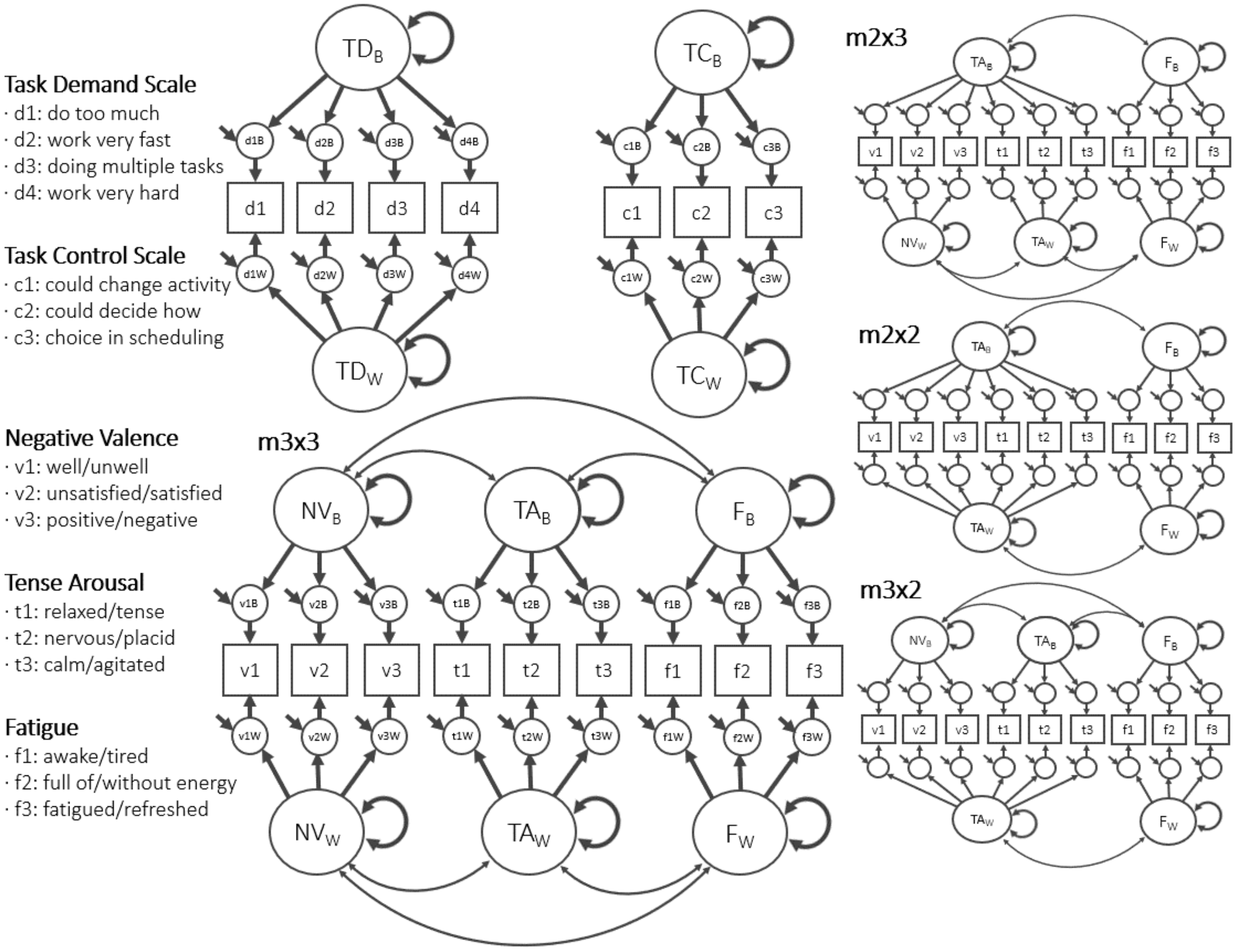


Figure S4.1A. Path diagram of the hypothesized factor models specified for Task Demand (TD), Task Control (TC), and Mood. The right side of the figure shows alternative models specified for Mood, in which we varied the number of latent factors at both levels. B, between; W, within; NV, Negative Valence; TA, Tense Arousal; F, Fatigue. Item wording is shown on the left side (see Supplementary Materials S3 for the exact wording, order, and Italian translations).

According to the results from Wilhelm & Schoebi (2007), we expected a strong correlation between Negative Valence and Tense Arousal at both levels, and especially at level 2, where the two dimensions could not be distinguished by the factor analysis conducted by the authors. In contrast, Dettmers et al. (2016) provided empirical support for a three-factor model at both levels, showing a better fit than a one-factor model, and a model assuming three factors with an higher-order factor. Consequently, as shown in Figure S4.1A, we compared the hypothesized model (m3×3) with three alternative models assuming two factors (i.e., Fatigue, and a Tense Arousal factor reflecting both Negative Valence and Tense Arousal item scores, in line with Thayer, 1990) either at level 2 (m2×3), at level 1 (m3×2) or both (m2×2).

2. Analytical strategy

A separate set of MCFA models was specified for each scale, following the guidelines provided by Kim et al. (2016), and using the *lavaan* R package (version 0.6.6; Rosseel, 2012). For each scale, a multilevel model was specified as shown in Figure S4.1A.

The analytic procedure was based on Hox (2010), according to which we specified a set of preliminary models to evaluate the factor structure at both levels. At level 1, a conventional one-level CFA was performed on the pooled within-cluster covariance matrix (S_{PW}) (see Muthén, 1994). At level 2, benchmark models were specified by imposing: (1) no specification (i.e., null model, implying no between-cluster structure at all), (2) only variances but no covariances (i.e., independence model, implying between-clusters variance but no substantively interesting structural model), and (3) a saturated model (implying that the construct only “exists” at the within level, whereas level 2 variation is just “spurious”) (Hox, 2010). We expected an acceptable fit for the conventional CFA on S_{PW}, whereas a poor fit was expected for the three benchmark models specified to evaluate the factor structure at level 2.

To validate our conceptualization of Task Demand, Task Control, and Mood as configural cluster constructs (Stapleton et al., 2016; Tay et al., 2014), and to evaluate measurement invariance across clusters (i.e., participants) (see Jak & Jorgensen, 2017), we tested cross-level isomorphism by fitting three increasingly restrictive models on the item scores of each scale, following Jak & Jorgensen (2017): (1) a configural invariance model, with the same factor structure but different loadings across levels (no constraints were imposed); (2) a weak invariance model, with factor loadings constrained to be equal across levels; and (3) a strong invariance model, with residual variance at level 2 constrained to zero (assuming both factor loadings and intercepts as invariant across clusters). Only models 2 and 3 are compatible with configural constructs (Stapleton et al., 2016; Tay et al., 2014).

The specified multilevel models, all fitted by standardizing the factor covariance matrix to avoid fixing the first indicator loadings, were compared by considering both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion weight (Aw) (Wagenmakers & Farrell, 2004), in addition to the root mean square error of approximation (RMSEA), the comparative fit index (CFI) and the standardized root mean squared residual (SRMR). As recommended by Hsu et al. (2015), the SRMR was separately computed for the within-individual (SRMR-W) and the between-participants measurement model (SRMR-B), since global fit indices are dominated by the fit information at level 1 (due to higher data numerosity on this level), and are likely to be unsensitive to misspecifications at level 2 (see also Ryu & West, 2009). According to the criteria proposed by (Hu & Bentler, 1999), we considered RMSEA ≤ .06, CFI ≥ .95, and SRMR ≤ .08 as indicative of adequate fit.

3. Results

All TDS, TCS and MDMQ items showed roughly normally distributed scores (see Figure S4.1B), with missing responses ranging from 0.00-1.31% (i.e., N = 0-20) for MDMQ and TDS items (presented at the beginning of ESM questionnaire) to 2.49-3.14% (i.e., N = 42-48) for TCS items (presented at the end of ESM questionnaire)[1]. Item scores ICCs ranged from .23 to .40, suggesting that most variance was at the within level, but level-2 variance was still substantial to justify a multilevel approach. Items t2 and t3 measuring Tense Arousal showed the highest level-2 variance (ICC = .39 and .40, respectively), whereas items f2 and f3 measuring Fatigue showed the highest variance at level 1 (ICC = .23 and .27, respectively).

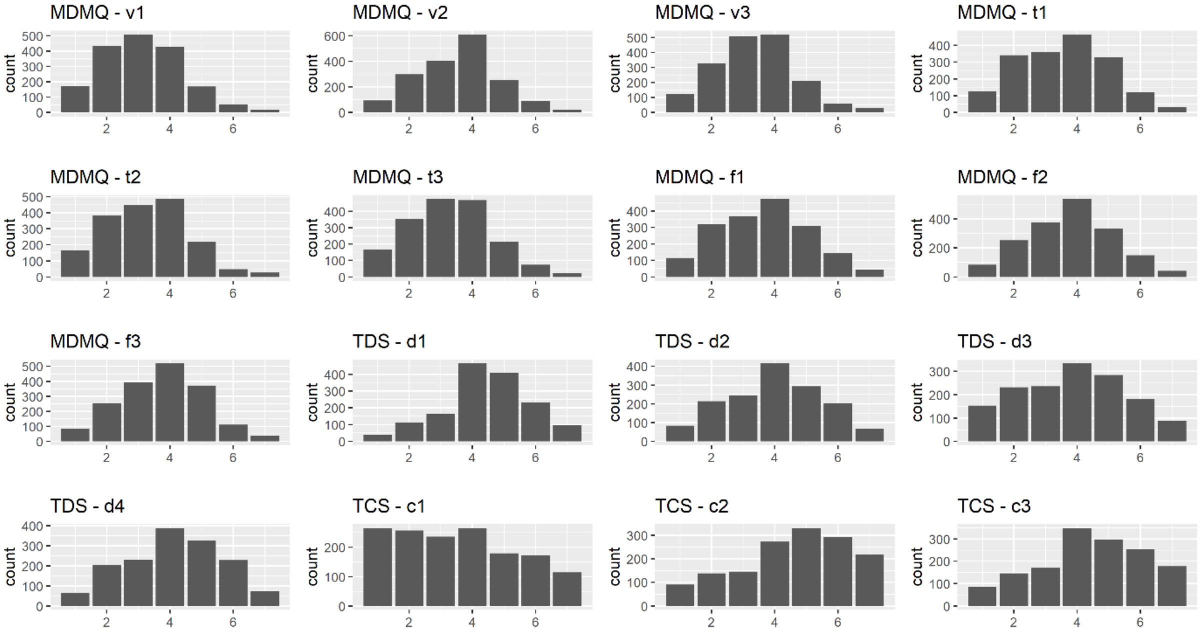


Figure S4.1B. Frequency distribution of ESM item scores.

[1] Whereas 'full-missing' observation (i.e., no response to any of the items) were more likely due to lack of participants' compliance, responses with missing data in one-to-14 items (i.e., N = 48, 3.14%) were caused by technical problems with the mobile app used for data collection, to be considered as missing-at-random.

Zero-order Pearson correlations between items scores were in the expected directions (see Figure S4.1C), with groups of scores associated with the same latent factor being highly inter-correlated (Pearson's *r* ranging from .47 to .76), and showing lower correlations with scores associated with different factors. TDS and TCS scores were more clearly distinguishable than Mood dimensions, with some correlations between items scores associated with Negative Valence and Tense Arousal being higher than those between scores assumed to reflect the same dimension. The latter also showed negative correlations with TCS item scores, and positive correlations with TDS scores, whereas Fatigue item scores showed weaker correlations with TDS items. Correlations between mean and mean-centered scores were in the same directions, but showed, respectively, higher and lower values than those computed from the raw scores.

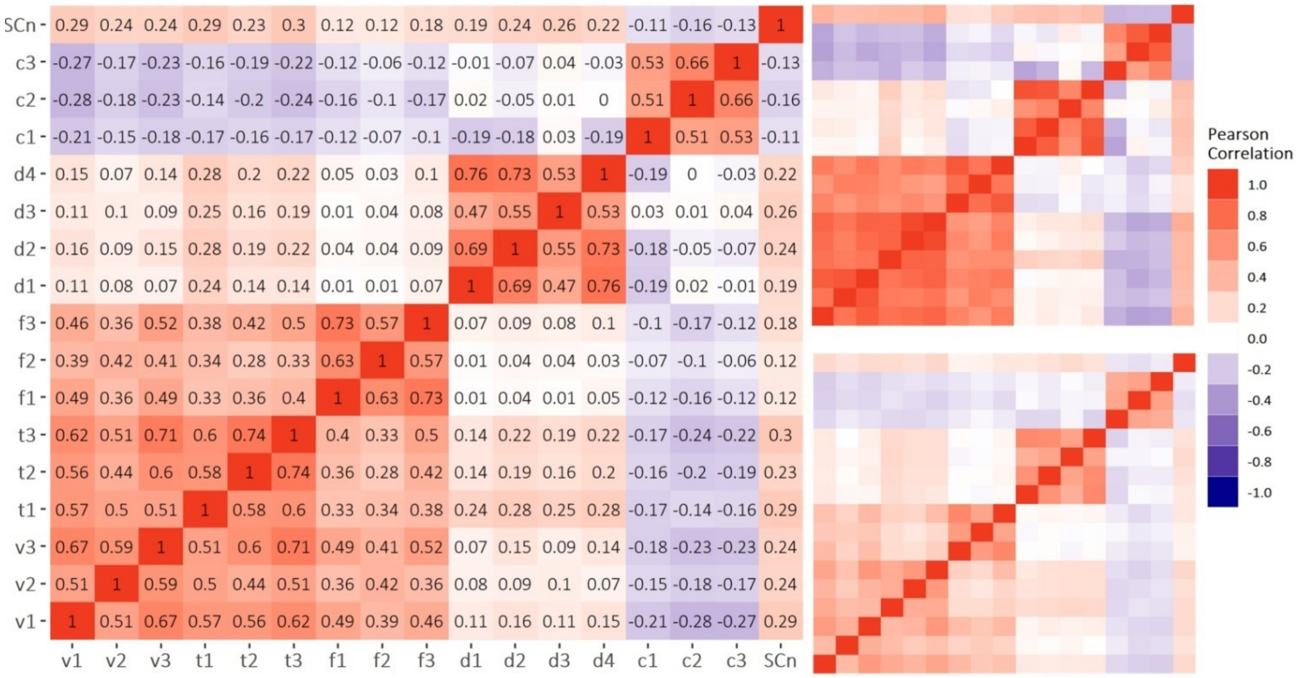


Figure S4.1C. Correlation matrix of raw item scores (on the left, i.e., considering all observations as independent), average scores (top-right, i.e., considering one mean value for each participant and item), and mean-centered scores (bottom-right, i.e., considering deviations from individual means). v1, v2, v3 = items measuring Hedonic Tone; t1, t2, t3 = items measuring Tense Arousal; f1, f2, f3 = items measuring Fatigue; d1, d2, d3, d4 = items measuring Task Demand; c1, c2, c3 = items measuring Task Control; SCn = number of reported Situational Constraints (not considered in the present work).

Preliminary models supported the existence of both within- (the conventional CFA specified on the pooled within-participant covariance matrix showed approximately acceptable fit, although RMSEA was .071 for Mood) and between-participants factorial structure (all benchmark models were rejected).

Four multilevel models showed improper solutions at level 2 (i.e., a negative variance was estimated for item c3 by the configural model of the TCS, and for item t3 measuring Tense Arousal in model m2×3, and both the configural and the weak invariance models m3×3). As we excluded problems of nonconvergence (all models converged), missing data (< 3.14% for each considered item), empirical underidentification (all loadings were > .60 at level 2, with correlations between MDMQ dimensions ranging from .58 to .91), and structural misspecification (all 95% CI around the negative variance estimates included zero), we imputed the improper solutions to sampling fluctuation, and we excluded participants associated with the lowest (negative) variance estimates (strategy A: sensitivity analysis). Participants were excluded one by one, until a positive variance was estimated. This procedure led to the exclusion of four participants (2.88%) to solve the problem in the configural model for TCS items, five participants (3.60%) to solve it in the configural and the MDMQ weak invariance models m3×3, and seven participants (5.04%) to solve it also in m2×3. As an alternative strategy (strategy B: fixed residual variance), we fixed the residual variance of the problematic items to the 15% of their total variance.

The model comparisons obtained for the three ESM scales under different constraints and using different subsamples of participants are reported in Tables S4.1A-C. In all comparisons, the hyphothesized configural model and the corresponding weak invariance model showed acceptable fit indices, with the only exception of the configural model of TCS items (RMSEA > .060). Since the configural model of the TCS with freely estimated parameters was saturated, χ^2-derived fit indices could not be evaluated. Thus, although the configural model showed higher Aw than the weak invariance model, the latter was preferred. Alternative models m2×3, m3×2 and m2×2 showed unacceptable fit across the four model comparisons for MDMQ scores, and were rejected, with the exception of m3×2 when we fixed the residual covariance of item t3 at level 2. Overall, the weak invariance models were selected as the best models, showing the lowest RMSEA and BIC in all comparisons, and the highest Aw in all TDS- and TCS-related model comparisons.

Table S4.1A. Model comparison and fit indices for the Task Demand Scale.

Model	n. par.	χ^2 (df)	RMSEA	CFI	SRMR-W	SRMR-B	AICw	BIC
Weak invariance	**16**	**32.33 (8)**	**.045**	**.991**	**.016**	**.061**	**.685**	**18223.89**
Configural	20	25.89 (4)	.060	.992	.013	.037	.315	18246.70
Strong invariance	12	462.55 (12)	.158	.829	.062	.218	.000	18624.83

n. par., number of estimated parameters, **df**, degrees of freedom associated with the χ^2 statistic; **RMSEA**, root mean square error of approximation; **CFI**, comparative fit index; **SRMR-W**, root mean squared residual within subject; **SRMR-B**, SRMR between subjects; **Aw**, Akaike Information Criterion weight; **BIC**, Bayesian Information Criterion; **TD**, Task Demand; **TC**, Task Control. Bold types indicate the selected model for each comparison.

Table S4.1B. Model comparisons and fit indices for the Task Control Scale.

	Model	n. par.	χ^2 (df)	RMSEA	CFI	SRMR-W	SRMR-B	Aw	BIC
Heywood	Configural (HC)[a]	15						.863	15353.27
	Weak invariance	**12**	**9.68 (3)**	**.039**	**.994**	**.010**	**.053**	**.137**	**15341.05**
	Strong invariance	9	443.61 (6)	.222	.616	.071	.232	.000	15753.09
Strategy A	Configural[a]	15						.315	14790.27
	Weak invariance	**12**	**5.38 (3)**	**.023**	**.998**	**.008**	**.035**	**.685**	**14773.84**
	Strong invariance	9	317.79 (6)	.190	.741	.058	.186	.000	15064.44
Strategy B	Configural	14	12.36 (1)	.088	.990	.002	.044	.034	15358.33
	Weak invariance	**12**	**9.68 (3)**	**.039**	**.994**	**.010**	**.053**	**.966**	**15341.05**
	Strong invariance	9	443.61 (6)	.222	.616	.071	.232	.000	15753.09

See the notes in Table S4.1A. **HC**, Heywood case; [a], The model was saturated, and the fit indices could not be evaluated. The table shows the model comparison conducted on the full sample (showing HC for item c3 on level 2 in the Configural model), and by using strategy A (i.e., excluding four participants based on sensitivity analysis: N = 135) or B (i.e., constraining the item c3 residual variance at level 2 to the 15% of its total variance at level 2).

Table S4.1C. Model comparison and fit indices for the Multidimensional Mood Questionnaire.

	Model	n. par.	χ^2 (df)	RMSEA	CFI	SRMR-W	SRMR-B	Aw	BIC
Heywood cases (N = 139)	m2×2	47	566.73 (52)	.075	.926	.045	.067	.000	44679.56
	m2×3 (HC)	49	474.59 (50)	.069	.939	.045	.052	.000	44602.37
	m3×2	49	418.75 (50)	.065	.947	.033	.058	.000	44546.53
	m3×3 CI (HC)	51	352.13 (48)	.060	.956	.033	.050	.999	44494.87
	m3×3 WI (HC)	**42**	**403.56 (57)**	**.059**	**.95**	**.036**	**.048**	**.000**	**44479.01**
	m3×3 SI	33	754.19 (66)	.077	.901	.044	.089	.000	44762.35
Strategy A (N = 134)	m2×2	47	502.81 (52)	.071	.933	.044	.053	.000	42513.49
	m3×2	49	355.20 (50)	.060	.955	.031	.048	.000	42380.75
	m3×3 CI	51	296.06 (48)	.055	.963	.031	.042	.999	42336.49
	m3×3 WI	**42**	**337.52 (57)**	**.054**	**.959**	**.034**	**.039**	**.001**	**42311.02**
	m3×3 SI	33	603.63 (66)	.069	.921	.039	.077	.000	42510.21
Strategy A (N = 132)	m2×2	47	484.63 (52)	.071	.935	.043	.052	.000	41750.22
	m2×3	49	408.37 (50)	.066	.946	.043	.043	.000	41688.80
	m3×2	49	348.00 (50)	.060	.955	.031	.047	.000	41628.43
	m3×3 CI	51	294.29 (48)	.055	.963	.030	.042	.999	41589.56
	m3×3 WI	**42**	**334.91 (57)**	**.054**	**.958**	**.033**	**.039**	**.001**	**41563.40**
	m3×3 SI	33	600.99 (66)	.070	.920	.039	.076	.000	41762.70
Strategy B (N = 139)	m2×2	47	566.73 (52)	.075	.926	.045	.067	.000	44679.56
	m2×3	49	367.26 (50)	.060	.954	.033	.045	.100	44495.05
	m3×2	49	418.75 (50)	.065	.947	.033	.058	.000	44546.53
	m3×3	51	359.09 (48)	.061	.955	.033	.044	.890	44501.82
	m3×3 WI	**42**	**398.47 (57)**	**.058**	**.951**	**.035**	**.046**	**.001**	**44473.92**
	m3×3 SI	33	754.19 (66)	.077	.901	.044	.089	.000	44762.35

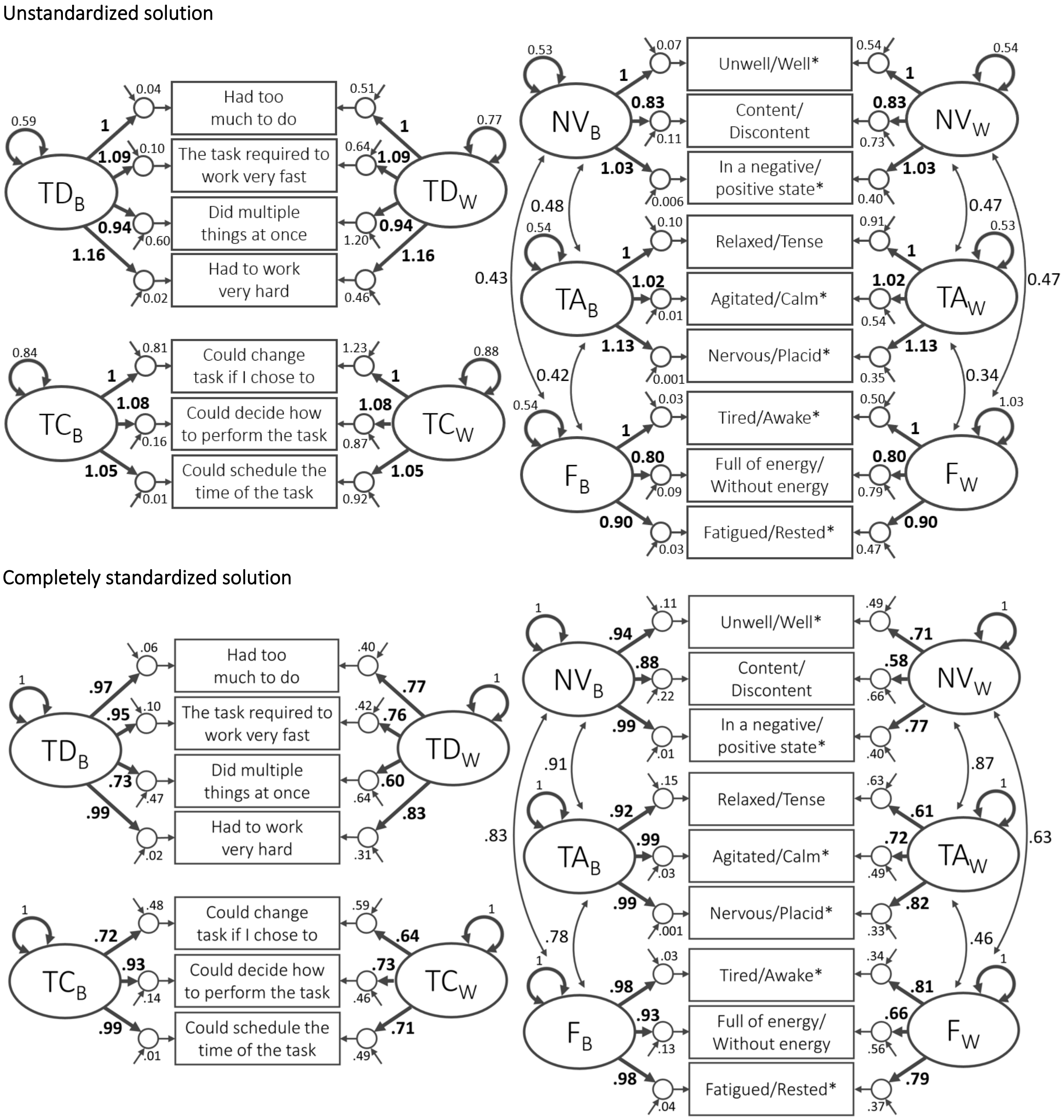
See the notes in Table S4.1A. **CI**, Configural invariance; **WI**, Weak invariance; **SI**, Strong invariance; **HC**, Heywood case. The table shows the model comparison conducted on the full sample (showing HC for item t3 on level 2 in m2×3, and the m3×3 CI and WI models), and by using either strategy A (i.e., sensitivity analysis) to solve the problem in models m3×3 (by excluding five participants, as reported in the main manuscript) and in all models (by excluding seven participants) or Strategy B (i.e., constraining the item t3 residual variance at level 2 to the 15% of its total variance at level 2).

In contrast, the model assuming strong factor invariance was rejected due to unsatisfactory fit across all model comparisons.

Figure S4.1D shows the unstandardized and the completely standardized solution for each scale. Standardized loadings estimated with weak configural models were all significant and ranged from .58 to .99, with estimated correlations between MDMQ dimensions ranging from .46 to .91.

4. Short discussion

The weak invariance model was selected based on satisfactory and overall better fit for all ESM scales, providing initial support to their ability of reflecting multilevel configural cluster constructs (Stapleton et al., 2016). This result also implies that weak measurement invariance holds across clusters (i.e., respondents), although strong invariance models were rejected, suggesting the presence of other factors than the hypothesized dimensions influencing item scores at level 2 (Jak & Jorgensen, 2017). Standardized loadings indicated stronger factor structure at level 2 than at level 1, a typical situation due to measurement error accumulating at the lower level (Hox, 2010). Coherently, reliability coefficients were higher for level 2, but adequate at both levels (see the main manuscript).


Unstandardized solution

Completely standardized solution

Figure S4.1D. Unstandardized and completely standardized parameter estimates at the between (B) and within (W) level from the selected two-level models with weak cross-level invariance. In the unstandardized solution, the first loading of each latent variable is fixed to 1 to freely estimate variances and covariances, whereas in the completely standardized solution latent variables are standardized to freely estimate all factor loadings, showing the correlations between MDMQ subfactors. **TD**, Task Demand; **TC**; Task Control; **NV**, Negative Valence; **TA**, Tense Arousal; **F**, Fatigue; *****, MDMQ items that were reversed prior to analyze the data.

5. References

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

Dettmers, J., Vahle-Hinz, T., Bamberg, E., Friedrich, N., & Keller, M. (2016). Extended work availability and its relation with start-of-day mood and cortisol. *Journal of Occupational Health Psychology*, *21*(1), 105–118. https://doi.org/10.1037/a0039602

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). Routledge.

Hsu, H. Y., Kwok, O. man, Lin, J. H., & Acosta, S. (2015). Detecting Misspecified Multilevel Structural Equation Models with Common Fit Indices: A Monte Carlo Study. *Multivariate Behavioral Research*, *50*(2), 197–215. https://doi.org/10.1080/00273171.2014.977429

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

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

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

Muthén, B. O. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

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

Ryu, E., & West, S. G. (2009). Level-Specific Evaluation of Model Fit in Multilevel Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 583–601. https://doi.org/10.1080/10705510903203466

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

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A Conceptual and Methodological Framework for Psychometric Isomorphism. *Organizational Research Methods*, *17*(1), 77–106. https://doi.org/10.1177/1094428113517008

Thayer, R. E. (1990). *The Biopsychology of Mood and Arousal*. Oxford University Press.

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

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