**Methods**

Statistical Analyses

We observed a negative variance on the between-level for item 14 in all preregistered analyses, which is common when a factor has a low number of indicators. Therefore, we constrained the loadings of the two indicators of the clarity factor to equality on the between-level. As a result, the same equality constraint applies to the within-level in models assuming cross-level invariance (i.e., equal loadings between and within individuals). Models were estimated using the MLR estimator because of the deviations from normality.

Dealing with Heywood cases: https://staskolenikov.net/papers/heywood-12.pdf

Level-specific fit based on the approach by Ryu and West (2009)!

Multilevel cross-level invariance: <https://www.tandfonline.com/doi/full/10.1080/10705511.2018.1534205>

Different number of factors occur in one third of models: Kim, Dedrick, Cao, and Ferron (Citation2016)

Use reporting guidelines: <https://www.tandfonline.com/doi/full/10.1080/00273171.2016.1228042>

* Good section on cross-level invariance! (?No cross-level invariance means unequal factor loadings across individuals? )

McNeish (2018) The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models -> Fit measures always indictae bad fit with very high loadings and minor misspecification

**Results**

**Item Descriptives**

Descriptive statistics for all items can be found in tables SX-SX. The within- and between person correlations for the full sample are shown in figure X and contrasted with the cross-sectional correlations we previously found following a mood induction (). Two important observations can be made here: First, the items of the nonacceptance and the modulate scales form a more strongly interconnected cluster. Second, item 10 shows noticeably low correlations with all other items. After inspecting the item’s content, we found this discrepancy is psychologically very reasonable, as the English translation is “I believe that I will continue feeling this way for a long time”. Such an item makes sense to indicate emotion regulation difficulties when it appears in a negative context (e.g., a mood induction or after a substantial stressor), but in contextually unrestricted momentary assessments, this is not the case. Therefore, we exclude this item from all analyses presented in this paper.

**Psychometrics**

***Factor Structure (H1) and Questionnaire Short Forms.***

In the construction sample, the four factor model with cross-loadings and cross-level equivalence had acceptable model fit on the SRMR and RMSEA (table X). While CFI and TLI were slightly below the preregistered threshold, the fit of the independence “null” model, used as a reference by CFI and TLI, already fit the data relatively well, imposing a mathematical ceiling (rmsea = 0.14; <https://davidakenny.net/cm/fit.htm>). Hence, the preregistered CFI and TLI thresholds should not be applied. In the validation sample, the same pattern emerged (table X). Therefore, the four factor model with cross-loadings and cross-level equivalence is accepted. This implies that the same latent variables underlie emotion regulation difficulties on both the within- and the between-person level.

The factor loadings only partially conformed to the expected structure. Two items of the Modulate scale were assigned to the Non-Acceptance scale with relatively low loadings (Figure X, Table X). Therefore, these two items should not be used. The same pattern was found in the validation sample (Figure X, Table X). This suggest that daily life studies should not use the standard 21 item S-DERS, but rather an 18-item version (table X). Based on factor loadings in the construction sample and theoretical considerations, we determined an 8-item and a 4-item version of the questionnaire (table X). Details on how these items were determined can be found in the supplements.

Fit measures and loadings for the full combined sample can be found in tables SX/SX.

***Convergent and discriminant validity???***

Notably, the correlation of the nonacceptance and the modulate factor were highly correlated on both the within- and between-person level. [hier könnte man auch die Cross-loadings und ggf. Netzwerkanalysen diskutieren]

***Variability within vs between individuals (H1).*** *MOVE THIS TO LATER POSITION*

With values mostly over 50%, the ICCs indicated that most variance in the S-DERS is between individuals, rather than within individuals (Total Score: 0.60 [0.56, 0.65], Nonacceptance: 0.52 [0.47, 0.57], Modulate: 0.49 [0.44, 0.54], Awareness: 0.60 [0.56, 0.65], Clarity: 0.50 [0.46, 0.55]. Item-wise descriptives statistics including ICCs can be found in table SX and show similar ratios. Hence, multilevel analyses are necessary for modelling the S-DERS in daily life. Such analyses can further elucidate which proportion of within- vs between-person variance is due to noise.

***Short Form Construction***

First, we inspected the factor loadings in the construction sample (figure SX, table SX). The final 4- and 8-item versions are marked in table X.

For the Non-Acceptance scale, creating a short form is most complex, as all items of the original S-DERS scale are good candidates from a statistical standpoint. Item 5 is the only item that did not belong to the non-acceptance scale in the trait DERS and we therefore eliminated this item from the short-version pool. All remaining items capture meta-affective processes, i.e., how a person feels about their emotions. Item 12 and 18 showed the highest loadings and could therefore be preferable from a statistical standpoint. Also, for psychological reasons, we suggest that the best 4-item measure of this scale might be item 18 (“I am irritated with myself for feeling this way”), as “irritation” is more likely to capture general negative affective components, while other items target more specific emotions (e.g., guilt [item 1], embarrassment [item 4], shame [item 8], anger [12]). We additionally suggest item 12 for the 8-item scale, as it has a similar loading pattern as item 18, which should lead to higher reliability. See the discussion section “recommendations for practical use” for further considerations.

For the modulate scale, only items 3 and 17 showed good loadings on the factor, as well as very low cross-loadings (|all| < .10) and good face validity. Hence, we suggest them for the 8-item version. For a 4-item version, both are possible, but we prefer the phrasing of item 17, as it is a little more specific and might be easier understood (e.g., in younger clinical samples).

For the awareness scale, all but item 19 are similarly sufficient candidates from a statistical standpoint. Item 16 might be suboptimal for momentary assessment settings, as “taking time to figure out what I am really feeling” might sometimes just not be possible in daily life for purely practical reasons and therefore introduce more complex context dependency. Item 2 is our suggestion for the 4-item scale, as it has the best loadings and captures the psychological construct of “awareness” with most face validity. Item 11 expresses more of a valuation than awareness. Therefore, we suggest item 6 as the second item for the 8-item version.

For the clarity scale, we suggest item 7 for the 4-item version, as item 14 had a considerable cross-loading on the non-acceptance factor. For an 8-item version, of course, both items must be used.

***Reliability (H3)***

Table X shows the reliabilities of the scales on the within-person level in the validation sample. For the full scale, the results from the mood induction studies is largely replicated: nonacceptance and modulate show excellent reliabilities, while awareness falls off and clarity is relatively low. Naturally, these reliabilities decrease further for the short versions, with very low reliability for the 1-item clarity scale. The between-person reliabilities are shown in figure X as a function of the average number of timepoints per person, which will differ from study to study. For the 19-item and the 8-item version, already 10 questionnaires per person lead to good reliabilities (>.80). Excellent reliabilities are found starting around 20 onwards (>.90). As an exception, the clarity scale generally had between-person reliabilities below .90 and remained below .80 for the 4-item version, regardless of how many time points were included.

Our full version from daily life showed better convergent and discriminant validity than the Lavender version in the hold-out sample (table SX). The correlations between versions was (XXX).

ENTSCHEIDUNGSDIAGRAMM ERSTELLEN

**Stress-Affect Relations (H4-H7)**

**Discussion**

Short form: Specific meta-emotions might be of interest! We checked with clinicians from children and youth psychiatry. Constructed with discriminant validity in mind, which reduces the reliability of the overall scale. The total scale reliability is better, when different scales are “contaminated” with shared variance. S-DERS merges three factor into one “modulate” factor. Our short version items both belong to the “impulse” factor. Network perspective might demand more items. Reliability oft two factors fort he 2-item version is really weak. But might be an underestimate, because it might just mean these constructs are heterogenious. Vielleicht shinyapp für die Berechnung der Reliabilität verschiedener Short-forms?

**Discussion**

-The S-DERS has context dependent information!

-Nonacceptance and Modulate are relatively muddled as factors!

-Modulate is already relatively incoherent in the original version