

Supplementary material 1: Description of multilevel analyses procedure

The level-1 model included the time variables, which capture within-person change over time. In the level-2 model, between-person characteristics such as intervention condition were used to predict the slope estimates representing change in the dependent variables. Subject-specific random effects (i.e., random intercept and slope) were retained whenever they significantly contributed to the model. Bayesian Information Criterion (BIC; Burnham & Anderson, 2004) was used to determine the best fitting model, with smaller values indicating a better fit of the model to the observed data. At level 1, three separate slopes using three time-varying covariates (McCoach & Kaniskan 2010) were modeled with the mid-treatment and 3-month follow up assessment as breakpoints. All participants started out receiving STAIR and the first slope modeled the general rate of change from the pre-treatment to mid-treatment assessment, i.e., in between STAIR and trauma-focused treatment. After STAIR participants received either NT or EMDR and the second slope modeled the rate of change following the mid-treatment assessment up to 3-months follow-up. The third slope modeled the rate of change following the 3-months follow-up up to 12-months follow-up. Contrast coding was used to include the evaluation of the categorical variable intervention condition (NT coded .5 and EMDR coded -.5). For the second and third slope, a slope by intervention condition interaction term was created to test a differential effect in the rate of change in the period following the mid-treatment assessment. At level 2, a diagonal covariance structure was selected.

For the CAPS, no mid-treatment assessment was performed. For this measure, we therefore estimated a linear trend indicating the direction and rate of change, and a quadratic trend indicating whether the rate of change increased or decreased over time.

Cohen's *d* was used as an effect size and computed from the multilevel estimated means and observed standard deviations. Within-group effect sizes for each outcome measure were calculated by dividing the difference between pretreatment and any subsequent means by the standard deviation of each mean. To correct for dependence among these means, we calculated the correlations between the pre-treatment and subsequent scores (Feingold, 2013).

Between-group effect sizes from mid-treatment to 3-month follow-up and from 3-month follow-up to 12-month follow-up were calculated by subtracting the means and dividing the result by the pooled standard deviation, adjusting the calculation of treatment standard deviation for weighting the differences of the pre-post-means as proposed by Morris (2008). In this way, the intervention does not influence the standard deviation. In both calculations, the STAIR/NT group was treated as the control group.

In addition to multilevel regression analyses, we determined the clinical significance of treatment effects by calculating the percentage of participants reaching the criteria for reliable and clinically significant change. The guidelines and recommendations outlined by Wise (2004) were applied on the CAPS continuous scores. Reliable change was calculated while considering test-retest reliability for the CAPS ($r = .90$; Weathers et al., 2001). The threshold for clinical significance was set at 48.96 (Beidel et al., 2019). A Cox regression survival analysis was performed to examine the effect of NT and EMDR following STAIR on the categorical outcome of a current diagnosis of PTSD as rated by a clinician on the basis of the CAPS interview. To prevent data from being biased due to measurement attrition, missing values were handled using a multiple imputation procedure with regression switching and predictive mean matching (MICE-PMM; Marshall, Altman, & Holder, 2010). To include all participants in the analysis, missing post-treatment, 3-month and 12-month follow-up data were estimated 5 times to generate 5 data sets with imputed data. These 5 imputed data sets

were then averaged to provide pooled frequency counts, hazard ratios, Wald χ^2 statistics, and p values.