Affective structure, measurement invariance, and reliability across different experience sampling protocols

Journal of Research in Personality 92 (2021) 104094

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Affective structure, measurement invariance, and reliability across different experience sampling protocols

Gudrun Eisele [a,*], Ginette Lafit [a,b], Hugo Vachon [a], Peter Kuppens [b], Marlies Houben [b], Inez Myin-Germeys [a], Wolfgang Viechtbauer [a,c]

[a] Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium
[b] Research Group for Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium
[c] Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, the Netherlands

ARTICLE INFO

Article history:
Received 19 November 2020
Revised 22 February 2021
Accepted 22 March 2021
Available online 26 March 2021

Keywords:
Mood
Positive and negative affect
Ecological momentary assessment
Measurement invariance
Structural validity

ABSTRACT

While affect is frequently measured with experience sampling methodology (ESM), the affective structure at the between- and within-person level has not been thoroughly investigated. We investigated the affective structure at the between- and within-person level, its invariance across different ESM protocols, and its reliability. Participants (N = 147) were randomly assigned to receive either a 30 or 60 item questionnaire three, six, or nine times per day, resulting in 72–75 participants per questionnaire length and 48–50 participants per sampling frequency. Momentary affect was assessed with 8 or 18 items. At both levels, a structure with two correlated factors showed the best fit compared to an orthogonal and a unidimensional model. A structure with additional freed residual correlations was invariant across protocols at the within-person level and showed high reliability. We observed indications of a more discrete affective structure within than between persons.

1. Introduction

[1]According to dimensional theories of affect, affective experiences vary along broad dimensions. The number and nature of these dimensions have been the subject of a long and ongoing debate (e.g., Schlosberg, 1952; Jacobson, Evey, Wright, & Newman, 2020). Although some researchers have suggested more affective dimensions (e.g., Schimmack & Grob, 2000), a large part of the discussion has centered on two-dimensional models. The circumplex model of affect (Russell, 1980) entails that affective states vary along the two dimensions valence, ranging from unpleasant to pleasant, and arousal, ranging from low (e.g., tired) to high (e.g., excited or alert). However, Watson and Tellegen (1985) have argued that variations in affective states across and within persons are best represented by a two-dimensional structure with the dimensions positive and negative affect (PA and NA, respectively) that map onto the space of the circumplex model of affect at a 45 degree angle (Watson, Wiese, Vaidya, & Tellegen, 1999). This means that Watson and Tellegen's PA dimension ranges from positive, high arousal to negative, low arousal states on the circumplex model of affect, while their NA dimension ranges from negative, high arousal to positive, low arousal states. Since its development, the PANAS (Watson, Clark, & Tellegen, 1988), an instrument designed to measure PA and NA, has been used extensively in applied research (over 37 000 citations on google scholar at the time of writing). Originally, PA and NA were expected to be largely independent of each other (Watson & Tellegen, 1985). However, research has indicated that the correlation between PA and NA varies for instance as a function of situation (Dejonckheere et al., 2019), and questionnaire characteristics, such as the set of items used to assess them (Russell & Carroll, 1999) and the time frame of the instructions (Diener & Emmons, 1984).

An additional factor that is expected to influence the independence of PA and NA is the level of analysis (Bleidorn & Peters, 2011; Brose, Voelkle, Lövdén, Lindenberger, & Schmiedek, 2015; Vansteelandt, Van Mechelen, & Nezlek, 2005; Zelenski & Larsen, 2000). In cross-sectional assessments, either trait affect or momentary affect is measured at one time point, both of which cannot

* Corresponding author at: Department of Neurosciences, KU Leuven/Research Group Psychiatry/Center for Contextual Psychiatry, Kapucijnenvoer 33 bus 7001 (blok h), 3000 Leuven, Belgium.
E-mail address: gudrunvera.eisele@kuleuven.be (G. Eisele).

[1] The reported confirmatory and exploratory analyses included in this paper were preregistered (https://osf.io/6t5ud) and deviations from the preregistration are indicated in the text. All materials are available on the OSF page of the project (https://osf.io/pzx8r/) and the analysis code is provided in the supplemental materials. Participants of this study did not agree for their data to be shared publicly, so supporting data is not available. However, the data will be shared with interested researchers upon request.

https://doi.org/10.1016/j.jrp.2021.104094
0092-6566/

disentangle within- and between-person (WP and BP, respectively) influences (Hamaker, Nesselroade, & Molenaar, 2007). In studies using experience sampling methodology (ESM) or diaries, repeated measurements are nested within individuals. While daily diaries are limited to one assessment per day, ESM studies typically include multiple assessments of affect per day. Both methods allow studying variability in the way that different people feel on average (at the BP level) and variability of the deviations from these personal means (at the WP level). Although positive and negative affect are expected to emerge at both levels (Watson & Tellegen, 1985), most support for the PA and NA structure has been obtained in analyses at the BP level. At this level, the broad dimensions of positive and negative affect have been found to relate in meaningful ways to other constructs, such as personality and clinical diagnoses (Brose, Schmiedek, Gerstorf, & Voelkle, 2020; Jacobson et al., 2020). At the WP level, however, momentary fluctuations in affect are expected to be triggered by specific situations that lead to distinct affective reactions (e.g., situations of loss lead to sadness, while situations of conflict lead to anger; as described by Brose et al., 2020). Based on these theoretical considerations, the affective structure is expected to be more discrete at the WP than BP level (Brose et al., 2015; Vansteelandt et al., 2005; Zelenski & Larsen, 2000).

Affect is a central feature of subjective experience and therefore crucial to many theories and research in psychology. Affective states are thought to guide behavior in reaction to external and internal events (Kuppens & Verduyn, 2017; Russell, 2003). When the internal or external situation changes, so does the affective state of a person. Therefore, affective states take a central role in many theories in psychology, such as theories of personality (e.g., Larsen & Ketelaar, 1991) or affective disturbances in psychopathology (e.g., Beck & Bredemeier, 2016; Telford, McCarthy-Jones, Corcoran, & Rowse, 2012). As affect fluctuates over time frames of minutes and hours as a function of the daily context it is embedded in, it is often assessed with ESM in the context of daily life (e.g., Houben, Van Den Noortgate, & Kuppens, 2015; Podsakoff, Spoelma, Chawla, & Gabriel, 2019). However, the BP and WP structures of affect assessed with ESM are still poorly understood.

Studies that were able to distinguish the WP and BP structure of affect have come to varying conclusions. Generally, a two-factor structure with correlated PA and NA showed adequate fit (Brose et al., 2015; Merz & Roesch, 2011), and fit better than a model with a single valence factor (Rappaport, Amstadter, & Neale, 2019; Rush & Hofer, 2014; Viechtbauer, Lataster, Rintala, Simons, Delespaul, Wichers, & Myin- Germeys, 2020) at both levels. Several studies also observed different correlations between PA and NA at the different levels of analysis (Bleidorn & Peters, 2011; Rush & Hofer, 2014; Schmukle, Egloff, & Burns, 2002). However, studies that have also included more complex structures (i.e., with more factors) have often found that these provided superior fit compared to a PA and NA structure. It has for instance been suggested that more differentiated NA factors are necessary to sufficiently explain the variance in affect ratings (Eadeh, Breaux, Langberg, Nikolas, & Becker, 2019; Möwisch, Schmiedek, Richter, & Brose, 2019), that three factors representing valence, calmness, and energetic arousal would be needed (Leonhardt, Könen, Dirk, & Schmiedek, 2016; Wilhelm & Schoebi, 2007), or that anxious mood, depressed mood, anger fatigue, and vigor should be distinguished (Cranford et al., 2006). In some of these cases, the structure depended on the level of analysis, with a more complex structure at the WP level (Charles, Mogle, Leger, & Almeida, 2019; Eadeh et al., 2019; Jacobson et al., 2020; Wilhelm & Schoebi, 2007). A recent study investigating daily affect suggested, for instance, that PA and NA emerge only at the BP level, while a seven-factor structure distinguishing discrete emotions provides the best fit at the WP level (Jacobson et al., 2020).

Most of these previous studies have investigated the structure of affect with variations of the PANAS in diaries. Daily diaries sample only once every day, typically at the same time and under the same circumstances, which likely reduced the types of affective states that can be sampled. Only four of the above studies investigated the factor structure of affect in an ESM context (Jacobson et al., 2020; Rappaport et al., 2019; Viechtbauer et al., 2020; Wilhelm & Schoebi, 2007) and it is currently unclear to what extent the findings from diary studies apply to affect measured with ESM. A further complicating factor is the wide variety of ESM protocols that are currently employed, with large variations in the number of items used and assessments per day. For instance, affect has been measured with nine unipolar items ten times per day in one of the above studies (Viechtbauer et al., 2020) and with six bipolar items presented four times per day in the other (Wilhelm & Schoebi, 2007). Cross-sectional studies have repeatedly shown how small variations in questionnaires, such as the addition of adjacent questions, can affect results (Schwarz, 1999). It is possible that more frequent assessments lead to the use of different reference points for ratings of affect (as suggested by Podsakoff et al., 2019), or that longer questionnaires exaggerate subtle differences between states (as suggested for instance by Wilhelm & Schoebi, 2007). Further, the proportion of error variance of affect measures may be influenced by the sampling protocol. Next to the structural validity of measures, the proportion of total variance of a measure that is considered to be true variance (i.e., reliability) is also an indication of the quality of a measure. As it is for instance possible to observe low reliability and excellent structural model fit, or vice versa, both structural validity and reliability are important to consider when evaluating the quality of a measure. However, whether variations in the ESM protocol influence the structure of affect or the reliability of measures of affect has not been investigated.

The current study extends previous research by investigating the factor structure and reliability of a large number of affect items with high and low arousal ratings (as opposed to the PANAS which includes only high arousal states) in ESM data. Specifically, we preregistered the comparison of structures that are frequently assumed in the ESM literature. Further, the special design of the current study allowed us to investigate the invariance of the affective structure across different sampling frequencies and questionnaire lengths that reflect the large diversity in sampling protocols used in ESM research. Although design characteristics of ESM studies may influence measures of affect, it has not been taken into account in previous investigations of the affective structure. In the current study, we investigated the structure of affect at the WP and BP level in a short and a long version of an ESM questionnaire. We hypothesized that at both levels and in both questionnaire versions, a two-factor structure with the correlated factors PA and NA[2] would provide the best fit compared to a single valence factor and to uncorrelated PA and NA factors. We then explored whether different sampling frequencies (three, six, or nine measurements per day) and questionnaire lengths (30 or 60 items), influence the structure of PA and NA. Finally, we investigated the reliability of PA and NA. We hypothesized that PA and NA would show satisfactory reliability in all groups but that reliability would be higher when based on more items. Both the confirmatory and exploratory analyses included in this paper were preregistered

[2] These two factors were based on the valence of the words, meaning that states with low, neutral, and high arousal were selected. The factors do thereby not correspond to the PA and NA factors defined by Watson and Tellegen (1985), but rather to the ends of the single valence factor defined by Russell (1980). Consequently, we expected a stronger negative correlation between the two factors than previously found with the PANAS items.

(https://osf.io/6t5ud) and deviations from the preregistration are indicated in the text.

2. Method

2.1. Sample

The sample consisted of 163 university students between 18 and 30 years old, who were required to be fluent in Dutch and to have never taken part in an ESM study before. Participants were recruited via social media and on campus sites. This study was powered for its main hypotheses that were investigated in a previous manuscript on changes in burden, compliance, and careless responding as a function of the experience sampling protocol (Eisele et al., 2020).

2.2. Protocol

Study procedures were approved by the Social and Societal Ethics Committee of KU Leuven. Upon arrival at the lab, participants provided informed consent. They were then randomly assigned to one of six experimental conditions (three, six, or nine beeps per day combined with a 30 or 60 item questionnaire). Participants were asked to fill in several baseline questionnaires and were briefed individually about the ESM protocol (more information about the baseline assessment and the briefing procedure is provided on the OSF page of this project). An ESM period of 14 days started the day after the briefing session. During this ESM period, participants were prompted three, six, or nine times per day to fill in an ESM questionnaire on a smartphone (Motorola DEFY+) that was provided by the researchers. We used the mobileQ app (Meers, Dejonckheere, Kalokerinos, Rummens, & Kuppens, 2020) for the ESM data collection. The prompts were randomly distributed in equal time windows lasting from 9 am until 10:30 pm. Participants had 90 s to start the questionnaire and 90 s to respond to each individual question before the questionnaire became unavailable. After the ESM period, participants returned to the lab to fill in a number of follow-up questionnaires. A random sample of participants also completed a qualitative interview about their experience of taking part in the ESM study. Then, participants received a remuneration of 40, 60, or 80 euros, depending on whether they had received three, six, or nine beeps per day.

2.3. Measures of affect[3]

Depending on their assigned condition, participants received either a short (30 items with full branching) or a long (60 items with full branching) ESM questionnaire. The short questionnaire contained four items measuring PA (“Right now, I feel happy/ relaxed/ energetic/ satisfied”) and four items measuring NA (“Right now, I feel stressed/ anxious/ irritated/ down”). In the long version of the questionnaire, there were additionally five items measuring PA (curious/ calm/ enthusiastic/ alert/ cheerful) and five items measuring NA (sad/ irritable/ tense/ disappointed/ restless). All items had Likert-type response scales ranging from 1, labelled with “Not at all”, to 7, labelled with “Very much”. The items were chosen from previous Dutch ESM studies and the Dutch version of the PANAS (Barge-Schaapveld, Nicolson, Berkhof, & Devries, 1999; Collip et al., 2011; Engelen, De Peuter, Victoir, Van Diest, & Van den Bergh, 2006; Geschwind, Peeters, Drukker, Van Os, & Wichers, 2011; Lataster et al., 2011; Matcham et al., 2019; Peeters, Nicholson, & Berkhof, 2003). Items were selected to cover a broad range of the affective space as described by both Russell and Watson and colleagues (Russell, 1980; Watson et al., 1999). We therefore selected items with positive and negative valence, ranging from low to high arousal. Whenever possible, the rating of individual adjectives on valence and arousal was confirmed with norm ratings of Dutch words (Moors et al., 2013; Verheyen, De Deyne, Linsen, & Storms, 2019). Some items were also selected for research questions unrelated to the current investigation, which lead to a larger proportion of high arousal terms.

2.4. Analysis

To answer our first research question, we compared three models of affect (one-factor model, two-factor model with uncorrelated NA/PA, and two factor model with correlated NA/PA) in all experimental groups simultaneously, separately for the WP and BP level. This was done by fitting partially saturated models (as suggested by Ryu & West, 2009; a saturated model imposes no structural restrictions, meaning that it fits the data perfectly). This means that when judging the fit of the WP model, a saturated BP model was specified and when selecting the BP model, a saturated WP model was specified. We chose this approach because misfit at the highest level may not be detected when testing both levels simultaneously when sample sizes at the BP level are small (Ryu & West, 2009). The fit of the partially saturated models was assessed by calculating level specific fit indices, namely a level specific Chi-square value, Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA). To compare models, the cut-off value for a significant ΔCFI was 0.005, and 0.01 for the ΔRMSEA (Chen, 2007). Additionally, a Chi-square test significant at $\alpha = 0.05$ was interpreted as showing a significant difference between two models. However, we only concluded that there is a significant difference between models when at least two of the three indices indicated that. The model showing the best fit was further evaluated by considering the level specific CFI, RMSEA, and Chi-square test. For these indices, the following values were interpreted as indicating acceptable fit: CFI $\geq$ 0.90, RMSEA $\leq$ 0.08, and a non-significant Chi-square test (Bentler, 1990; Hu & Bentler, 1999). However, in line with the model comparisons, we only concluded inadequate fit if at least two of these three fit indices indicated that. In all models, the metric of the latent construct was fixed with the effects coding method (Brown, 2015). When the preregistered structure did not result in acceptable fit, we proceeded to freeing error correlations one by one based on modification indices until an acceptable fit was reached. The preregistered strategy of adding cross loadings and removing the original loading of the item did not result in acceptable fit for any of the models and is therefore not reported. The code for all preregistered and exploratory model modifications that were conducted is provided on the OSF page of this project (https://osf.io/pzx8r/).

Subsequently, we tested for measurement invariance of the short questionnaire items across different sampling frequencies and different questionnaire lengths. These measurement invariance analyses were also conducted separately per level. At the WP level, we tested for configural (i.e., same structure) and metric invariance (i.e., same factor loadings) and for the invariance of latent variances and covariances (for more information on the types of measurement invariance, see Vandenberg & Lance, 2000). At the BP level, we tested for configural, metric, and scalar invariance (i.e., same item intercepts) and for the invariance of latent means, variances and covariances.

[3] Only variables used in the current analysis are reported. The reader is referred to the OSF page for a full overview of the ESM questionnaire and cross-sectional measures.

Finally, the reliability measure omega was calculated for NA and PA per level based on the results of the factor analyses (Bolger & Laurenceau, 2013).

All analyses were conducted in R, using the lavaan package for the CFAs (Rosseel, 2012). The lme4, sjstats (Lüdecke, 2020), and esmpack (Viechtbauer & Constantin, 2019) packages were used for preprocessing. Exploratory factor analyses (not preregistered) were conducted with the psych package (Revelle, 2018). More detail on the analyses can be found in the preregistration and the code is provided in the supplemental materials.

3. Results

3.1. Sample and item characteristics

Of 163 participants who took part in the baseline session, 3 had to be excluded because they did not meet the inclusion criteria and 2 participants dropped out. Additionally, 4 participants were identified as careless responders in a previous analysis of the dataset (Eisele et al., 2020) and for 2 participants the timing of the beeps was more than one hour off. These participants were therefore excluded from the current analyses. In case of technical problems that led to beeps not being delivered for more than one full day, the data gathered after this disruption was also excluded from the analyses. Further, 5 participants had to be excluded from the analyses because they showed no variance on at least one of the items. The final sample included 147 participants (short questionnaire 3, 6, and 9 times with 26, 23, and 23 participants, respectively, and the long questionnaire 3, 6, and 9 times with 25 participants in each frequency group). Beeps with missing values on affect items were excluded from the analyses (N = 2084, 17% of all beeps). Intraclass correlations (ICCs) of the items ranged from 0.20 to 0.39. ICCs refer to the proportion of variance in an item that is due to differences between individuals as opposed to differences within individuals. An ICC of 0.20 indicates that 20% of the variance of this item is explained by differences between individuals, while the majority of variance, namely 80%, is explained by differences within individuals. Several items showed skewed distributions (skewness and kurtosis in Table 1 give an indication of the deviations from a normal distribution of responses). CFA models operate under the distributional assumption of multivariate normality. Violations of this assumption are observed and even expected for many variables in the current study. Such violations have for instance been shown to lead to an inflation of $\chi 2$ values and thereby an over rejection of correctly specified models (Curran, West, & Finch, 1996). Although the maximum likelihood estimator used for the CFAs is robust against small violations of the normality assumption, such violations can influence results, especially in small samples (Brown, 2015). Estimators that control for the effects of violations of the normality assumption exist but are not currently implemented in the multilevel part of the lavaan package. This limitation should be kept in mind when interpreting the results, especially at the BP level, where fit indices may be biased. The descriptive statistics and ICCs of all items are reported in Table 1 and covariance matrices for the different experimental groups are reported in the supplemental materials.

3.2. Structure of affect

3.2.1. Short questionnaire: within-person structure

For the short version of the questionnaire, a two-factor model with correlated factors provided the best fit at the WP level (see Table 2 for the fit of all tested models). This means that it fitted significantly better than a model with orthogonal PA and NA factors and a model with a single valence factor. However, the fit measures did not indicate acceptable fit according to the preregistered cut-offs (RMSEA = 0.117; CFI = 0.892; $\chi^2(19) = 2654.39$, $p = 0.000$). Subsequently, the original model was extended in an exploratory way by letting the residuals of the items stressed and relaxed correlate (following the suggested modifications), which resulted in acceptable fit (RMSEA = 0.073; CFI = 0.961; $\chi^2(18) = 980.277$, $p = 0.000$). The structure with the standardized loadings is depicted in Fig. 1, the full model output is provided in Table S2 in the supplemental materials. Notably, the correlation of PA and NA was very high (−0.851). Further, modification indices as high as 195.91 (for the correlated residual between stressed and anxious) indicated that letting further residuals correlate could improve the fit of the model substantially.

3.2.2. Short questionnaire: between-person structure

At the BP level of the short questionnaire, a two factor structure with correlated PA and NA also fit the data best but did not result

Table 1
Descriptive statistics of affect items.

English item	Original Dutch item	Mean	SD Person-level means	Skewness Person-level means	Kurtosis Person-level means	SD Deviations from person-level means	Skewness Deviations from person-level means	Kurtosis Deviations from person-level means	ICC
Positive affect									
Happy	Gelukkig	5.029	0.718	−0.527	0.632	0.923	−0.538	1.496	0.362
Energetic	Energiek	4.081	0.734	−0.265	0.034	1.102	−0.234	0.295	0.298
Satisfied	Tevreden	4.801	0.796	−0.509	0.108	1.063	−0.475	1.047	0.337
Relaxed	Ontspannen	4.622	0.802	−0.181	−0.434	1.248	−0.417	0.409	0.276
Alert[1]	Alert	3.921	1.006	−0.249	−0.260	1.217	−0.103	0.552	0.393
Cheerful[1]	Opgewekt	4.044	0.778	−0.313	−0.915	1.209	−0.094	0.067	0.288
Enthusiastic[1]	Enthousiast	3.966	0.900	−0.316	−0.698	1.228	0.121	0.170	0.328
Curious[1]	Nieuwsgierig	3.163	0.987	0.053	−0.548	1.284	0.369	0.345	0.372
Calm[1]	Kalm	4.838	0.678	0.443	0.440	1.200	−0.615	0.572	0.210
Negative affect									
Stressed	Gestrest	2.384	0.852	1.041	1.502	1.179	0.913	1.298	0.351
Anxious	Angstig	1.510	0.558	2.039	5.588	0.785	2.021	7.153	0.358
Irritated	Geïrriteerd	1.942	0.629	0.967	0.486	1.126	1.405	2.726	0.227
Down	Somber	1.740	0.666	1.430	2.964	0.959	1.579	4.065	0.305
Sad[1]	Bedroefd	1.720	0.547	1.698	4.729	0.977	1.822	4.679	0.264
Irritable[1]	Prikkelbaar	2.248	0.661	0.696	0.077	1.210	0.978	1.259	0.204
Tense[1]	Gespannen	2.693	0.867	0.439	−0.521	1.231	0.535	0.259	0.321
Disappointed[1]	Teleurgesteld	2.032	0.713	1.300	2.283	1.185	1.388	2.881	0.261
Restless[1]	Rusteloos	2.242	0.766	0.813	0.877	1.109	0.903	1.403	0.296

Notes. [1] Item was only part of the long questionnaire; the remaining items were part of both questionnaires. ICC = Intraclass correlation.

Table 2
Model fit and comparisons.

Model	CFI	RMSEA	df	AIC	BIC	χ^2	*p*	Diff.χ^2	df diff.	*p* diff.	Diff. RMSEA	Diff. CFI
Short Questionnaire												
Within Person (N = 10154 observations)												
Model A: Correlated PA & NA	0.892	0.117	19	219951.749	220392.512	2654.390	0.000					
Model B: Valence	0.882	0.119	20	220209.030	220642.567	2913.671	0.000	259.281	1	0.000	−0.002	0.011
Model C: Orthogonal PA & NA	0.672	0.199	20	225352.723	225786.260	8057.363	0.000	5402.974	1	0.000	−0.082	0.220
Model A with correlated residuals [a]	0.961	0.073	18	218279.637	218727.625	980.277	0.000					
Between Person (N = 147 persons)												
Model A: Correlated PA & NA	0.996	0.189	19	217416.216	217856.979	118.857	0.000					
Model B: Valence	0.992	0.260	20	217514.554	217948.091	219.195	0.000	100.338	1	0.000	−0.071	0.004
Model C: Orthogonal PA & NA	0.992	0.253	20	217504.077	217937.614	208.718	0.000	89.861	1	0.000	−0.064	0.004
Model A with correlated residuals [b]	1.000	0.064	15	217329.488	217799.154	24.129	0.063					
Combined Within and Between Person Model A with correlated residuals at both levels	0.960	0.054	33	218274.879	218614.483	1005.520	0.000					
Long Questionnaire												
Within Person (N = 5045 observations)												
Model A: Correlated PA & NA	0.786	0.104	134	258221.292	259696.202	7480.288	0.000					
Model B: Valence [e]												
Model C: Orthogonal PA & NA	0.710	0.121	135	260820.609	262288.994	10081.605	0.000	5402.974	1	0.000	−0.017	0.076
Model A with correlated residuals [c]	0.901	0.073	127	254294.128	255814.721	3539.124	0.000					
Between Person (N = 75 persons)												
Model A: Correlated PA & NA	0.987	0.211	134	251321.459	252796.369	580.455	0.000					
Model B: Valence	0.981	0.250	135	251509.150	252977.535	770.146	0.000	189.691	1	0.000	−0.040	0.006
Model C: Orthogonal PA & NA	0.986	0.215	135	251342.443	252810.828	603.439	0.000	22.984	1	0.000	−0.004	0.001
Model A with correlated residuals [d]	0.998	0.079	113	250948.329	252560.289	165.325	0.001					

Notes. Model comparisons were conducted with Model A of the respective level and questionnaire version. [a] stressed-relaxed; [b] relaxed-stressed, happy-down, stressed-down, relaxed-anxious; [c] stressed-relaxed, sad-down, irritable-irritated, calm-relaxed, tense-stressed, relaxed-tense, alert-energetic; [d]: sad-down, happy-satisfied, calm-relaxed, irritable-irritated, sad-anxious, anxious- down, tense-restless, tense-stressed, alert-energetic, disappointed-restless, cheerful-enthusiastic, enthusiastic-curious, satisfied-disappointed, curious-irritated, cheerful-energetic, enthusiastic-energetic, curious-energetic, alert-curious, happy-stressed, happy-tense, cheerful-curious; [e] Model did not converge. PA = Positive affect; NA = Negative affect; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation.

in an acceptable fit (RMSEA = 0.189; CFI = 0.996; $\chi^2(19) = 118.857$, $p = 0.000$). Subsequently, residuals of items were allowed to correlate in an exploratory way. We proceeded in a step-wise manner, freeing the residual correlation with the highest modification index, testing model fit, then proceeding to the next residual correlation. This approach resulted in a fitting model with residual correlations between relaxed and stressed, happy and down, stressed and down, and relaxed and anxious (see Fig. 1 & Table S2 in the supplemental materials; RMSEA = 0.064; CFI = 1; $\chi^2(15) = 24.129$, $p = 0.063$).

3.2.3. Long Questionnaire: WP structure

For the long version of the questionnaire, the structure with correlated PA and NA resulted in the best fit at the WP level (RMSEA = 0.104; CFI = 0.786; $\chi^2(134) = 7480.288$, $p = 0.000$). However, the fit was not acceptable based on the pre-specified criteria. Subsequently, we proceeded to freeing residual correlations in an exploratory way, as described above. Freeing seven residual correlations between closely related items or antonyms (stressed-relaxed, sad-down, irritable-irritated, calm-relaxed, tense-stressed, relaxed-tense and alert-energetic) resulted in an acceptable fit (RMSEA = 0.073; CFI = 0.901; $\chi^2(127) = 3539.124$, $p = 0.000$; see Table 2 for all fitted models; coefficients of the model with correlated residuals are provided in Table S3 in the supplemental materials).

3.2.4. Long questionnaire: BP structure

At the BP level of the long questionnaire, the model with two correlated factors did not fit significantly better than a model with independent PA and NA factors. For consistency, the correlated factor structure was retained. However, the fit was not acceptable based on the pre-specified criteria (RMSEA = 0.211; CFI = 0.987; $\chi^2(134) = 580.455$, $p = 0.000$). As described above, residual correlations were then freed in the original model in an exploratory way, which resulted in an acceptable fit (freed residual correlations: sad-down, happy-satisfied, calm-relaxed, irritable-irritated, sad-anxious, anxious-down, tense-restless, tense-stressed, alert-energetic, disappointed-restless, cheerful-enthusiastic, enthusiastic-curious, satisfied-disappointed, curious-irritated, cheerful-energetic, enthusiastic-energetic, curious-energetic, alert-curious, happy-stressed, happy-tense, cheerful-curious; RMSEA = 0.079; CFI = 0.998; $\chi^2(113) = 165.325$, $p = 0.001$; see Table S3 in the supplemental materials for the coefficients).

3.3. Exploratory factor analyses

The high number of residual correlations necessary to achieve only borderline acceptable fit in the long version of the questionnaire (7 for WP and 21 for BP level) led us to doubt the tenability of the preregistered structures. We therefore deviated from the preregistration by conducting exploratory factor analyses (EFAs) on the person-level means and deviations from these means (i.e., the within-person mean centered data) of the long questionnaire. To detect the number of factors to retain, parallel analyses were conducted with maximum likelihood factor extraction based on the correlation matrices of the person-level means and the deviations from these means. The suggested number of factors was two at the BP level and five at the WP level. The detected structure at the BP level resembled the planned PA and NA structure with the exception that calm and relaxed loaded (negatively) on the NA factor. At the WP level items seemed to form clusters of different degrees of arousal and valence. Additionally, a factor with items relating to irritation emerged. The loadings of the items on the individual factors and correlations between factors are shown in Tables 3 and 4.

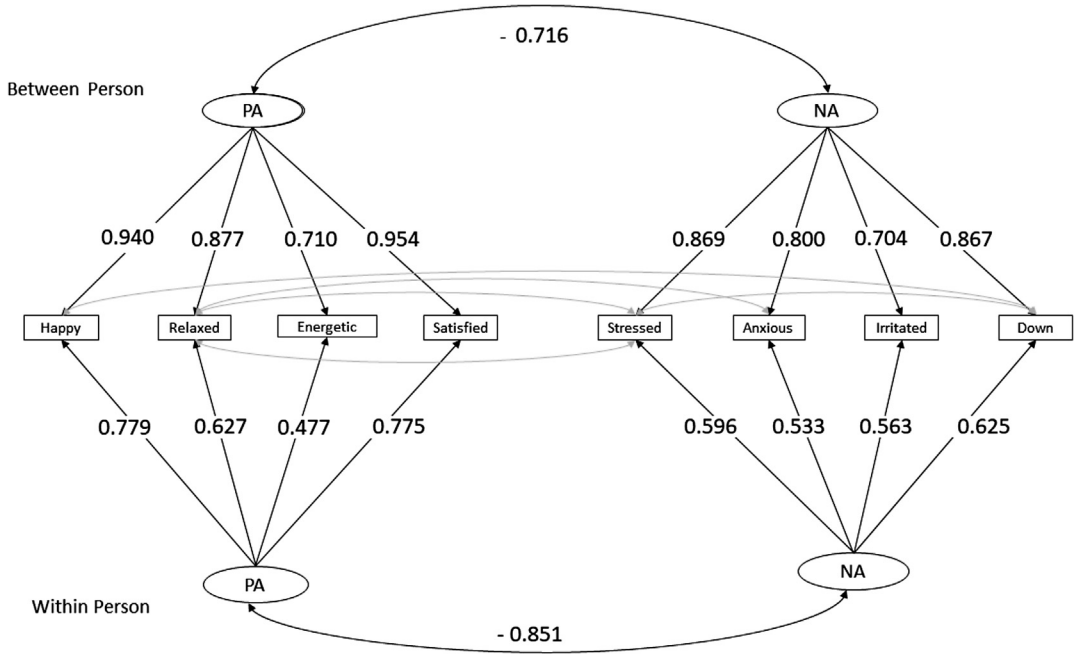

Fig. 1. Structures at the within- and between-person level with standardized loadings and factor correlations. Residual correlations are indicated with grey arrows. Residual variances and factor variances are omitted for simplicity. PA = positive affect; NA = negative affect. All coefficients significant at $p < 0.005$.

Table 3
Exploratory factor analysis of the long questionnaire person-level means.

Item	Factor 1: NA	Factor 2: PA
Happy	−0.370	**0.636**
Energetic	0.018	**0.915**
Satisfied	−0.373	**0.662**
Relaxed	**−0.669**	0.356
Alert	0.251	**0.634**
Cheerful	−0.035	**0.920**
Enthusiastic	−0.045	**0.922**
Curious	0.210	**0.766**
Calm	**−0.568**	0.004
Stressed	**0.888**	0.001
Anxious	**0.728**	0.159
Irritated	**0.567**	−0.032
Down	**0.757**	−0.047
Sad	**0.781**	0.053
Irritable	**0.567**	−0.054
Tense	**0.870**	0.024
Disappointed	**0.779**	−0.114
Restless	**0.703**	0.063
Correlations	Factor 1	Factor 2
Factor 1	1	−0.48
Factor 2		1

Notes. Highest loading per item highlighted in bold. NA = Negative affect; PA = Positive affect; Fit indices: RMSEA = 0.212; Tucker Lewis Index = 0.647; $\chi 2$ (118) = 453.52, $p < 0.001$; N = 75 persons.

3.4. Measurement invariance across sampling frequencies and questionnaire lengths

Next, we investigated whether differences in the sampling protocol (three, six, or nine beeps per day; 30 or 60 items) would have an influence on the structure of affect. We therefore conducted measurement invariance analyses of the best fitting WP structure of the short questionnaire (i.e., using the affect items that are common to both the 30- and the 60-item questionnaires). This structure was found to meet configural and metric invariance and invariance of the latent variances and correlations across both different sampling frequencies and different questionnaire lengths (see Table 5). This means that the overall structure of affect, the loadings of items, and the relationship between PA and NA were not influenced by variations in the sampling protocol. The originally planned measurement invariance analyses of the BP structure should be interpreted cautiously due to the small sample sizes of the groups (sampling frequency groups included 51, 48, and 48 participants; questionnaire length groups 75 and 72 participants) and are reported in the supplemental materials.

3.5. Reliabilities

The reliabilities of PA and NA for the short questionnaire in the model with correlated residuals specified at both levels were high (see Table 2 for fit indices of this combined model). The reliability was $\omega = 0.818$ for PA and $\omega = 0.815$ for NA at the WP level and $\omega = 0.934$ for PA and $\omega = 0.929$ for NA at the BP level. The reliabilities for the individual frequency and questionnaire length groups are reported in Table S1 in the supplemental materials, as well as the reliabilities for the long questionnaire version.

4. Discussion

In the current study, we investigated the structure of affect measured with ESM, the invariance of this structure across different questionnaire lengths and sampling frequencies, and the reliability of PA and NA. As hypothesized, a two-factor structure with correlated PA and NA showed the best fit compared to a model with orthogonal factors and a unidimensional model. However, the simple PA and NA structure did not result in a satisfactory fit. Results of the CFAs and EFAs indicated the presence of additional shared variance between items that could be accounted for by freeing residual correlations or adding more nuanced factors. For the short version of the questionnaire, an exploratory structure accounting for these additional relationships with freed residual correlations was invariant across different sampling protocols at the WP level and showed high reliability. Differences between levels in the strength of loadings, correlations between PA and NA, and the number of suggested factors in EFAs could be observed

Table 4
Exploratory factor analysis of the long questionnaire deviations from the person-level means.

Item	Factor 1: PA high arousal	Factor 2: PA low arousal	Factor 3: NA high arousal	Factor 4: NA low arousal	Factor 5: NA irritation
Happy	0.202	**0.450**	−0.061	−0.203	−0.128
Energetic	**0.676**	0.025	0.082	−0.020	0.010
Satisfied	0.287	**0.311**	−0.170	−0.093	−0.149
Relaxed	0.061	0.223	**−0.650**	0.050	0.012
Alert	**0.528**	−0.158	0.107	−0.027	0.022
Cheerful	**0.695**	0.047	−0.087	−0.040	0.012
Enthusiastic	**0.696**	0.057	−0.075	0.009	−0.055
Curious	**0.531**	−0.079	0.035	0.055	−0.002
Calm	−0.093	0.068	**−0.465**	0.029	−0.103
Stressed	0.010	0.017	**0.794**	0.000	0.005
Anxious	0.061	0.008	0.375	**0.381**	−0.090
Irritated	0.013	−0.010	0.018	0.024	**0.732**
Down	−0.035	−0.021	0.009	**0.799**	−0.038
Sad	0.008	0.001	−0.022	**0.812**	0.044
Irritable	−0.015	0.003	0.009	−0.014	**0.773**
Tense	−0.084	0.098	**0.670**	0.062	0.088
Disappointed	−0.043	−0.039	0.027	**0.436**	0.262
Restless	0.037	0.088	**0.412**	0.129	0.160
Correlations	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1	1	0.50	−0.32	−0.45	−0.44
Factor 2		1	−0.44	−0.28	−0.32
Factor 3			1	0.44	0.55
Factor 4				1	0.52
Factor 5					1

Notes. Highest loading per item highlighted in bold. NA = Negative affect; PA = Positive affect; Fit indices: RMSEA = 0.033; Tucker Lewis Index = 0.975; $\chi 2(73) = 467.36$, $p < 0.001$; N = 5045 observations

Table 5
Measurement invariance analyses of short questionnaire within-person structure.

Model	CFI	RMSEA	df	AIC	BIC	χ^2	*p*	diff. χ^2	df diff.	*P* diff.	diff. RMSEA	diff. CFI
Across sampling frequencies (3 beeps N = 1786; 6 beeps N = 3390; 9 beeps N = 4978)												
Configural	0.959	0.074	54	218141.142	219485.108	1052.085	0.000					
Metric	0.959	0.067	66	218131.685	219388.943	1066.627	0.000	14.542	12	0.267	0.007	0.000
(Co)Variances	0.957	0.066	72	218173.463	219387.368	1120.405	0.000	53.778	6	0.000	0.001	0.002
Across questionnaire lengths (short questionnaire N = 5067; long questionnaire N = 5087)												
Configural	0.960	0.074	36	218104.986	219000.963	1025.772	0.000					
Metric	0.959	0.068	42	218107.731	218960.354	1040.517	0.000	14.745	6	0.022	0.005	0.000
(Co)Variances	0.959	0.066	45	218104.331	218935.278	1043.117	0.000	2.601	3	0.457	0.002	0.000

Notes. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation.

which point towards a more discrete structure at the WP than BP level.

4.1. Support for PA & NA structure

In line with our hypotheses, a PA and NA structure fit the data better than a single valence factor or two orthogonal factors in all cases, thereby offering support for previous findings from both diary and ESM data (Rappaport et al., 2019; Rush & Hofer, 2014; Viechtbauer et al., 2020). However, it is important to note that the fit of a simple PA and NA structure was not satisfactory in any of the cases, a fact that is discussed in more detail below. The high correlation of PA and NA at the WP level of the short questionnaire indicated a high degree of redundancy between the two factors. This is likely due to the inclusion of lexical antonyms in the questionnaire (e.g., stressed and relaxed, happy and sad; see Russell & Carroll, 1999) and the definition of positive and negative affect based on valence and hence across arousal levels (as opposed to the definition based on only high arousal terms used by Watson and Tellegen, 1985). However, model comparisons showed that two separate but correlated factors resulted in significantly better fit than one factor, indicating that PA and NA should not be collapsed into one factor in the current data. Further, relaxed and calm showed high cross-loadings and were found to load on the NA factor in the exploratory analyses of the long questionnaire. This indicates that they are more characteristic of the absence of negative than of the presence of other positive affective states (in line with Watson et al., 1999). For the short version of the questionnaire, a WP PA and NA structure that allowed a correlation between the residuals of stressed and relaxed was found to be invariant across protocols, suggesting that variations in sampling frequency and questionnaire length do not influence the measures of affect. At the BP level, the structure did not fit in all groups, which may indicate that sampling frequency and questionnaire length do influence the relationships between individual items at this level. However, the small sample size at the BP level may also explain the misfit. This result should therefore be interpreted cautiously and replicated in a larger sample in the future. Even if the functioning of items is indeed not affected, it is important to note that the sampling frequency does have an effect on the timing of changes in affect that can be detected (Ebner-Priemer & Sawitzki, 2007) and questionnaire length has previously been observed to lead to higher burden, reported careless responding, and lower compliance (Eisele et al., 2020). PA and NA were also found to show high internal consistency at both levels. The values observed at the WP level were at the higher end of reliabilities reported in the literature, which may partly be caused by the method used to calculate reliability (Brose et al., 2020).

4.2. Additional shared variance between items

Notably, the simple PA and NA structure did not show acceptable fit for any of the levels and questionnaire versions. Modification indices suggested that items shared additional variance that was not accounted for by PA, NA, or their correlation. Although this tendency was present in both questionnaire versions, additional relations between items became most apparent in the long version. Compared to the short version, this questionnaire was characterized by generally increased redundancy of items (e.g., irritated and irritable) and the inclusion of synonyms (e.g., sad and down). The resulting misfit of the simple PA and NA structure was addressed with two different approaches (the association between item redundancy and model misfit has for instance been discussed by Brown, 2015). First, residual correlations were freed based on modification indices. Allowing residual correlations is in line with previous work using the PANAS, where items belonging to the same content domains are often allowed to correlate at both levels to achieve good fit (e.g., Brose et al., 2015; Merz & Roesch, 2011; Rush & Hofer, 2014). However, the allowed residual correlations had not been specified in advance for the current study. Modifying a confirmatory model in such a data-driven way has been criticized as being a-theoretical and therefore especially susceptible to sample specific noise (Hermida, 2015). This means that even though the conducted modifications may seem in line with previous work on affect and may inform about interesting tendencies in the data, they should be interpreted in light of how they were achieved and warrant replication in a new dataset. As discussed in the literature, residual correlations may also hint towards the existence of additional factors (see Gerbing & Anderson, 1984). Indeed, an even more exploratory approach revealed that the additional relations between items could also be accounted for by specifying extra factors at the WP level. This is in line with findings from previous studies that have suggested a more nuanced affective structure within persons (Eadeh et al., 2019; Jacobson et al., 2020; van Halem, Van Roekel, Kroencke, Kuper, & Denissen, 2020). In the current data, this approach resulted in five correlated factors at the WP level that roughly reflected clusters of items characterized by varying degrees of valence and arousal (high arousal PA, high arousal NA, low arousal PA, low arousal NA). Additionally, the NA factor irritation emerged, which revealed a clustering of affective states that typically co-occur in emotional episodes with specific antecedents, appraisals, and behavioral consequences (Russell, 2003). All factors were therefore justifiable based on theoretical accounts of the affective space (Russell, 2003; Russel, 1980; Zevon & Tellegen, 1982). Notably, several items showed high cross-loadings on other factors (e.g., anxious, satisfied). This may be explained by the fact that they are centrally located on the arousal spectrum and do not belong clearly to the group of high or low arousal terms. Taken together, the two approaches reflect discrepancies in the literature, where additional associations between items have sometimes been treated as a nuisance, and hence accounted for by freeing correlated residuals, and sometimes as meaningful additional factors.

4.3. Differences in affective structure between levels

The current findings also highlight differences in the structure of affective states between persons and within persons that are consistent with a more discrete affective structure at the WP level (i.e., individual affective states co-occur less at WP than BP level). In line with this, a relatively large part of the variance of affective states between persons was accounted for by the dimensions PA and NA (e.g., individuals who frequently feel sad also tend to feel stressed often), while these broad dimensions explained a smaller part of the variance in affective states within persons over time (e.g., if an individual feels sad at a given moment they are only slightly more likely to also feel stressed at the exact same moment). This was apparent in the strength of loadings at the two levels (e.g., only 23% of WP variance of energetic was accounted for by WP PA, compared to 50% of BP variance accounted for by BP PA; consistent with e.g., Eadeh et al., 2019; Möwisch et al., 2019). Another sign of a more discrete affective structure was that the correlation of PA and NA was more negative at the WP level, similar to previous findings (Bleidorn & Peters, 2011; Rush & Hofer, 2014; Schmukle et al., 2002). This indicates that affective states of opposite valence are also less likely to co-occur at the WP than BP level. In addition, the number of factors suggested in the parallel analyses differed between levels, with a more nuanced structure at the WP level. The small sample size at the BP level may partly explain this discrepancy, as parallel analysis, the method used to determine the number of factors, is known to underestimate the number of factors in situations with correlated factors and small sample sizes (Auerswald & Moshagen, 2019). However, this finding may also point in the direction of a more nuanced structure of affect at WP than BP level (in line with Jacobson et al., 2020). Taken together, these observations are in line with theories and empirical findings describing a different affective structure at the WP than BP level (Brose et al., 2015, 2020; Vansteelandt et al., 2005; Zelenski & Larsen, 2000). However, the tenability of a more discrete WP structure should be tested in a confirmatory way in the future. It is important to highlight that while WP affect may be more discrete, a more nuanced organization of affective states was also prevalent at the BP level, as indicated by the residual correlations necessary to achieve acceptable model fit in the current study.

4.4. Limitations & future research

While the current study furthers our understanding of the WP and BP affective structure, we also faced a number of limitations that should be addressed in the future. The sample size at the BP level was small. This has consequences for the results of the CFAs, as the precision and power of model parameters, as well as the reliability of fit statistics is lower for small sample sizes (Brown, 2015). The power of measurement invariance tests is also known to be affected by the sample size (Chen, 2007; Meade, 2005). The results of the current BP analyses should therefore be replicated in larger samples in the future before drawing firm conclusions.

Another limitation was the item set, which was not fully balanced in terms of arousal and did not cover all affective states. The field would benefit from more systematic approaches including even larger item pools. Additionally, the investigation was conducted in a young student sample and the results may not generalize to other groups.

Further, the application of multilevel CFA/SEM is relatively recent and many methodological choices have not been investigated in detail. This applies for instance to cut-offs for fit indices. Recent work has highlighted that these cut-offs are highly dependent on the characteristics of the model (McNeish & Wolf, 2020). However, simulation work for multilevel CFAs is sparse. Relatedly, using good model fit as a proxy of the quality of a measure can be misleading, as synonyms, antonyms, or cross loading items can lead to model misfit, but may be desirable from a reliability or content validity point of view (see also Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011 for a discussion of the inverse relationship between reliability and model fit). Moreover, we responded to the misfit of the confirmatory models by engaging in exploratory post hoc model modification. This data-driven approach has been criticized as being too susceptible to sample specific noise (Hermida, 2015) and the resulting structures should therefore be interpreted as exploratory and warrant replication in new samples

before being given the same weight as the confirmatory factor analyses reported in this paper. Alternatives such as the Bayesian lasso (Pan, Ip, & Dubé, 2017), that relax some of the stringent assumptions of CFA, may be a valuable approach in the future.

As noted earlier, beeps with missing data on the affect items were excluded from the analyses. In the vast majority of cases (N = 1994), these stem from beeps that were not responded to by the participants (in a few rare cases [N = 60], a participant has started to respond to a beep but stopped while filling out the affect items). Missing data always have the potential to bias the results and we cannot exclude this possibility for the present analyses. Future research should consider alternative ways of handling these types of missing data.

Further, the development of models that do integrate the temporal dependency of the data with variations from time series analysis is an active area of research, and first implementations in software exist (McNeish & Hamaker, 2019). Investigations in the future could benefit from taking the temporal dependency of ESM data into account. As the implementation of multilevel EFA in standard software is still limited, we used standard analytical tools on person-level means and deviations from these means separately. Although the separate analysis of person-level means and deviations thereof does produce similar results to more complex statistical approaches in many cases, it is possible that this influenced the results.

Finally, previous research has indicated that there are meaningful differences in affective structures between persons (Brose et al., 2015) and within persons over time (Dejonckheere et al., 2019; Vogelsmeier, Vermunt, Bülow, & De Roover, 2020; Vogelsmeier, Vermunt, van Roekel, & De Roover, 2019). In the current study, we investigated the average WP structure but allowing the structure to be different across and within individuals may provide additional valuable information. New methodological advancements like latent Markov factor analysis (Vogelsmeier et al., 2019) allow to investigate such temporal changes in factor structures in more detail. First results using this method have indicated that a simple PA and NA structure may alternate with more nuanced structures over time (Vogelsmeier et al., 2019).

4.5. Conclusions and recommendations

Our results suggest that PA and NA emerge at both levels in ESM data and show high reliability, also when measured with few items. However, the fit of a simple PA and NA structure was not acceptable, which indicates that variance in the individual items is lost if one only considers PA and NA. This was especially apparent at the WP level, suggesting that more nuanced factors are a more accurate description of affective states within individuals. Therefore, when affect is the main interest of the ESM study, it may be better to include more items to be able to detect these more nuanced factors. Ideally, a strong theoretical foundation is needed for relying on PA and NA, more nuanced factors, or even individual items, at both levels. As pointed out by Brose and colleagues (2020), reasons of covariation on dimensions at the WP level are often difficult to explain theoretically, but there are certainly situations where these broad dimensions are of interest. In cases when a theory allows only vague predictions, comparing results obtained with different sub-factors and general PA/NA may be a powerful way to refine the existing theory (see Möwisch et al., 2019 for an example). Such comparisons could further clarify if the subdivision of affect into more nuanced factors is necessary. The low WP loadings found in the current work also highlight that the comparability of sum scores of different item sets is likely limited, which is worrisome considering the large diversity in items used to measure WP affect in previous studies (Brose et al., 2020). Additionally, the combination and number of items has a direct influence on the structure that can be detected. As discussed by Möwisch and colleagues (2019) and evident in the current results, more nuanced factors can only be detected when sufficient items are included. The dependency of the structure on the specific item set also highlights the need of constant assessment of the adequacy of a structure. Therefore, its structural validity should ideally be a routine check with every new item set and researchers should rely as much as possible on item sets with known structures. Finally, there are a number of methodological challenges associated with conducting factor analyses in ESM data that should be addressed in the future.

5. Funding Information

This research was funded by an Odysseus grant (Grant G0F8416N) allocated to Inez Myin-Germeys by Fonds voor Wetenschappelijk Onderzoek (FWO).

CRediT authorship contribution statement

Gudrun Eisele: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Software, Visualization, Writing - original draft. **Ginette Lafit:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Hugo Vachon:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Peter Kuppens:** Conceptualization, Writing - review & editing. **Marlies Houben:** Conceptualization, Writing - review & editing. **Inez Myin-Germeys:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing. **Wolfgang Viechtbauer:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded by an Odysseus grant (Grant G0F8416N) allocated to Inez Myin-Germeys by Fonds voor Wetenschappelijk Onderzoek (FWO).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2021.104094.

References

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods, 24*(4), 468–491. https://doi.org/10.1037/met0000200.

Barge-Schaapveld, D., Nicolson, N., Berkhof, J., & Devries, M. (1999). Quality of life in depression: Daily life determinants and variability. *Psychiatry Research, 88*(3), 173–189. https://doi.org/10.1016/S0165-1781(99)00081-5.

Beck, A. T., & Bredemeier, K. (2016). A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. *Clinical Psychological Science, 4*(4), 596–619. https://doi.org/10.1177/2167702616628523.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238.

Bleidorn, W., & Peters, A. L. (2011). A multilevel multitrait-multimethod analysis of self-and peer-reported daily affective experiences. *European Journal of Personality, 25*(5), 398–408. https://doi.org/10.1002/per.804.

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research (Methodology in the social sciences).* New York, NY.: Guilford Press.

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within- person affect variation. *Emotion, 20*(4), 677–699. https://doi.org/10.1037/emo0000583.

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between-person and within-person structures of affect are a matter of degree. *European Journal of Personality, 29*(1), 55–71. https://doi.org/10.1002/per.1961.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY.: Guilford Press.

Charles, S. T., Mogle, J., Leger, K. A., & Almeida, D. M. (2019). Age and the factor structure of emotional experience in adulthood. *The Journals of Gerontology: Series B, 74*(3), 419–429. https://doi.org/10.1093/geronb/gbx116.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834.

Collip, D., Nicolson, N., Lardinois, M., Lataster, T., Van Os, J., & Myin-Germeys, I. (2011). Daily cortisol, stress reactivity and psychotic experiences in individuals at above average genetic risk for psychosis. *Psychological Medicine, 41*(11), 2305–2315. https://doi.org/10.1017/S0033291711000602.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably?. *Personality and Social Psychology Bulletin, 32*(7), 917–929. https://doi.org/10.1177/0146167206287721.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16. https://doi.org/10.1037/1082-989X.1.1.16.

Dejonckheere, E., Mestdagh, M., Verdonck, S., Lafit, G., Ceulemans, E., Bastian, B., & Kalokerinos, E. K. (2019). The relation between positive and negative affect becomes more negative in response to personally relevant events. *Emotion*. https://doi.org/10.1037/emo0000697.

Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology, 47*(5), 1105–1117. https://doi.org/10.1037/0022-3514.47.5.1105.

Eadeh, H. M., Breaux, R., Langberg, J. M., Nikolas, M. A., & Becker, S. P. (2019). Multigroup multilevel structure of the child and parent versions of the positive and negative affect schedule (PANAS) in adolescents with and without ADHD. *Psychological Assessment*. https://doi.org/10.1037/pas0000796.

Ebner-Priemer, U. W., & Sawitzki, G. (2007). Ambulatory assessment of affective instability in borderline personality disorder: The effect of the sampling frequency. *European Journal of Psychological Assessment, 23*(4), 238–247. https://doi.org/10.1027/1015- 5759.23.4.238.

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*.

Engelen, U., De Peuter, S., Victoir, A., Van Diest, I., & Van den Bergh, O. (2006). Verdere validering van de "Positive and Negative Affect Schedule" (PANAS) en vergelijking van twee Nederlandstalige versie. *Gedrag en Gezondheid, 34*, 89–102.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research, 11*(1), 572–580. https://doi.org/10.1086/208993.

Geschwind, N., Peeters, F., Drukker, M., Van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 79*(5), 618–628. https://doi.org/10.1037/a0024595.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait–state model. *Journal of Research in Personality, 41*(2), 295–315. https://doi.org/10.1016/j.jrp.2006.04.003.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*(3), 319–336. https://doi.org/10.1037/a0024917.

Hermida, R. (2015). The problem of allowing correlated errors in structural equation modeling: Concerns and considerations. *Computational Methods in Social Sciences, 3*(1), 5–17.

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin, 141*(4), 901–930. https://doi.org/10.1037/a0038822.

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.

Jacobson, N. C., Evey, K. J., Wright, A. G., & Newman, M. G. (2020). Integration of discrete and global structures of affect across three large samples: specific emotions within-persons and global affect between-persons. https://doi.org/10.31234/osf.io/gb5up.

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology, 17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004.

Larsen, R. J., & Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *Journal of Personality and Social Psychology, 61*(1), 132. https://doi.org/10.1037/0022-3514.61.1.132.

Lataster, J., Myin-Germeys, I., Wichers, M., Delespaul, P., Van Os, J., & Bak, M. (2011). Psychotic exacerbation and emotional dampening in the daily life of patients with schizophrenia switched to aripiprazole therapy: A collection of standardized case reports. *Therapeutic Advances in Psychopharmacology, 1*(5), 145–151. https://doi.org/10.1177/2045125311419552.

Leonhardt, A., Könen, T., Dirk, J., & Schmiedek, F. (2016). How differentiated do children experience affect? An investigation of the within-and between-person structure of children's affect. *Psychological Assessment, 28*(5), 575–585. https://doi.org/10.1037/pas0000195.

Lüdecke, D. (2020). _sjstats: Statistical Functions for Regression Models (Version 0.17.9). https://doi.org/10.5281/zenodo.1284472.

Matcham, F., Pietro, C., Barattieri Bulgari di San, V., De Girolamo, G., Dobson, R., Eriksson, H., & Hotopf, M. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry, 19*(1). https://doi.org/10.1186/s12888-019-2049-z.

McNeish, D., & Hamaker, E. L. (2019). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*. https://doi.org/10.1037/met0000250.

McNeish, D., & Wolf, M. G. (2020). Dynamic fit index cutoffs for confirmatory factor analysis models. https://doi.org/10.31234/osf.io/v8yru.

Meade, A. W. (2005). Sample size and tests of measurement invariance. In Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Meers, K., Dejonckheere, E., Kalokerinos, E. K., Rummens, K., & Kuppens, P. (2020). mobileQ: A free user-friendly application for collecting experience sampling data. *Behavior Research Methods, 52*, 1510–1515. https://doi.org/10.3758/s13428-019- 01330-1.

Merz, E. L., & Roesch, S. C. (2011). Modeling trait and state variation using multilevel factor analysis with PANAS daily diary data. *Journal of Research in Personality, 45*(1), 2–9. https://doi.org/10.1016/j.jrp.2010.11.003.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A. L., ... Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45*(1), 169–177. https://doi.org/10.3758/s13428-012-0243-8.

Möwisch, D., Schmiedek, F., Richter, D., & Brose, A. (2019). Capturing affective well-being in daily life with the Day Reconstruction Method: A refined view on positive and negative affect. *Journal of Happiness Studies, 20*(2), 641–663. https://doi.org/10.1007/s10902-018-9965-3.

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods, 22*(4), 687–704. https://doi.org/10.1037/met0000112.

Peeters, F. A., Nicholson, N., & Berkhof, J. (2003). Cortisol responses to daily events in major depressive disorder. *Psychosomatic Medicine, 65*(5), 836–841. https://doi.org/10.1097/01.PSY.0000088594.17747.2E.

Podsakoff, N. P., Spoelma, T. M., Chawla, N., & Gabriel, A. S. (2019). What predicts within- person variance in applied psychology constructs? An empirical examination. *Journal of Applied Psychology, 104*(6), 727–754.

Rappaport, L., Amstadter, A., & Neale, M. (2019). Model fit estimation for multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(2), 318–329. https://doi.org/10.1080/10705511.2019.1620109.

Revelle, W. (2018). psych: Procedures for personality and psychological research. Evanston, Illinois, USA: Northwestern University. https://CRAN.R-project.org/package=psych Version = 1.8.12.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36 http://www.jstatsoft.org/v48/i02/.

Rush, J., & Hofer, S. M. (2014). Differences in within-and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment, 26*(2), 462–473. https://doi.org/10.1037/a0035666.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145–172. https://doi.org/10.1037/0033-295x.110.1.145.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*(1), 3–30. https://doi.org/10.1037/0033-2909.125.1.3.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16*(4), 583–601. https://doi.org/10.1080/10705510903203466.

Schimmack, U., & Grob, A. (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality, 14*(4), 325–345. https://doi.org/10.1002/1099- 0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44*(4), 229–237. https://doi.org/10.1037/h0055778.

Schmukle, S. C., Egloff, B., & Burns, L. R. (2002). The relationship between positive and negative affect in the Positive and Negative Affect Schedule. *Journal of Research in Personality, 36*(2002), 463–475. https://doi.org/10.1016/S0092-6566(02)00007-7.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93.

Telford, C. R. R. G., McCarthy-Jones, S., Corcoran, R., & Rowse, G. (2012). Experience sampling methodology studies of depression: The state of the art. *Psychological Medicine, 42*(6), 1119. https://doi.org/10.1017/S0033291711002200.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

van Halem, S., Van Roekel, E., Kroencke, L., Kuper, N., & Denissen, J. (2020). Moments that matter? On the complexity of using triggers based on skin conductance to sample arousing events within an experience sampling framework. *European Journal of Personality*.. https://doi.org/10.1002/per.2252.

Vansteelandt, K., Van Mechelen, I., & Nezlek, J. B. (2005). The co-occurrence of emotions in daily life: A multilevel approach. *Journal of Research in Personality, 39*(3), 325–335. https://doi.org/10.1016/j.jrp.2004.05.006.

Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2019). Lexicosemantic, affective, and distributional norms for 1,000 Dutch adjectives. *Behavior Research Methods, 52*, 1108–1121. https://doi.org/10.3758/s13428-019-01303-4.

Viechtbauer, W., & Constantin, M. (2019). Esmpack: Functions that facilitate preparation and management of ESM/EMA data. R package version 0.1-14.

Viechtbauer, W., Lataster, T., Rintala, A., Simons, C., Delespaul, P., Wichers, M., & Myin- Germeys, I. (2020). Evidence for a two-factor positive and negative affect structure in daily life: The Maastricht Momentary Mood Questionnaire (3MQ). Manuscript in preparation.

Vogelsmeier, L. V. D. E., Vermunt, J. K., Bülow, A., & De Roover, K. (2020). Evaluating covariate effects on ESM measurement model changes with latent Markov factor analysis: A three-step approach. https://doi.org/10.31234/osf.io/6ufrc.

Vogelsmeier, L. V., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(4), 557–575. https://doi.org/10.1080/10705511.2018.1554445.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037//0022- 3514.54.6.1063.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*(2), 219–235. https://doi.org/10.1037/0033-2909.98.2.219.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology, 76*(5), 820–838. https://doi.org/10.1037/0022-3514.76.5.820.

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.

Zelenski, J. M., & Larsen, R. J. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality, 34*(2), 178–197. https://doi.org/10.1006/jrpe.1999.2275.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43*(1), 111.