Supplementary Material: Assessing the Internal Consistency Reliability of Ecological Momentary Assessment Measures: Insights from the WARN-D study

Sebastian Castro-Alvarez Di Jody Zhou Laura F. Bringmann
Rayyan Tutunji Ricarda K. K. Proppert Carlotta L. Rieble
Eiko I. Fried Siwei Liu

This supplementary material of the article titled "Assessing the Internal Consistency Reliability of Ecological Momentary Assessment Measures: Insights from the WARN-D study" presents the detailed analyses conducted in the study along with the corresponding `R` code.

In this study, we estimated the reliability of the different scales of questionnaires measured during the stage 2 of the WARN-D study. The questionnaire data were collected through the weekly measurements on depression and general anxiety disorder, and the ESM measurements on positive and negative affect. For the weekly measurements, the reliability coefficients were estimated based on the generalizability theory approach (Cranford et al., 2006), the multilevel modeling approach (Nezlek, 2017), and the multilevel confirmatory factor analysis model (Geldhof et al., 2014; Lai, 2021). For the ESM measurements, in addition to the previous approaches, we also estimated the reliability coefficients by means of the p-technique factor analysis (Hu et al., 2016), the dynamic multilevel factor model (Xiao et al., 2023), and the latent state-trait theory (Castro-Alvarez et al., 2022).

To conduct these analyses, the following packages are required: `psych`, `misty`, `lavaan`, `MplusAutomation`, `lme4`, `dplyr`, and `esmpack`. Additionally, we wrote several custom functions which are available in OSF.

WARN-D Data

The WARN-D study(Fried et al., 2023) is a multi-stage multi-cohort longitudinal study, which aims to develop an early warning system to prevent depression in university students in the

Netherlands. The study includes four stages: Baseline assessments (Stage 1), 3-month experience sampling design (Stage 2), follow-ups every 3 months for 21 months (Stage 3), and a second 3-month experience sampling design (Stage 4). The data we analyzed in the present study was from a subsample collected during stage 2, wherein approximately 30% of the sample is on holdout to test predictive models and to implement cross-validation techniques(Tutunji et al., 2024). Participants were allowed to choose between Dutch and English as their language during the study.

Stage 2 of the WARN-D study included self-report questionnaires and passive wearable measures. Self-report questionnaires for the daily experiences were administered four times a day in time intervals of 225 minutes on average. In these questionnaires, participants reported the following 10 emotions on a 7-point Likert scale: Sad, stressed, overwhelmed, nervous, ruminate, irritable, tired, cheerful, motivated, and relaxed. In addition, participants were asked to fill weekly questionnaires, which were administered every Sunday. The weekly questionnaires include adapted versions of the 9-item Patient Health Questionnaire (PHQ-9, Kroenke & Spitzer, 2002) as well as the 7-item measure to assess generalized anxiety disorder (GAD-7, Spitzer et al., 2006), both were on a 4-point Likert scale.

For the analyses presented in this study, we split the sample by language and analyzed the data from each subsample separately. In the process of prepossessing and cleaning the raw data (this `R` script is available in OSF), we removed observations with total or partial missingness as well as observations of the less-used language for participants who used both languages to fill in their questionnaires. The raw data of the WARN-D project will become openly available in the future. In the mean time, we shared in OSF the IDs of the participants included in this study. We loaded the cleaned data into `R` as `data.frame` objects named `warn_d_ema` for the ESM data, and `warn_d_week` for the weekly data.

In what follows, we present the procedure to estimate the reliability based on different approaches with detailed code. First, we present the results for the weekly questionnaires followed by the results of the ESM questionnaires.

Reliability for weekly questionnaires

Participants completed the two weekly questionnaires (i.e., PHQ-9 and the GAD-7) on Sundays for up to 12 weeks. In the WARN-D study, some questions of the PHQ-9 were split into two questions. For example, the question "Trouble falling or staying asleep, or sleeping too much" was split into two questions: One that indicates insomnia and other that indicates hypersomnia. For our analyses, we only considered the one with the highest response score. Furthermore, the GAD-7 was reduced to five items because two items overlapped with other questions and were thus repetitive.

To assess the reliability of the PHQ-9 and the GAD-7, we used the following approaches: Generalizability theory (Cranford et al., 2006), the multilevel modeling (Nezlek, 2017), and

the multilevel confirmatory factor analysis (Geldhof et al., 2014; Lai, 2021). These approaches account for the nested structure of the data and estimate at least one reliability coefficient for each level (within- and between-level). Other approaches such as p-technique factor analysis (Hu et al., 2016) were not considered for these questionnaires because weekly questionnaires did not yield enough measurement occasions per participant.

Weekly data description.

Before computing the reliability for each scale, we briefly described the data by computing the compliance rate per participant and plotting the distribution of each item. In this section, we also split the data by language.

To easily manipulate the data in later steps, we created two vectors with the variable names for the items of each scale.

```
phq <- grep("phq", names(warn_d_week), value = TRUE)
gad <- grep("gad", names(warn_d_week), value = TRUE)
```

With the following code, we computed the compliance rate per participant over the whole sample.

```
compliance_week <- esmpack::calc.nomiss(id, id, warn_d_week)
```

Generally, participants completed most of the weekly assessments. The median compliance was 10, with a minimum of 1, and a maximum of 12. Figure S1 presents the distribution of the compliance rate per participant for the weekly questionnaires.

As mentioned before, in this study, we report the results based on the language used to fill in the questionnaires. For this reason, we split the data and compute the compliance rate per participant by language.

```
warn_d_week_en <- warn_d_week[warn_d_week$lang == "en", ]
warn_d_week_nl <- warn_d_week[warn_d_week$lang == "nl", ]

compliance_week_en <- esmpack::calc.nomiss(id, id, warn_d_week_en)
compliance_week_nl <- esmpack::calc.nomiss(id, id, warn_d_week_nl)
```

We also plotted the distribution of the compliance rate for the weekly questionnaires by language as shown in Figure S2. Considering that the Dutch subsample is smaller than the English subsample, we do not think there is any major difference between the two language subsamples.

Figure S1: Distribution of the compliance for the weekly questionaires

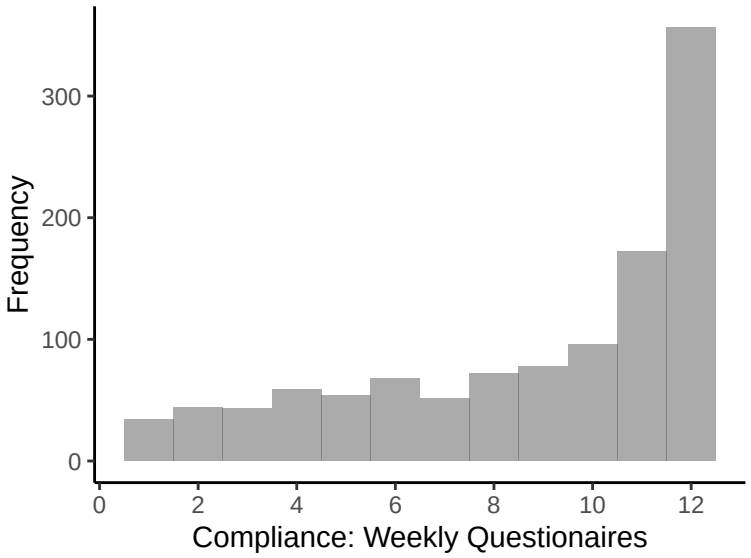
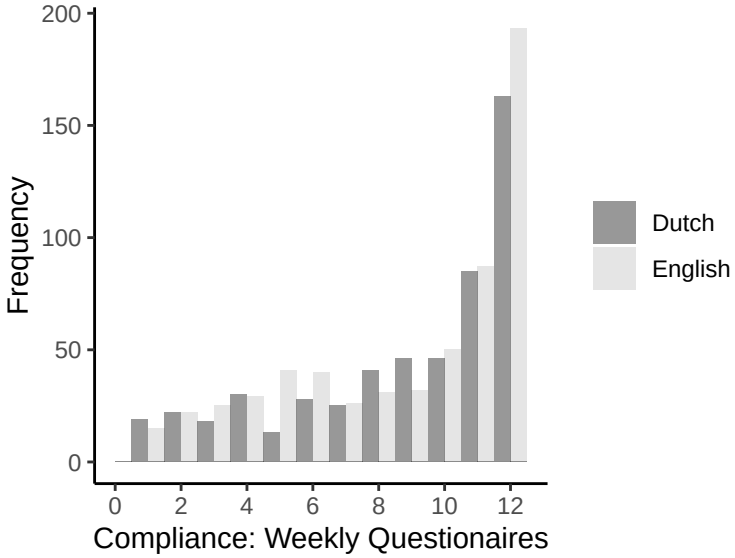


Figure S2: Distribution of the compliance for the weekly questionaires by language



To plot the distribution per item given the language used, we reshaped the data into long format:

```
warn_d_week_en_long <- reshape(warn_d_week_en,
                               direction = "long",
                               idvar = c("id", "week_num"),
                               timevar = "item",
                               times = c(phq, gad),
                               varying = list(c(phq, gad)),
                               v.names = "resp")

warn_d_week_nl_long <- reshape(warn_d_week_nl,
                               direction = "long",
                               idvar = c("id", "week_num"),
                               timevar = "item",
                               times = c(phq, gad),
                               varying = list(c(phq, gad)),
                               v.names = "resp")
```

Figure S3 and Figure S4 present the distribution of the five items of the GAD-7 and the nine items of the PHQ-9 for the English and the Dutch subsamples, respectively. Overall, the items' distributions across the two subsamples tend to be positively skewed. Notice that there seem to be some floor effects for items *gad7*, *phq8*, and *phq9* in both languages.

Figure S3: Distribution of the responses per item for the weekly questionaires in English

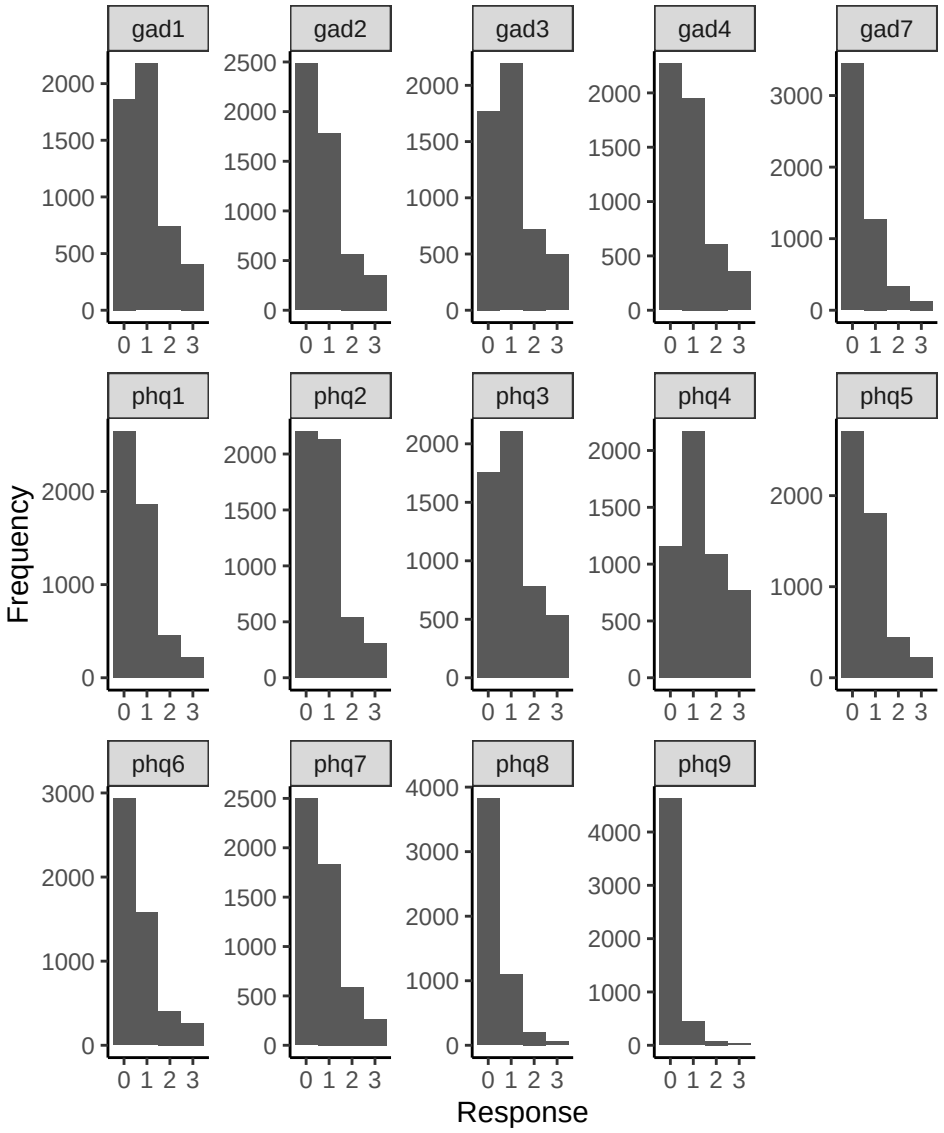
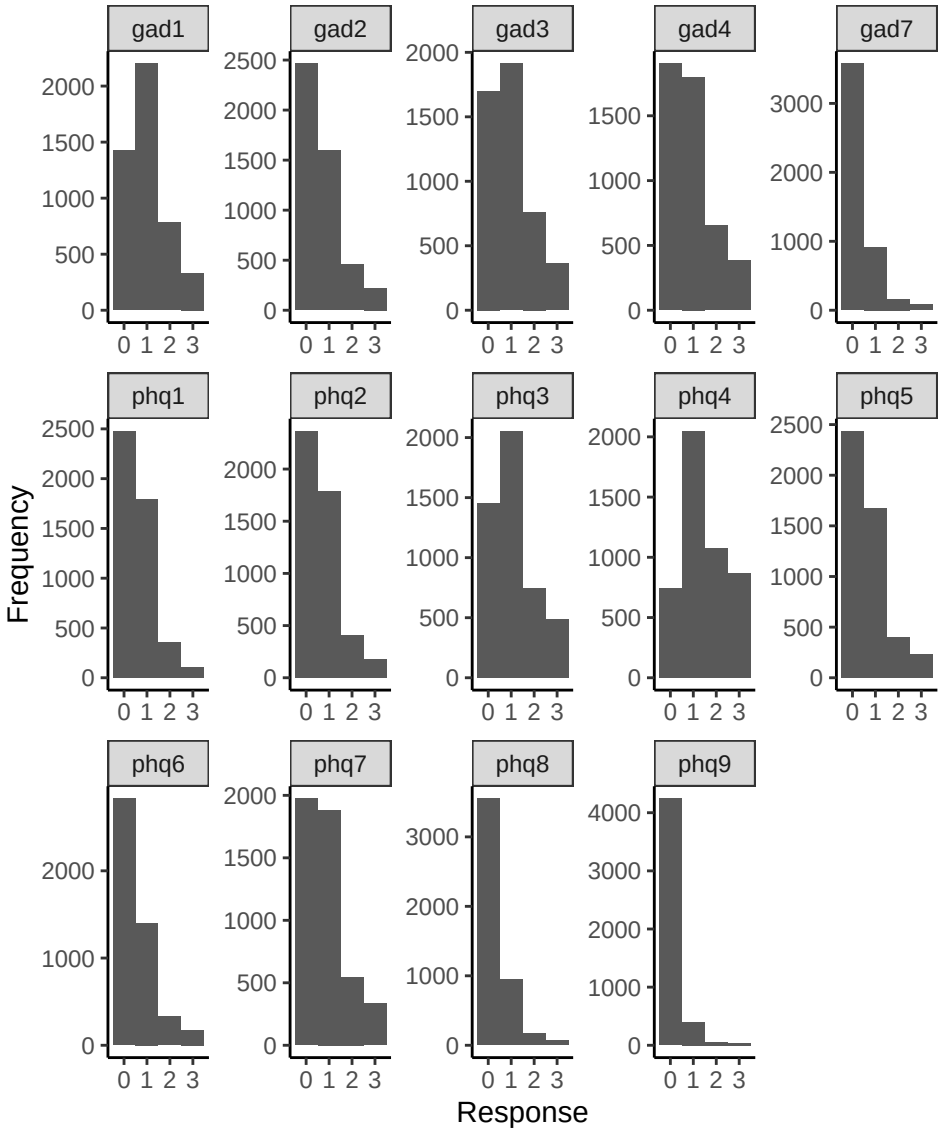


Figure S4: Distribution of the responses per item for the weekly questionaires in Dutch



Generalizability theory.

The first approach was suggested by Cranford et al. (2006), which is encompassed within the generalizability theory framework. In this approach, the goal is to identify different sources of variance and error. In longitudinal data, these sources are the persons, the measurement occasions (weeks), the items, and their two-way interactions. Then, the estimated variances are used to compute different reliability coefficients. For these analyses, we focus on the R_{kr} and R_c, which are the reliability of the average taken over K random occasions and the reliability of change, respectively. R_{kr} can be interpreted as the reliability at the between-level, and R_c can be interpreted as the reliability at the within-level.

This procedure is implemented in the function `mlr` of the `psych` package. In the following code, we used the `mlr` function to compute the R_{kr} and R_c for the PHQ-9 and the GAD-7 given the language used by the participants.

```
# GT PHQ English
mlr_phq_en <- psych::mlr(warn_d_week_en,
                         grp   = "id",
                         Time  = "week_num",
                         items = phq,
                         lmer  = TRUE,
                         lme   = FALSE,
                         alpha = FALSE,
                         aov   = FALSE)

# GT PHQ Dutch
mlr_phq_nl <- psych::mlr(warn_d_week_nl,
                         grp   = "id",
                         Time  = "week_num",
                         items = phq,
                         lmer  = TRUE,
                         lme   = FALSE,
                         alpha = FALSE,
                         aov   = FALSE)

# GT GAD English
mlr_gad_en <- psych::mlr(warn_d_week_en,
                         grp   = "id",
                         Time  = "week_num",
                         items = gad,
                         lmer  = TRUE,
                         lme   = FALSE,
                         alpha = FALSE,
```

Table S1: Reliability coefficients of the weekly questionnaires given the generalizability theory.

	PHQ-9		GAD-7	
	EN	NL	EN	NL
R_{kr}	0.96	0.97	0.95	0.95
R_c	0.70	0.64	0.77	0.69

```
                          aov   = FALSE)

# GT GAD Dutch
mlr_gad_nl <- psych::mlr(warn_d_week_nl,
                          grp   = "id",
                          Time  = "week_num",
                          items = gad,
                          lmer  = TRUE,
                          lme   = FALSE,
                          alpha = FALSE,
                          aov   = FALSE)
```

The estimated R_{kr} and R_c for each weekly questionnaire and subsample are presented in Table S1. Overall, R_{kr} was very high for both questionnaires and subsamples. It varied between 0.95 and 0.97. R_c was generally lower and it tended to be lower for the Dutch subsample in comparison to the English subsample. It varied between 0.64 and 0.77.

Multilevel modeling.

Another approach similar to the generalizability theory is to estimate the reliability based on multilevel modeling techniques (Nezlek, 2017). The idea in this procedure is to fit an empty three-level multilevel model to the data in which responses to the items are nested within occasions, which are nested within persons. Then, the estimated variances are used to compute two reliability coefficients: R_{krn} and R_{cn}. R_{krn} is the reliability at the between-level, which indicates how reliable the persons' means are. R_{cn} is the reliability at the within-level, which indicates how consistent are the responses to the items within occasions and persons. According to Nezlek (2017), R_{cn} is equivalent to Cronbach's alpha but accounting for the nested structure in the data.

To estimate the reliability coefficients based on Nezlek (2017) procedure, we also used the function `mlr` of the `psych` package. Hence, we simply extracted the results from the `mlr` objects fitted in the previous section. Table S2 presents the estimated R_{krn} and R_{cn} for each weekly questionnaire and each language. The between-level reliability is quite similar to the estimated coefficients of the generalizability theory and varies between 0.95 and 0.97.

Table S2: Reliability coefficients of the weekly questionnaires given the three-level null multilevel model.

	PHQ-9		GAD-7	
	EN	NL	EN	NL
R_{krn}	0.96	0.97	0.95	0.95
R_{cn}	0.45	0.22	0.65	0.46

In contrast, the within-level reliability coefficients are much lower when compared with the R_c in the generalizability theory. Furthermore, there are clear differences for the R_{cn} across questionnaire and language. The R_{cn} was lower for the PHQ-9 than for the GAD-7, and it was also lower for the Dutch subsample than for the English subsample.

Multilevel confirmatory factor analysis.

The last approach that we used to estimate the reliability of the weekly questionnaires is the multilevel confirmatory factor analysis (Geldhof et al., 2014; Lai, 2021). In this procedure, one fits a multilevel confirmatory factor model (with equal factor loadings across levels) to the data and computes three McDonald's omega reliability coefficients as defined by Lai (2021). Here, we focused on two of these coefficients, the between-level omega (ω^b) and the within-level omega (ω^w).

The procedure as presented by Lai (2021) has been implemented in the function `multilevel.omega` of the package `misty`. The following code fits the multilevel model and estimates the reliability coefficients for each weekly questionnaire and language.

```
# ML-CFA PHQ English
mlcfa_phq_en <- misty::multilevel.omega(
  warn_d_week_en[, phq],
  cluster = warn_d_week_en$id,
  missing = "listwise")

# ML-CFA PHQ Dutch
mlcfa_phq_nl <- misty::multilevel.omega(
  warn_d_week_nl[, phq],
  cluster = warn_d_week_nl$id,
  missing = "listwise")

# ML-CFA GAD English
mlcfa_gad_en <- misty::multilevel.omega(
  warn_d_week_en[, gad],
  cluster = warn_d_week_en$id,
```

Table S3: Reliability coefficients of the weekly questionnaires given multilevel confirmatory factor analysis.

	PHQ-9		GAD-7	
	EN	NL	EN	NL
ω^b	0.84	0.83	0.86	0.84
ω^w	0.73	0.66	0.78	0.72
CFI	0.91	0.87	0.99	0.96
TLI	0.89	0.85	0.98	0.94
RMSEA	0.05	0.05	0.04	0.05

```
  missing = "listwise")

# ML-CFA GAD Dutch
mlcfa_gad_nl <- misty::multilevel.omega(
  warn_d_week_nl[, gad],
  cluster = warn_d_week_nl$id,
  missing = "listwise")
```

The estimated between- and within-omegas as well as a few fit indices for each model are presented in Table S3. Overall, ω^b was high and very similar across weekly questionnaires and samples. It varied between 0.84 and 0.84. ω^w was lower, varying between 0.73 and 0.73. Also, ω^w of the English subsample tended to be higher than ω^w of the Dutch subsample.

Reliability for EMA questionnaires

During the EMA portion of the study, participants reported how they were feeling up to four times a day for 85 days (for a maximum of 340 measurements). The assessed emotions were sad, stressed, overwhelmed, nervous, ruminate, irritable, tired, cheerful, motivated, and relaxed.The first seven aimed to measure **negative affect** and the last three aimed to measure **positive affect**. Here, we analyze each set of items separately.

To assess the reliability of the EMA items, we used six different approaches. Firstly, we used the same three methods used to assess the reliability of the weekly questionnaires (i.e., generalizability theory, multilevel modeling, and multilevel confirmatory factor analysis). Secondly, we considered three additional approaches, which especially designed for intensive longitudinal data. These approaches include the p-technique factor analysis (Hu et al., 2016), the dynamic multilevel factor model (Xiao et al., 2023), and the latent state-trait theory (Castro-Alvarez et al., 2022).

EMA data description.

As similarly done with the weekly questionnaires, we first explore the data by looking at the compliance rate and the distribution of the responses per item.

We created the following two vectors to easily select the items of interest in this section.

```
pa <- c("cheer", "motiv", "relax")
na <- c("sad", "stres", "overw", "nervo", "rumin", "irrit", "tired")
```

Next, we computed the compliance rate per participant:

```
compliance_ema <- esmpack::calc.nomiss(id, id, warn_d_ema)
```

Overall, the median number of completed measurements were 187, with a minimum of 1, and a maximum of 338. Figure S5 presents the distribution of the compliance over the whole sample. The distribution is relatively flat with a peak around 260 measurements.

For the EMA questionnaires, we also separated the data by language. We split the data and computed the compliance for each subsample as follows:

```
warn_d_ema_en <- warn_d_ema[warn_d_ema$lang == "en", ]
warn_d_ema_nl <- warn_d_ema[warn_d_ema$lang == "nl", ]

compliance_ema_en <- esmpack::calc.nomiss(id, id, warn_d_ema_en)
compliance_ema_nl <- esmpack::calc.nomiss(id, id, warn_d_ema_nl)
```

Figure S5: Distribution of the compliance for the EMA questionaires

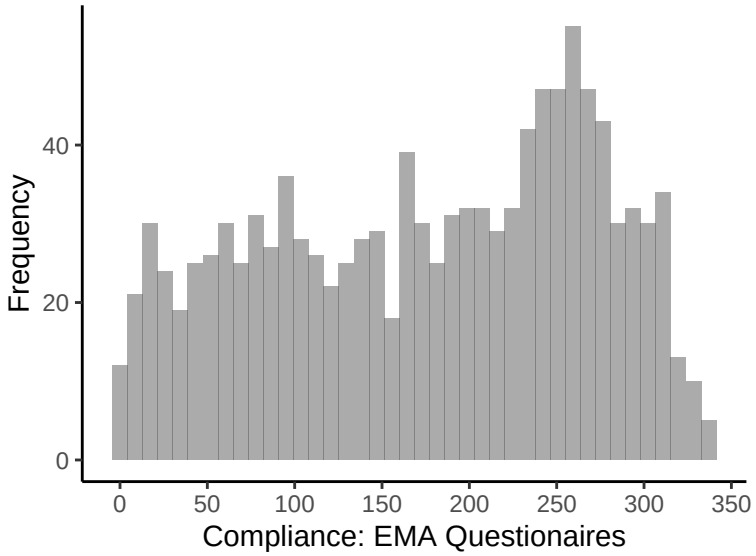


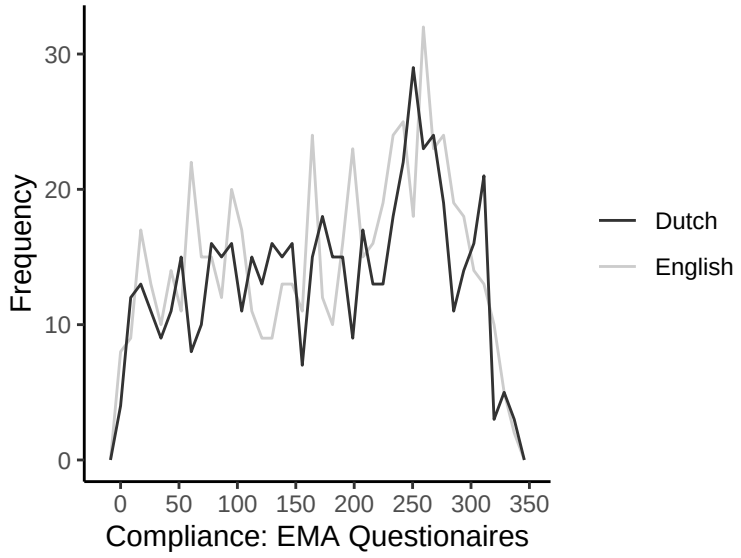
Figure S6 presents the distribution of the compliance by language. It does not seem that the compliance to the planned measurements was affected by the preferred language of the participants. For the English subsample, the median compliance was 190 measurements with a minimum of 1 and a maximum of 337. For the English subsample, the median compliance was 185 measurements with a minimum of 1 and a maximum of 338.

Lastly, we also explored the distribution of the responses to each emotion across the two subsamples. To do this, we first structured the data into long format with the following code:

```
warn_d_ema_en_long <- reshape(warn_d_ema_en,
                              direction = "long",
                              idvar = c("id", "bn"),
                              timevar = "item",
                              times = c(na, pa),
                              varying = list(c(na, pa)),
                              v.names = "resp")

warn_d_ema_nl_long <- reshape(warn_d_ema_nl,
                              direction = "long",
                              idvar = c("id", "bn"),
                              timevar = "item",
                              times = c(na, pa),
                              varying = list(c(na, pa)),
```

Figure S6: Distribution of the compliance for the EMA questionaires by language




```
                                    v.names = "resp")
```

The distributions of the responses per item and for each subsample are shown in Figure S7 and Figure S8. Generally speaking, the items of positive affect tend to have a symmetric distribution while the items of negative affect tend to be positively skewed. Interestingly, the distribution of the item tired, which measures negative affect, is more symmetric in both samples. This might be explained by considering that feeling tired can also be influenced by the time of the day at which participants completed the questionnaires. Naturally, participants will report lower levels of tiredness in the first measurements of the day a higher level by the end of the day.

Figure S7: Distribution of the responses per item for the EMA questionaires in English

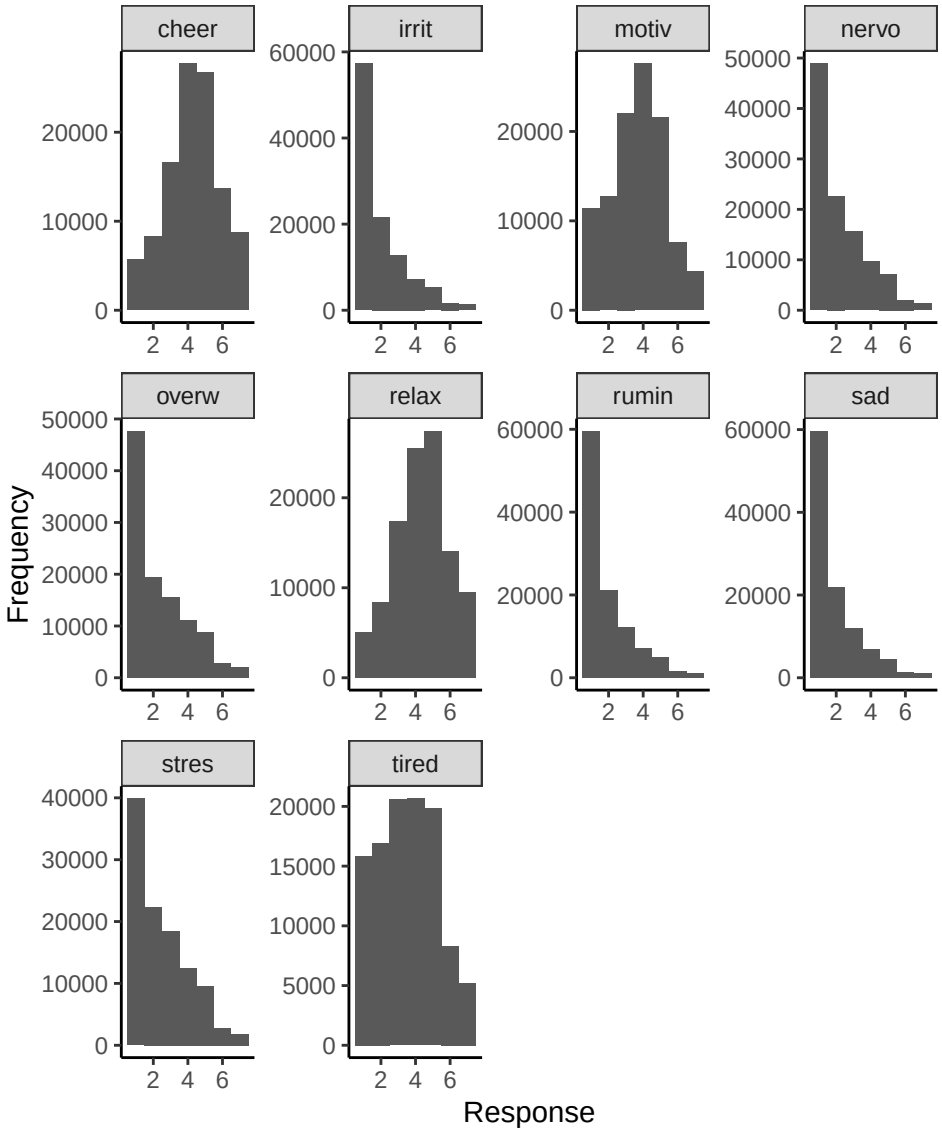


Figure S8: Distribution of the responses per item for the EMA questionaires in Dutch

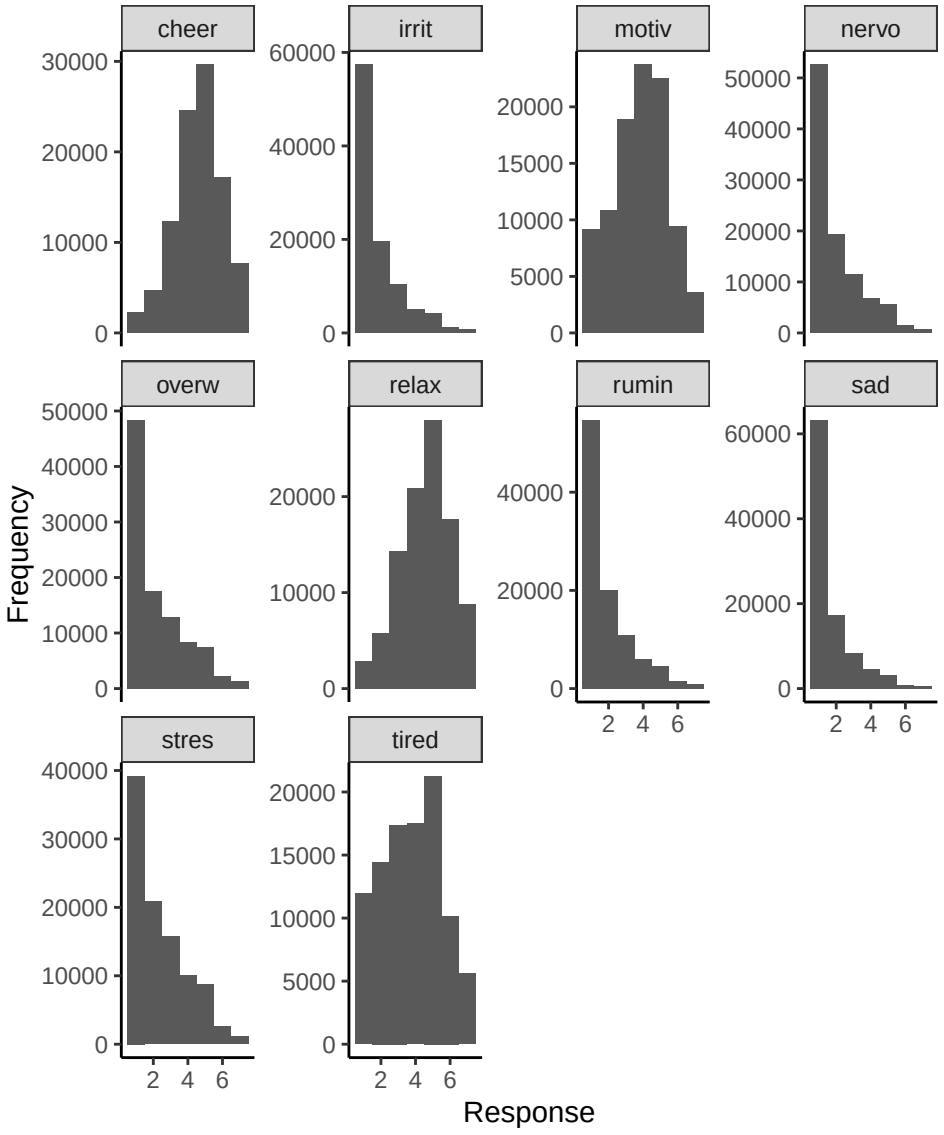


Table S4: Reliability coefficients of the ema data given the generalizability theory.

	PA		NA	
	EN	NL	EN	NL
R_{kr}	1.00	1.00	1.00	1.00
R_c	0.63	0.61	0.79	0.74

Generalizability theory.

Similarly to the weekly data, we calculate reliability of the EMA data following the generalizability theory framework (Cranford et al., 2006). For these analyses, We used the estimated variances, including persons, the measurement occasions (beeps), the items (7 negative affect and 3 positive affect items), and their two-way interactions.) to compute different reliability coefficients. we focus on the R_{kr}, which is the reliability at the between-level, and R_c, which is the reliability at the within-level. In the following syntax, we used the function `mlr` of the `psych` package to compute the R_{kr} and R_c for each language subsample.

```
mlr_en_pa <- psych::mlr(warn_d_ema_en, grp = "id", Time = "bn", items = pa,
lmer = TRUE, lme = FALSE, alpha = FALSE, aov = FALSE)

mlr_en_na <- psych::mlr(warn_d_ema_en, grp = "id", Time = "bn", items = na,
lmer = TRUE, lme = FALSE, alpha = FALSE, aov = FALSE)
# Warning message: Model failed to converge with max|grad| = 0.00494617
# (tol = 0.002, component 1)

mlr_nl_pa <- psych::mlr(warn_d_ema_nl, grp = "id", Time = "bn", items = pa,
lmer = TRUE, lme = FALSE, alpha = FALSE, aov = FALSE)
# Warning: Model failed to converge with max|grad| = 0.00224557
# (tol = 0.002, component 1)


mlr_nl_na <- psych::mlr(warn_d_ema_nl, grp = "id", Time = "bn", items = na,
lmer = TRUE, lme = FALSE, alpha = FALSE, aov = FALSE)
# Warning: unable to evaluate scaled gradientWarning: Model failed to converge:
# degenerate  Hessian with 1 negative eigenvalues
```

The estimated R_{kr} and R_c are presented in Table S4. Overall, R_{kr} was very high for both the positive and negative affect and for both subsamples. It varied between 1 and 1. R_c was generally lower and it tended to be lower for the positive affect in comparison to the negative affect. It varied between 0.61 and 0.79.

Table S5: Reliability coefficients of the ema data given the three-level null multilevel model.

	PA		NA	
	EN	NL	EN	NL
R_{krn}	1.0	1.00	1.00	1.00
R_{cn}	0.5	0.43	0.68	0.53

Multilevel modeling.

The second approach we used to estimate the reliability in ESM data is to use the multilevel modeling framework following (Nezlek, 2017). The idea is to estimate variance by fitting an empty three-level multilevel model wherein item nested within measurement occasions nested within persons to compute two reliability coefficient We then utilized these estimated variances compute two reliability coefficients: R_{krn} which is the between-level reliability, and R_{cn}, which is the within-level reliability. This procedure can be automatically executed using the same function `mlr` of the `psych` package as we used to estimate the generalizability theory approach. Hence, we simply extracted the results from the `mlr` objects fitted in the previous section.

Table S5 presents the estimated R_{krn} and R_{cn} for each language. Similarly to what we found in the weekly questionnaire dataset, the between-level reliability is quite similar to the estimated coefficients of the generalizability theory and varies between 1 and 1. Also similarly to the weekly questionnaire dataset, the within-level reliability coefficients are much lower when compared with the R_c in the generalizability theory. Furthermore, the R_{cn} was lower for the positive affect compared to the negative affect, and it was also lower for the Dutch subsample than for the English subsample.

Multilevel confirmatory factor analysis.

Similar to the weekly data, We also used the the function `multilevel.omega` of the package `misty` to compute the reliability through the multilevel confirmatory factor analysis (Geldhof et al., 2014; Lai, 2021). In the procedure, we fit separate multilevel confirmatory factor model for each language and for PA and NA, and then compute the between-level omega (ω^b) and the within-level omega (ω^w) McDonald's omega reliability coefficients as defined by Lai (2021).

```
mlcfa_en_pa <- misty::multilevel.omega(
  warn_d_ema_en[,pa],
  cluster = warn_d_ema_en$id,
  missing = "listwise")

mlcfa_en_na <- misty::multilevel.omega(
  warn_d_ema_en[,na],
  cluster = warn_d_ema_en$id,
```

Table S6: Reliability coefficients of the EMA questionnaires given multilevel confirmatory factor analysis.

	PA		NA	
	EN	NL	EN	NL
ω^b	0.75	0.71	0.93	0.93
ω^w	0.68	0.66	0.80	0.75
CFI	1.00	1.00	0.89	0.87
TLI	0.99	0.99	0.87	0.84
RMSEA	0.03	0.02	0.08	0.08

```
  missing = "listwise")

mlcfa_nl_pa <- misty::multilevel.omega(
  warn_d_ema_nl[,pa],
  cluster = warn_d_ema_nl$id,
  missing = "listwise")

mlcfa_nl_na <- misty::multilevel.omega(
  warn_d_ema_nl[,na],
  cluster = warn_d_ema_nl$id,
  missing = "listwise")
```

```
mlcfa_en_pa_v02 <- misty::multilevel.omega(
  warn_d_ema_en[,pa],
  cluster = warn_d_ema_en$id,
  fix.resid = c("all"),
  missing = "listwise")

mlcfa_nl_pa_v02 <- misty::multilevel.omega(
  warn_d_ema_nl[,pa],
  cluster = warn_d_ema_nl$id,
  fix.resid = c("all"),
  missing = "listwise")
```

The estimated between- and within-omegas as well as a few fit indices for each model of the EMA data are presented in Table S6.

Overall, ω^b was similar across the two language subsamples for both PA and NA. ω^w was slightly lower than ω^b. Moreover, both of the ω^w and ω^b of the English subsample tended to be higher than those of the Dutch subsample.

P-technique factor analysis.

By having plenty of time points per participants, we can also use other methods that are appropriate to estimate the reliability of time series data. One of the approaches that allows to do this is the p-technique factor model. In this model, one fits a confirmatory factor model for each participant and computes McDonald's omega based on the estimated parameters as illustrated by Hu et al. (2016). For this and the following approaches, we decided to only include participants that completed at least 100 of the planned measurements. Thus, in the following code, we create a new `data.frame` for each language selecting the participants who satisfy our criteria.

```
ids_excl_en <- names(which(compliance_ema_en < 100))
ids_excl_nl <- names(which(compliance_ema_nl < 100))

warn_d_ema_en_100 <- warn_d_ema_en[!(warn_d_ema_en$id %in% ids_excl_en), ]
warn_d_ema_nl_100 <- warn_d_ema_nl[!(warn_d_ema_nl$id %in% ids_excl_nl), ]
```

By doing this, 166 and 140 were excluded from the English and Dutch samples, respectively. Also, for the following analysis, we excluded participants that show no variability in one or more emotion items throughout the study.

```
nullvar_excl_en <- apply(
  apply(warn_d_ema_en_100[, c(pa, na)], 2,
        function(xx) ave(xx, warn_d_ema_en_100$id, FUN = sd)), 1,
  function(x) all(x > 0))

nullvar_excl_nl <- apply(
  apply(warn_d_ema_nl_100[, c(pa, na)], 2,
        function(xx) ave(xx, warn_d_ema_nl_100$id, FUN = sd)), 1,
  function(x) all(x > 0))

warn_d_ema_en_100 <- warn_d_ema_en_100[nullvar_excl_en, ]
warn_d_ema_nl_100 <- warn_d_ema_nl_100[nullvar_excl_nl, ]
```

This excluded 4 and 15 participants for the English and Dutch subsamples, respectively. Thus, the number of participants included in these analyses was 441 for the English subsample and 401 for the Dutch subsample.

To loop through all the participants of each set, we turn the data into a list, in which each element is the data from each participant.

```
ids_tmp <- unique(warn_d_ema_en_100$id_n)

warn_d_list_en <- list()

for (i in 1:length(ids_tmp)) {
warn_d_list_en[[i]] <- warn_d_ema_en_100[warn_d_ema_en_100$id_n == ids_tmp[i], ]
names(warn_d_list_en)[i] <- as.character(ids_tmp[i])
}
rm(i, ids_tmp)

ids_tmp <- unique(warn_d_ema_nl_100$id_n)

warn_d_list_nl <- list()

for (i in 1:length(ids_tmp)) {
warn_d_list_nl[[i]] <- warn_d_ema_nl_100[warn_d_ema_nl_100$id_n == ids_tmp[i], ]
names(warn_d_list_nl)[i] <- as.character(ids_tmp[i])
}
rm(i, ids_tmp)
```

Next, we imported the lavaan models into R. The models for each set of items were unidimensional. The factor model of the items of negative affect included the seven items of negative affect and was identified by fixing the variance of the latent factor at 1. The factor model of the items of positive affect included three items. Due to the few number of items, the model was unidentified when only fixing the variance of the latent factor or the first factor loading. For this reason, we defined a tau-equivalent model for the items of positive affect. This means that all the factor loadings were fixed at 1 and that the variance was freely estimated.

```
pfa_pa_model <- paste0(readLines("Lavaan/pa_ptechnique_tau.txt"),
                       collapse = "\n")
pfa_na_model <- paste0(readLines("Lavaan/na_ptechnique.txt"), collapse = "\n")
```

To perform the analyses, we looped through all the participants by language and construct (PA and NA). In these loops, we fitted the corresponding model per person, saved the `lavaan` object with the results. Then, we checked whether the model converged and finished successfully. Finally, we saved the estimated reliability coefficient for the participants for which the model converged.

PT-FA Loop: PA for the English subsample.

```
pfa_pa_en <- pfa_pa_en_omg <- list()

for (i in 1:length(warn_d_list_en)) {
  cat(sprintf("This is %d-th participant in the data list.\n", i))
  tmp_dat <- na.omit(warn_d_list_en[[i]][, pa])

  if (!any(eigen(cov(tmp_dat, use = "complete.obs"))$values <= 0) &
      all(diag(var(tmp_dat)) > 0)) {
    pfa_pa_en[[i]] <- lavaan::cfa(pfa_pa_model, tmp_dat)

    if (lavaan::lavInspect(pfa_pa_en[[i]], what = "converged") &
        lavaan::lavInspect(pfa_pa_en[[i]], what = "post.check")) {
      tmp_est <- lavaan::parameterEstimates(pfa_pa_en[[i]])
      pfa_pa_en_omg[[i]] <- tmp_est$est[tmp_est$label == "omega"]
    } else {
      pfa_pa_en_omg[[i]] <- NA
    }
  } else {
    pfa_pa_en[[i]] <- NULL
    pfa_pa_en_omg[[i]] <- NA
  }
}
```

PT-FA Loop: PA for the Dutch subsample.

```
pfa_pa_nl <- pfa_pa_nl_omg <- list()

for (i in 1:length(warn_d_list_nl)) {
  cat(sprintf("This is %d-th participant in the data list.\n", i))
  tmp_dat <- na.omit(warn_d_list_nl[[i]][, pa])

  if (!any(eigen(cov(tmp_dat, use = "complete.obs"))$values <= 0) &
      all(diag(var(tmp_dat)) > 0)) {
    pfa_pa_nl[[i]] <- lavaan::cfa(pfa_pa_model, tmp_dat)

    if (lavaan::lavInspect(pfa_pa_nl[[i]], what = "converged") &
        lavaan::lavInspect(pfa_pa_nl[[i]], what = "post.check")) {
      tmp_est <- lavaan::parameterEstimates(pfa_pa_nl[[i]])
```

```
      pfa_pa_nl_omg[[i]] <- tmp_est$est[tmp_est$label == "omega"]
    } else {
      pfa_pa_nl_omg[[i]] <- NA
    }
  } else {
    pfa_pa_nl[[i]] <- NULL
    pfa_pa_nl_omg[[i]] <- NA
  }
}
```

PT-FA Loop: NA for the English subsample.

```
pfa_na_en <- pfa_na_en_omg <- list()

for (i in 1:length(warn_d_list_en)) {
  cat(sprintf("This is %d-th participant in the data list.\n", i))
  tmp_dat <- na.omit(warn_d_list_en[[i]][, na])

  if (!any(eigen(cov(tmp_dat, use = "complete.obs"))$values <= 0) &
      all(diag(var(tmp_dat)) > 0)) {
    pfa_na_en[[i]] <- lavaan::cfa(pfa_na_model, tmp_dat)

    if (lavaan::lavInspect(pfa_na_en[[i]], what = "converged") &
        lavaan::lavInspect(pfa_na_en[[i]], what = "post.check")) {
      tmp_est <- lavaan::parameterEstimates(pfa_na_en[[i]])
      pfa_na_en_omg[[i]] <- tmp_est$est[tmp_est$label == "omega"]
    } else {
      pfa_na_en_omg[[i]] <- NA
    }
  } else {
    pfa_na_en[[i]] <- NULL
    pfa_na_en_omg[[i]] <- NA
  }
}
```

PT-FA Loop: NA for the Dutch subsample.

```r
pfa_na_nl <- pfa_na_nl_omg <- list()

for (i in 1:length(warn_d_list_nl)) {
  cat(sprintf("This is %d-th participant in the data list.\n", i))
  tmp_dat <- na.omit(warn_d_list_nl[[i]][, na])

  if (!any(eigen(cov(tmp_dat, use = "complete.obs"))$values <= 0) &
      all(diag(var(tmp_dat)) > 0)) {
    pfa_na_nl[[i]] <- lavaan::cfa(pfa_na_model, tmp_dat)

    if (lavaan::lavInspect(pfa_na_nl[[i]], what = "converged") &
        lavaan::lavInspect(pfa_na_nl[[i]], what = "post.check")) {
      tmp_est <- lavaan::parameterEstimates(pfa_na_nl[[i]])
      pfa_na_nl_omg[[i]] <- tmp_est$est[tmp_est$label == "omega"]
    } else {
      pfa_na_nl_omg[[i]] <- NA
    }
  } else {
    pfa_na_nl[[i]] <- NULL
    pfa_na_nl_omg[[i]] <- NA
  }
}
```

The p-technique factor model for PA converged to a proper solution for most participants but it did not for 0 in the English subsample and 3 in the Dutch subsample. There were more convergence issues or improper solutions for the p-technique factor model of NA. In this case, the model did not converged or resulted in an improper solution for 11 participants in the English subsample and 18 participants in the Dutch subsample. Regarding the estimated reliability coefficients for the participants for whom the models converged to a proper solution, the estimated coefficients were lower for PA than for NA. This can be explained by the number of items, as generally a shorter scale tends to have lower reliability than longer scales. Summary statistics for the estimated reliability coefficients are presented in Table S7. Furthermore, the distribution of the estimated reliability coefficients comparing by language are shown in Figure S9 and Figure S10.

Two-level random dynamic model-based approach.

We performed the analyses of dynamic multilevel factor model (2RDM, Xiao et al., 2023) and the Mixed-Effects Trait-State-Occasion model (ME-TSO, Castro-Alvarez et al., 2022) in Mplus. For both models, we exported the data including participants that completed at least 100 measurement to Mplus in the following syntax:

Table S7: Summary statistics for the estimated reliability coefficients based on PT-FA.

	Median	Q1	Q3	$N \geq 0.8$
PA English	0.62	0.55	0.73	36
PA Dutch	0.61	0.53	0.72	34
NA English	0.76	0.71	0.84	182
NA Dutch	0.70	0.63	0.81	101

Figure S9: Distribution pfa reliability PA

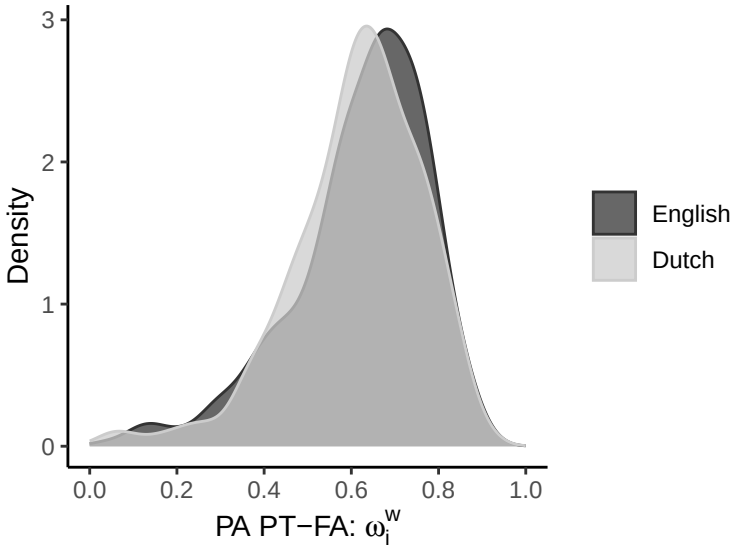
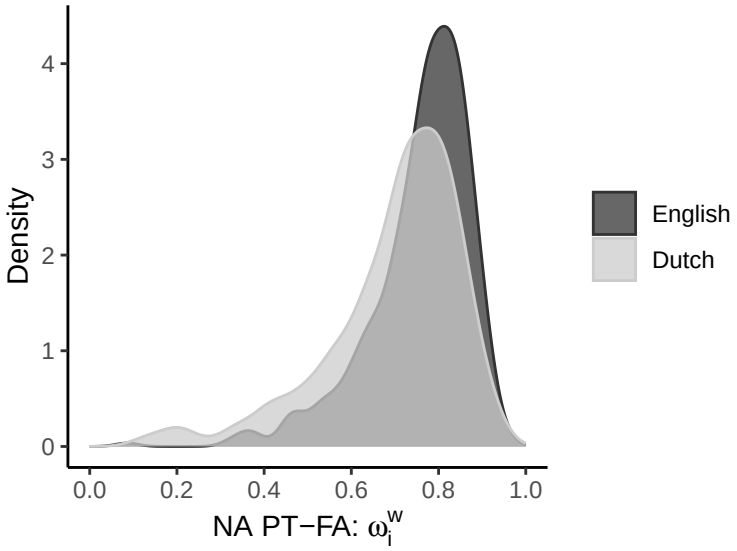


Figure S10: Distribution pfa reliability NA




```
if (!file.exists("Mplus/ema_en.dat")) {
  MplusAutomation::prepareMplusData(
  warn_d_ema_en_100[, c("id_n", "bn_grid", pa, na)],
  filename = "Mplus/ema_en.dat",
  inpfile = FALSE
  )
}

if (!file.exists("Mplus/ema_nl.dat")) {
  MplusAutomation::prepareMplusData(
  warn_d_ema_nl_100[, c("id_n", "bn_grid", pa, na)],
  filename = "Mplus/ema_nl.dat",
  inpfile = FALSE
  )
}
```

We developed the custom function `TwordmOutcomes` to read the Mplus results for the 2RDM method, and to compute reliability based on factor scores estimated by Mplus.

```
file_names <- c("Mplus/2rdm_nl_na_1d.out",
                "Mplus/2rdm_nl_pa_1d.out",
                "Mplus/2rdm_en_pa_1d.out")
# file_n <- c("Mplus/2rdm_en_na_1d.out")
```

```r
I_values <- c(7,3,3) # item size for each
results <- mapply(TwordmOutcomes, file_names, I_values, SIMPLIFY = FALSE)

# results1 <- TwordmOutcomes(file_n, 7)

# TwordmOutcomes(file_n, 7)
# Access results for each dataset
#results[[1]]  # Results for "2rdm_nl_na"
#results[[2]]  # Results for "2rdm_nl_pa"
#results[[3]]  # Results for "2rdm_en_pa"


en_pa_2rdm.out <- results$`Mplus/2rdm_en_pa_1d.out`
nl_na_2rdm.out <- results$`Mplus/2rdm_nl_na_1d.out`
nl_pa_2rdm.out <- results$`Mplus/2rdm_nl_pa_1d.out`
```

The 2rdm model converged for most cases except of the NA for the English subsample.

The median between-person reliability across iterations for the positive affect and the English subsample was 0.881 The lower CI and the upper CT of this between-person reliability coefficients were 0.86 and 0.9,respectively.

The median between-person re-liability across iterations for the positive affect and the Dutch subsample was 0.851 The lower CI and the upper CT of this between-person reliability coefficients were 0.823 and 0.876, respectively.

The median between-person re-liability across iterations for the negative affect and the Dutch subsample was 0.953 The lower CI and the upper CT of this between-person reliability coefficients were 0.944 and 0.96, respectively.

For the within-person reliability coefficients, we plotted their posterior distributions across 200 iterations for each person and the estimated within-person reliability coefficients (i.e., the median of the posterior distribution across-iteration) In the following syntax Figure S11, Figure S13, and Figure S12. The densities of the person-specific posterior distributions of the within-level reliability coefficients are shown on the left side of the plot and the distribution of the median within-level reliability coefficientsis is shown on the right side. The NA of the Dutch subsample has the lowest within-person reliability.

Table S8 shows the summary statistics for the within-person reliability based on the 2RDM model. In general, the negative affect of the Dutch subsample has the lowest within-person reliability, and the positive affect of the English subsample has the highest within-person reliability.

Figure S11: Posterior distribution of the within-reliability coefficients for positive affect of per person in the English subsample

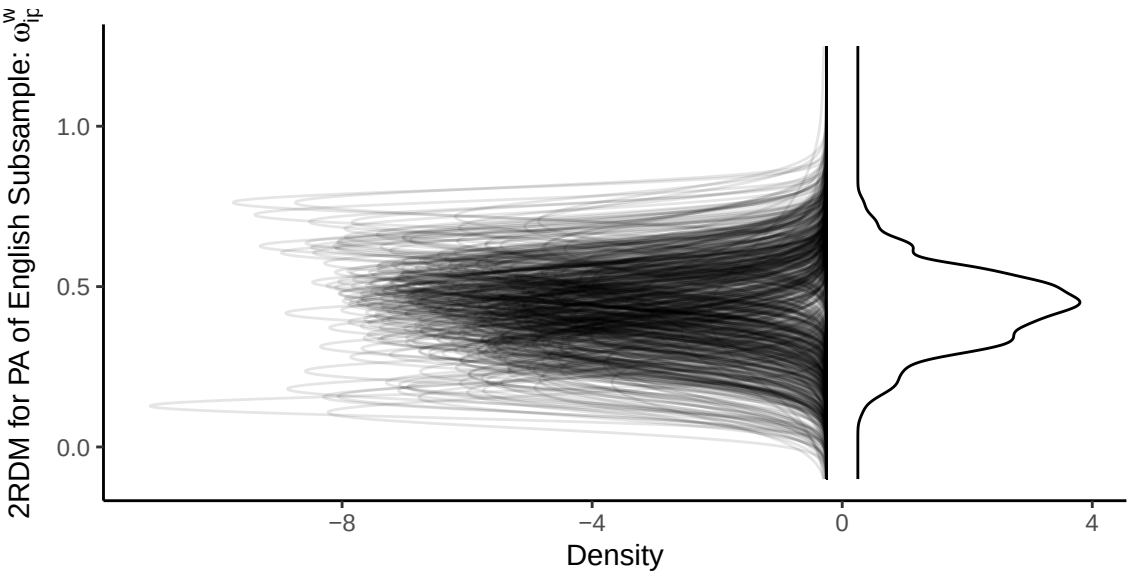


Figure S12: Posterior distribution of the within-reliability coefficients for positive affect of per person in the Dutch subsample

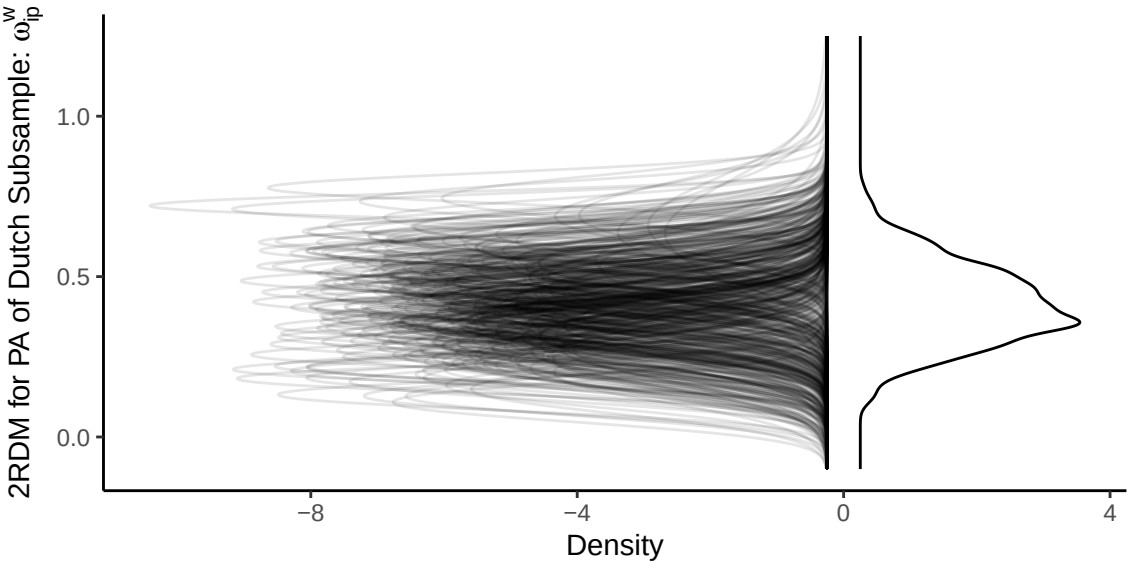


Figure S13: Posterior distribution of the within-reliability coefficients for negative affect of per person in the Dutch subsample

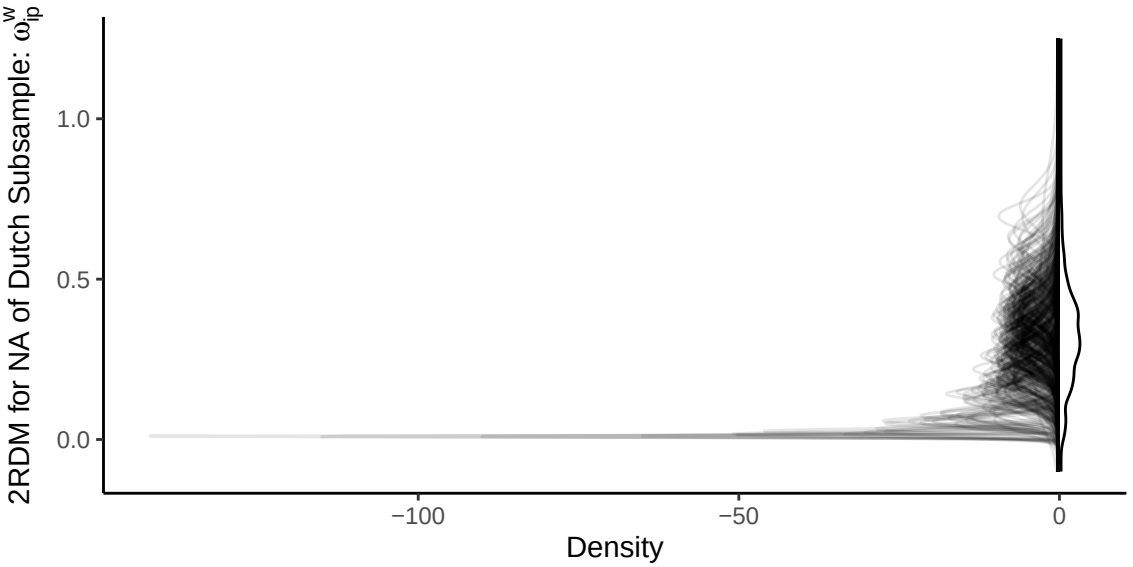


Table S8: Summary statistics for the estimated within-person reliability coefficients based on the 2rdm model.

	Median	Q1	Q3	$N \geq 0.6$
PA English	0.43	0.35	0.51	34
PA Dutch	0.41	0.33	0.50	31
NA Dutch	0.31	0.21	0.40	9

Latent state-trait theory.

The last approach we used to estimate the reliability of the items of PA and NA is the Mixed-Effects Trait-State-Occasion model (ME-TSO, Castro-Alvarez et al., 2022) encompassed within the Latent State-Trait theory and the Dynamic Structural Equation Modeling framework. This model can be described as a multilevel dynamic factor analysis, which have been especially developed to distinguish between the trait and the state components of the construct of interest. Within this model, five variance coefficients are defined as proportions of the total variance for each variable of each participant, the largest of which is referred to as the reliability. Note that the reliability coefficients are estimated per person and per item.

After fitting the models in Mplus, we imported the output into R. In the following loop, we extracted the MCMC of the parameters of interest for each model and we computed all the variance coefficients as defined by the ME-TSO.

```
# Get files names and other information to loop through.
outfiles <- c("metso_en_na.out", "metso_en_pa.out",
              "metso_nl_na.out", "metso_nl_pa.out")
I <- c(7, 3, 7, 3)
item_names <- list(na, pa, na, pa)

# Create empty list to store results.
metso_var_samples <- list(EN_NA = NULL,
                          EN_PA = NULL,
                          NL_NA = NULL,
                          NL_PA = NULL)

# Loop to read ME-TSO output into R.
for (i in 1:length(outfiles)) {
  mplus_out_tmp <- MplusAutomation::readModels(paste0("Mplus/", outfiles[i]))
  mplus_out_tmp <- mplus_out_tmp$bparameters$factor.score.imputations

  fscores_tmp <- readFScores(paste0("Mplus/", outfiles[i]),
                             clus = "ID_N",
                             ranef.vars = "AR")

  for (j in 1:nrow(mplus_out_tmp)) {
    # Extract the j-th samples of the parameter estimates
    bpar_tmp <- extract.metso.bpar(mplus_out_tmp[j, ], I[i])
    # Extract the j-th samples of the individual AR effects
    ar_tmp <- fscores_tmp[, j]
    # Compute ME-TSO variance coefficients for the j-th samples
    coeff_tmp <- metso.var.coeff(
```

```
      within.parameters  = bpar_tmp$within,
      between.parameters = bpar_tmp$between,
      ind.ar = ar_tmp,
      id     = names(ar_tmp)
    )
  metso_var_samples[[i]][[j]] <- coeff_tmp
  }
  rm(bpar_tmp, ar_tmp, coeff_tmp, j)
}
rm(mplus_out_tmp, fscores_tmp, i)
```

The result of the previous loop was a list of lists (`metso_var_samples`) that stores the computed MCMC for the variance coefficients based on the last 200 iterations. We further simplified this object by extracting MCMC of the reliability coefficients.

```
# Loop to extract the ME-TSO reliability.
metso_rel_samples <- list(EN_NA = NULL,
                          EN_PA = NULL,
                          NL_NA = NULL,
                          NL_PA = NULL)

for (i in 1:length(metso_var_samples)) {
  # Extract only the reliability samples
  metso_rel_samples[[i]] <- lapply(
    metso_var_samples[[i]], function(x) {
      x$random.situations$Xi_j0$reliability
    }
  )
  # Restructure reliability samples to array
  metso_rel_samples[[i]] <- array(
    data = unlist(metso_rel_samples[[i]]),
    dim = c(nrow(metso_rel_samples[[i]][[1]]),
            ncol(metso_rel_samples[[i]][[1]]),
            length(metso_rel_samples[[i]])),
    dimnames = list(id = rownames(metso_rel_samples[[i]][[1]]),
                    items = item_names[[i]],
                    iter  = paste0("it_", 1:length(metso_rel_samples[[i]]))
                    )
  )
}
rm(i)
```

Next, we computed the estimated reliability per person and item by taking the median of the MCMC reliability samples by person and item.

```
metso_rel_sum <- lapply(
  metso_rel_samples,
  function (x) {
    apply(x, 1, function(y) {
      apply(y, 1, median, na.rm = TRUE)
    })
    }
)

metso_rel_sum <- lapply(metso_rel_sum, function(x) {
  as.data.frame(t(x))
})
```

Table S9 presents the median reliability across persons per item
and language. For the items of PA, **cheerful** was the most reliable in both subsamples, with reliability estimates around 0.8. The median reliability for the items **motivated** and **relaxed** varied between 0.45 and 0.5. Based on these estimates, there do not seem to be major differences between the two languages. On the other hand, for the items of NA, there were quite some differences in the reliability estimates between the two languages. **Nervous** was the item with the lowest differences, and **sad** and **tired** were the items with the largest differences in the median between the two subsamples. In the English subsample, the items with the largest median reliability estimates were **stressed** and **overwhelmed**. In the Dutch subsample, the item with the largest median reliability was **ruminate**. The items **irritable** and **tired** had the lowest median reliability in both subsamples. In particular, the median reliability **tired** was very low, which might indicate that the item is not a good indicator of NA.

Figure S14 and Figure S15 present the posterior distributions of the reliability coefficients per person and the distribution of the estimated reliability coefficients over persons for the items **cheerful** and **sad**, respectively. Overall, the posterior distributions are very narrow around the median and the distribution of the estimated reliability coefficients tends to be positively skewed. To easily compare between the two languages, we also plotted the distributions of the estimated reliabilities of these two items overlapped on each other by language, as shown in Figure S16 and Figure S17.

Table S9: Summaries of the estimated reliability coefficients for the items of NA and PA by language.

	English			Dutch		
	Median	Min	Max	Median	Min	Max
Cheerful	0.80	0.77	0.88	0.80	0.77	0.93
Motivated	0.46	0.44	0.55	0.45	0.43	0.64
Relaxed	0.49	0.46	0.62	0.50	0.47	0.74
Sad	0.50	0.48	0.71	0.57	0.54	0.76
Stressed	0.69	0.65	0.87	0.65	0.61	0.84
Overwhelmed	0.69	0.66	0.86	0.65	0.63	0.81
Nervous	0.66	0.63	0.84	0.64	0.61	0.81
Ruminate	0.64	0.61	0.81	0.70	0.68	0.86
Irritable	0.48	0.45	0.69	0.45	0.42	0.66
Tired	0.31	0.30	0.39	0.38	0.38	0.44

Figure S14: Posterior distribution per person and overall distribution of the reliability coefficients for the item 'Cheerful' by language

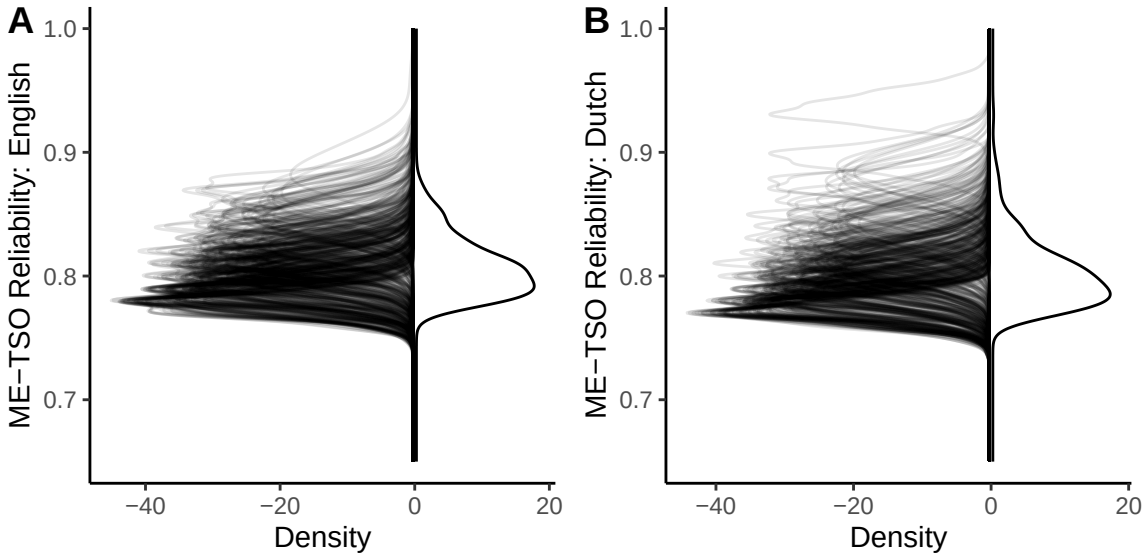


Figure S15: Posterior distribution per person and overall distribution of the reliability coefficients for the item 'Sad' by language

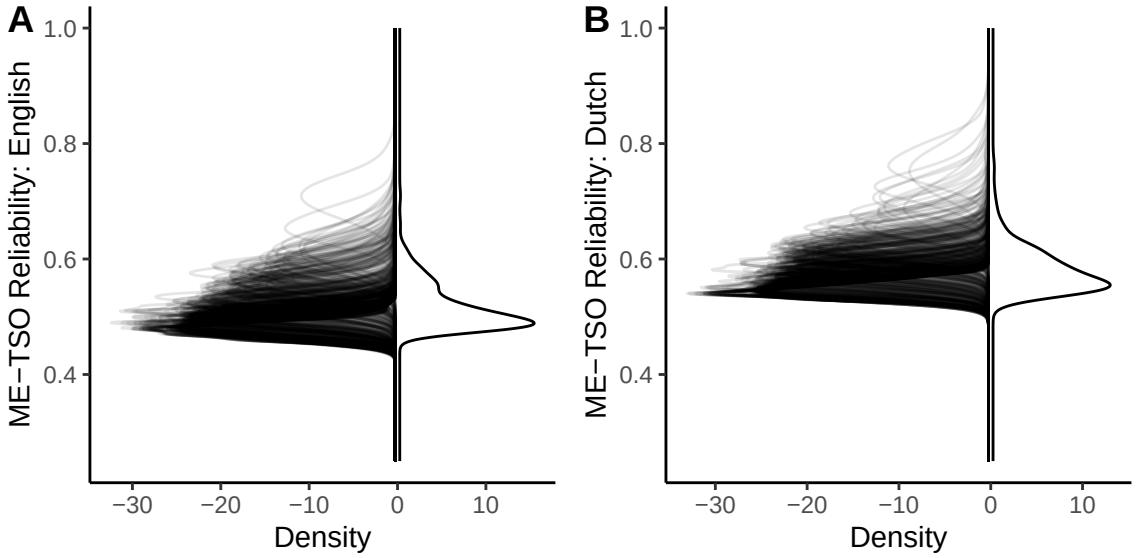


Figure S16: Distribution ME-TSO reliability: Cheerful

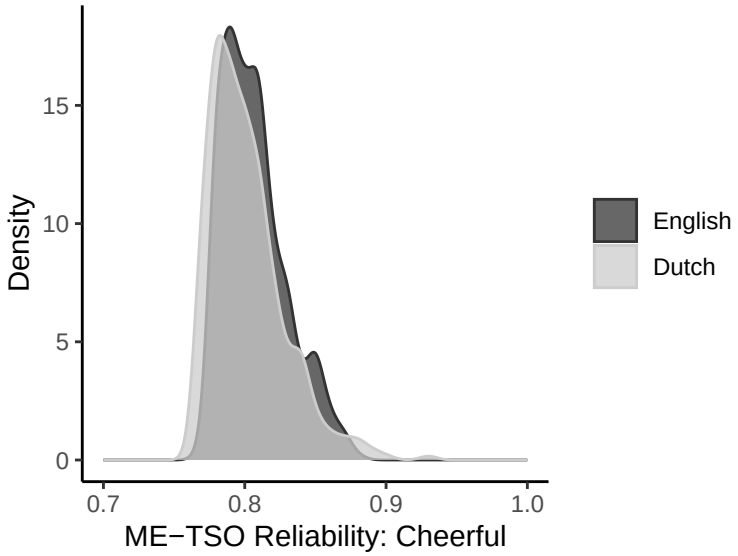
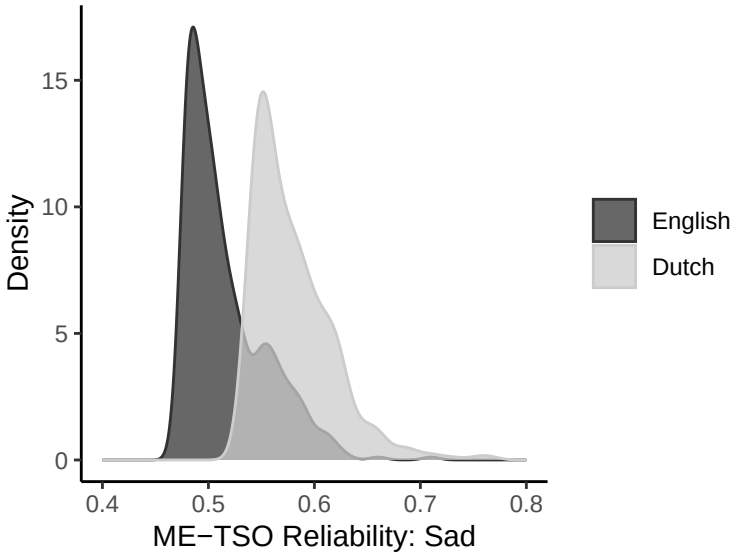


Figure S17: Distribution ME-TSO reliability: Sad



Comparing Reliabilities across Person-Specific Methods

Figure S18 presents the Within-person PA reliability comparison between PT-FA, 2RDM and ME-TSO and Figure S19 presents the Within-person NA reliability comparison between PT-FA, 2RDM and ME-TSO. The median values of the within-person reliability estimated by the PT-FA method are generally highest for both PA and NA. On the contrary, 2RDM has the lowest median values compared to the other two methods for both PA and NA. The differences between 2RDM and PT-FA are more pronounced for NA compared to PA, as PT-FA has higher values for negative affect than positive affect, but 2RDM has lower values for negative affect. The variabilities of the item-specific
within-person reliability estimated by ME-TSO are smaller than those estimated by PT-FA and 2RDM, but many of them, especially the item-specific reliability for NA, are positively skewed.

Figure S18: Within-person PA reliabilty comparison between PT-FA, 2RDM and ME-TSO

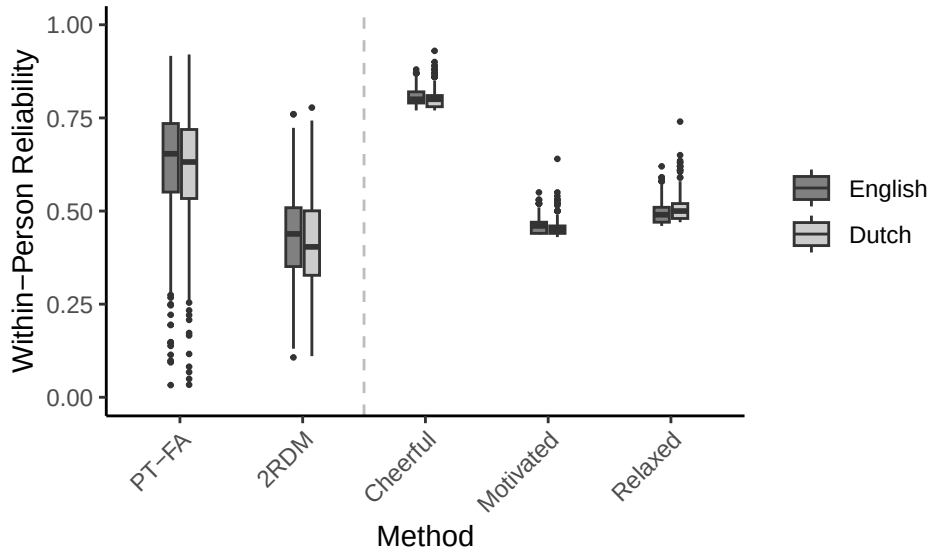
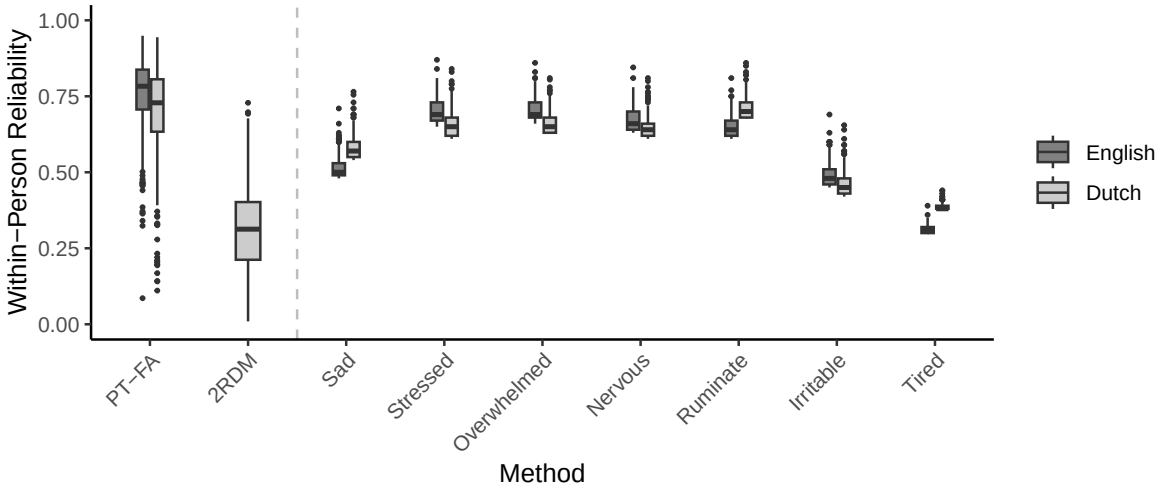


Figure S19: Within-person NA reliabilty comparison between PT-FA, 2RDM, and ME-TSO



R Session

```
R version 4.4.2 (2024-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)

Matrix products: default


locale:
[1] LC_COLLATE=English_Netherlands.utf8  LC_CTYPE=English_Netherlands.utf8
[3] LC_MONETARY=English_Netherlands.utf8 LC_NUMERIC=C
[5] LC_TIME=English_Netherlands.utf8

time zone: Europe/Amsterdam
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] kableExtra_1.4.0 ggplot2_3.5.1    knitr_1.49

loaded via a namespace (and not attached):
 [1] gtable_0.3.6      xfun_0.50         psych_2.4.12      rstatix_0.7.2
 [5] lattice_0.22-6    quadprog_1.5-8    vctrs_0.6.5       tools_4.4.2
 [9] Rdpack_2.6.2      generics_0.1.3    stats4_4.4.2      parallel_4.4.2
[13] tibble_3.2.1      pkgconfig_2.0.3   esmpack_0.1-20    Matrix_1.7-1
[17] lifecycle_1.0.4   compiler_4.4.2    farver_2.1.2      stringr_1.5.1
[21] munsell_0.5.1     mnormt_2.1.1      carData_3.0-5     htmltools_0.5.8.1
[25] yaml_2.3.10       Formula_1.2-5     pillar_1.10.1     car_3.1-3
[29] ggpubr_0.6.0      nloptr_2.1.1      tidyr_1.3.1       MASS_7.3-61
[33] reformulas_0.4.0  boot_1.3-31       abind_1.4-8       nlme_3.1-166
[37] lavaan_0.6-19     tidyselect_1.2.1  digest_0.6.37     stringi_1.8.4
[41] reshape2_1.4.4    dplyr_1.1.4       purrr_1.0.4       labeling_0.4.3
[45] splines_4.4.2     cowplot_1.1.3     fastmap_1.2.0     grid_4.4.2
[49] colorspace_2.1-1  cli_3.6.3         magrittr_2.0.3    broom_1.0.7
[53] pbivnorm_0.6.0    withr_3.0.2       scales_1.3.0      backports_1.5.0
[57] rmarkdown_2.29    lme4_1.1-36       ggsignif_0.6.4    evaluate_1.0.3
[61] rbibutils_2.3     viridisLite_0.4.2 rlang_1.1.5       Rcpp_1.0.14
[65] glue_1.8.0        xml2_1.3.6        svglite_2.1.3     rstudioapi_0.17.1
[69] minqa_1.2.8       jsonlite_1.8.9    plyr_1.8.9        R6_2.5.1
```

[73] systemfonts_1.2.1

References

Castro-Alvarez, S., Tendeiro, J., Jonge, P. de, Meijer, R. R., & Bringmann, L. (2022). Mixed-effects trait-state-occasion model: Studying the psychometric properties and the person–situation interactions of psychological dynamics. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 438–451. https://doi.org/10.31234/osf.io/4ext3

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

Fried, E. I., Proppert, R. K. K., & Rieble, C. L. (2023). Building an Early Warning System for Depression: Rationale, Objectives, and Methods of the WARN-D Study. *Clinical Psychology in Europe*, *5*(3), 1–25. https://doi.org/10.32872/cpe.10075

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

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., Neale, M. C., Sisk, C. L., & Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 532–543. https://doi.org/10.1080/10705511.2016.1148605

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. https://doi.org/10.1037/met0000287

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, *69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Tutunji, R., Fried, E. I., Proppert, R. K. K., & Rieble, C. (2024). *05_preregistration of holdout sample*. OSF. https://doi.org/10.17605/OSF.IO/NYD84

Xiao, Y., Wang, P., & Liu, H. (2023). Assessing intra- and inter-individual reliabilities in intensive longitudinal studies: A two-level random dynamic model-based approach. *Psychological Methods*. https://doi.org/10.1037/met0000608