**Online Supplementary Material**

|  |  |
| --- | --- |
| **Appendix A.** Feedback reports and feedback session | Page 2 |
| **Appendix B.**  Comparison between dropouts and completers on the four outcome measures at baseline | Page 6 |
| **Appendix C.**  Feasibility and usability of the intervention | Page 8 |
| **Appendix D.**  Statistical analyses for depressive symptom severity | Page 10 |
| **Appendix E.**  Statistical analyses for disturbances in interpersonal relations | Page 15 |
| **Appendix F.**  Statistical analyses for disturbances in social role functioning | Page 20 |
| **Appendix G.**  Statistical analyses for empowerment | Page 26 |
| **Appendix H.**  Post-hoc analyses | Page 32 |

# Appendix A

Feedback reports and feedback session

Standardized feedback reports were generated based on individual data and emailed to the participant after each of the four weeks of EMA measurements. Exemplary feedback reports (in Dutch) are available online: <https://osf.io/m6hvg/>. There was ample room in the feedback reports for participants to write their remarks on what caught their eye in the graphs and how they would explain the patterns. Feedback reports in the Do-module showed data related to positive affect (PA), activity patterns, and the relations between the two. Conversely, feedback reports in the Think-module showed data related to negative affect (NA), minor daily events, thoughts, and the relations between them. Each successive report contained richer information and each report ended with an announcement of the information that would be added in the following report. The feedback reports predominantly comprised descriptive graphs, which are listed below:

**Descriptive graphs**

**Do-module**

*Week 1*

Bar graph: Number of completed measurements in week 1

Bar graph: Average intensity of the six separate positive emotions

Line graph: PA peaks and valleys from moment-to-moment in the preceding week

Bar graphs: Average intensity of PA, physical activity, time spent outdoors, time spent in pleasant company, and time spent at home in the preceding week

*Week 2*

Bar graph: Number of completed measurements in week 2

Line graph: PA and physical activity peaks and valleys from moment-to-moment in the preceding week

Bar graphs: Average intensity of PA, physical activity, time spent outdoors, time spent in pleasant company, and time spent at home for week 1 and week 2

Bar graph: Top 3 activities (based on frequency in the preceding week)

*Week 3*

Bar graph: Number of completed measurements in week 3

Line graphs: PA peaks and valleys from moment-to-moment in the preceding week in combination with lines for physical activity, time spent outdoors, time spent in pleasant company, and time spent at home

Bar graphs: Average intensity of PA, physical activity, time spent outdoors, time spent in pleasant company, and time spent at home for week 1-3

Bar graph: Frequency of all listed activities in the preceding week

Bar graph: Top 3 activities (based on concurrent PA in the preceding week)

*Week 4*

Bar graph: Total number of missed and completed measurements

Line graph: PA intensity averaged across time of day (for all time points)

Line graph: PA peaks and valleys from day-to-day across the intervention period

Bar graphs: Average intensity of PA, physical activity, time spent outdoors, time spent in pleasant company, and time spent at home for week 1-4

Pie chart: Amount of time spent doing nothing versus doing at least 1 activity

Bar graph Average PA intensity after having done nothing or something

Bar graph: Frequency of all listed activities across the intervention period

Bar graph: Average PA after listed activities across the intervention period

**Think-module**

*Week 1*

Bar graph: Number of completed measurements in week 1

Bar graph: Average intensity of the six separate negative emotions

Line graph: NA peaks and valleys from moment-to-moment in the preceding week

Bar graphs: Average intensity of NA, rumination, negative thoughts, and positive thoughts in the preceding week

*Week 2*

Bar graph: Number of completed measurements in week 2

Line graph: NA and rumination peaks and valleys from moment-to-moment in the preceding week

Bar graphs: Average intensity of NA, rumination, negative thoughts, and positive thoughts for week 1 and week 2

Bar graph: Number of negative and positive events in the preceding week

*Week 3*

Bar graph: Number of completed measurements in week 3

Line graphs: NA peaks and valleys from moment-to-moment in the preceding week in combination with lines for rumination, negative thoughts, and positive thoughts

Bar graphs: Average intensity of NA, rumination, negative thoughts, and positive thoughts for week 1-3

Pie chart: Amount of measurements that were preceded by only positive events, only negative events, both, or none (in the preceding week)

Bar graph: Number of negative and positive events in the preceding week

Bar graphs: Intensity of NA and rumination after negative and positive events

*Week 4*

Bar graph: Total number of missed and completed measurements

Line graph: NA intensity averaged across time of day (for all time points)

Line graph: NA peaks and valleys from day-to-day across the intervention period

Line graph: Average daily momentary well-being in combination with retrospective daily well-being (to visualize potential retrospective bias)

Bar graphs: Average intensity of NA, rumination, negative thoughts, and positive thoughts for week 1-4

Pie chart: Amount of all measurements that were preceded by only positive events, only negative events, both, or none

Bar graph: Number of negative and positive events in the 4-week intervention period

Bar graphs: Intensity of NA, rumination, positive thoughts, and negative thoughts after periods with only positive events, only negative events, both, or none (across the intervention period)

**Statistical tests (week 4)**

Participants were encouraged to complete at least 28 of the 35 (80%) self-assessments each week. Hence, the first graph in each report showed the number of completed assessments in reference to the 80-percent line. For those participants who eventually filled out 75% or more self-assessments, the fourth and final feedback report additionally included feedback on temporal relationships between multiple sets of two variables: between PA and four activity types (physical activity, time spent outdoors, time spent in pleasant company, time spent at home) for the Do-module, and between NA and negative thoughts, positive thoughts, and rumination for the Think-module. To determine the temporality of effects within a specific person (*e.g.*, whether increased PA precedes or follows an increase in physical activity) we used vector autoregression (VAR) models. The R package Autovar (Emerencia, 2018) allowed us to automatically estimate and optimize VAR models for each individual participant and to generate two-variable networks for personalized feedback. Procedures were adapted from the HowNutsAreTheDutch study (Blaauw et al., 2014). Each feedback report comprised a general explanation of the network graphs, which was further clarified by the researcher in the feedback session.

**Feedback session**

The final feedback report was discussed with a research assistant according to a standardized protocol. First, the research assistant inquired whether there had been any singularities during the intervention period and emphasized that the report only reflected (what had been measured during) that specific time period and that statistical results are not necessarily ‘true’ and should hence be interpreted with caution. Second, the research assistant explained how to read the graphs and helped participants provide meaning by asking them what they found striking, whether they recognized any patterns, and how they would explain them. Participants were encouraged to make notes and to share the feedback report with their (future) therapist. Feedback sessions lasted approximately 15-30 minutes.

**References**

Blaauw, F., van der Krieke, L., Bos, E., Emerencia, A., Jeronimus, B. F., Schenk, M., … de Jonge, P. (2014). HowNutsAreTheDutch: Personalized feedback on a national scale. *Expanding the Boundaries of Health Informatics Using AI: Papers from the 2014 AAAI Fall Symposium.* Retrieved from <http://www.cs.rug.nl/~ando/pdfs/blaauw_2014.pdf>

Emerencia, A. (2018). autovarCore: Automated Vector Autoregression Models and Networks (Version 1.0-4). Retrieved from <https://CRAN.R-project.org/package=autovarCore>

# Appendix B

Comparison between dropouts and completers on the four outcome measures at baseline

We examined whether there was a significant difference between dropouts (*n* = 20) and completers (*n* = 90) on the four outcome variables at baseline: the IDS-SR for depressive symptom severity, the NEL for empowerment, and the IR and SR subscales of the OQ-45 for functioning in interpersonal relations and social roles, respectively. Inspection of histograms and Q-Q plots revealed some deviations from normality in the dropout group (*n* = 20). Moreover, variances were unequal between dropouts and completers for the IDS-SR and OQ-45 IR scale as assessed by Levene’s test for equality of variances (Table 1). Therefore, we used Welch’s test for unequal variances instead of the planned independent samples t-tests and found no statistically significant differences between the groups (Table 2, Figure 1a-d).

Table 1 *Levene’s test of equal variances between dropouts and completers for the four main outcome measures*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Outcome | *F* | *df* | *p* | *CI95* | |
| Lower bound | Upper bound |
| Depressive symptoms | 1.86 | 19 / 89 | .05 | 1.0 | 4.2 |
| Disturbances in interpersonal relations | 1.93 | 19 / 89 | .04 | 1.0 | 4.3 |
| Disturbances in social roles | 1.22 | 19 / 89 | .53 | 0.6 | 2.7 |
| Empowerment † | 0.70 | 19 / 89 | .38 | 0.37 | 1.6 |

*Note.* † NEL total score without the professional help scale

Table 2 *Welch’s two sample t-test between dropouts and completers for the four main outcome measures*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Outcome | *t* | *df* | *p* | *CI95* | |
| Lower bound | Upper bound |
| Depressive symptoms | 1.0 | 23.7 | .31 | -3.4 | 10.1 |
| Disturbances in interpersonal relations | 0.4 | 23.6 | .69 | -2.5 | 3.7 |
| Disturbances in social roles | -0.1 | 26.4 | .95 | -2.7 | 2.5 |
| Empowerment † | 0.0 | 32.3 | .98 | -6.5 | 6.6 |

*Note.* † NEL total score without the professional help scale



*Figure 1.* Comparison between dropouts and completers on the four outcome measures at baseline. (a) depressive symptom severity (IDS-SR total score), (b) disturbances in interpersonal relations (OQ-45 IR scale), (c) disturbances in social role functioning (OQ-45 SR scale), (d) empowerment (NEL total score without the professional help scale).

# Appendix C

Feasibility and usability of the intervention

For information on patient-perceived feasibility and usability of the interventionsee Table 1. Participants indicated in the post-EMA evaluation questionnaire (n = 90) that they already used a smartphone quite often before study start (78.3 ± 19.6); 86 participants used their own smartphone for the self-assessments and 4 were lent one for study purposes. Almost all participants (*n* = 89) received the feedback reports at their own e-mail address; one participant had to create an e-mail address for the study. Most participants preferred receiving the interim reports via e-mail or post (*n* = 82), while some would rather have additionally discussed the interim reports (*n* = 4) and others had rather not received the interim reports (*n* = 4). For the final feedback report, a majority preferred discussing the feedback either directly (n = 19) or after having received the report digitally (n =35). A substantial subgroup (n =35) would have preferred merely receiving the final feedback report per e-mail or post and one participant would rather have not received the final feedback report at all. These results are striking given the high ratings on the question how useful it was to discuss the final report with the researcher (see Table 1).

Table 1*Patient-perceived feasibility and usability of the intervention (post-EMA evaluation questionnaire)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total  (n=90)a | Do-module  (n = 47) | Think-module  (n = 43) |
| *Web application* |  |  |  |
| Was the app easy to handle? | 87.8 ± 15.7 | 90.0 ± 13.0 | 85.4 ± 18.0 |
| Did technical problems hinder you in filling out questions? | 18.5 ± 22.6 | 23.3 ± 25.6 | 13.2 ± 17.5 |
| Did you find the app annoying or stressful? |  |  |  |
| With respect to the number of measurements per day | 42.2 ± 30.6 | 41.4 ± 28.8 | 43.0 ± 32.7 |
| With respect to the length of the daily measurements | 18.6 ± 22.1 | 16.1 ± 21.2 | 21.4 ± 22.9 |
| With respect to the duration of the measurement period | 27.0 ± 27.0 | 24.3 ± 27.1 | 30.0 ± 26.8 |
| Did you find that structure (5 measurements per day for 28 days) pleasant? | 69.3 ± 21.6 | 67.6 ± 22.4 | 71.2 ± 20.8 |
| Did you have to adjust your normal activities or daily routine due to the measurements? | 28.5 ± 26.0 | 25.2 ± 24.4 | 32.2 ± 27.5 |
| Did you find the app annoying or stressful with respect to the content of the questions? | 18.8 ± 20.5 | 16.0 ± 21.6 | 21.9 ± 19.1 |
| Were any of the questions in the app difficult or unclear? | 14.9 ± 19.7 | 14.8 ± 22.4 | 15.0 ± 16.7 |
| Did you experience feelings that were not covered by the questions? | 28.7 ± 24.4 | 24.6 ± 23.8 | 33.3 ± 24.4 |
| Could you fit all your daily experiences in the measurements? | 60.0 ± 23.4 | 61.0 ± 23.2 | 57.9 ± 23.8 |
| *Feedback reports* |  |  |  |
| Were the interim reports comprehensible? | 68.8 ± 20.2 | 70.9± 21.0 | 66.7 ± 19.5 |
| Was the final report comprehensible? | 70.0 ± 20.5 | 72.2 ± 21.5 | 67.4 ± 19.1 |
| Were the interim reports useful? | 62.5 ± 22.1 | 62.9 ± 20.7 | 62.0 ± 23.8 |
| Was the final report useful? | 70.6 ± 20.5 | 69.5 ± 21.2 | 71.9 ± 19.7 |
| What did you think of the amount of feedback in the final report? |  |  |  |
| Too little | 13% | 11% | 16% |
| Exactly right | 86% | 87% | 84% |
| Too much | 1% | 2% | 0% |
| Did you find it useful to discuss the final report with the researcher? | 79.7 ± 17.0 | 79.2 ± 17.5 | 80.4 ± 16.5 |
| *Overall* |  |  |  |
| Recommend intervention to others | 86% | 85% | 86% |

*Note.* Questions were mainly rated on visual analogue scales (ranging from not at all (0) to very much (100)), and in some cases on dichotomous scales. Numbers represent mean ± standard deviation and percentages for the dichotomous variables. a One Think-module dropout did complete the evaluation questionnaire, bringing the total to 90.

# Appendix D

Statistical analyses for depressive symptom severity

**Completed measurementsDescriptive statistics**

The number of completed IDS-SR assessments per measurement wave and corresponding means (SD) per group are provided in Table 1a and Table 1b, respectively.

Table 1a*Number of completed assessments for depressive symptom severity (IDS-SR)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | Do-module | Think-module | Control |
| Baseline | 161 | 55 | 55 | 51 |
| Post-EMA | 133 | 47 | 45 | 41 |
| FU1 | 110 | 38 | 40 | 32 |
| FU2 | 99 | 32 | 32 | 35 |
| FU3 | 96 | 38 | 30 | 28 |
| FU4 | 89 | 31 | 30 | 28 |

*Note.* Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

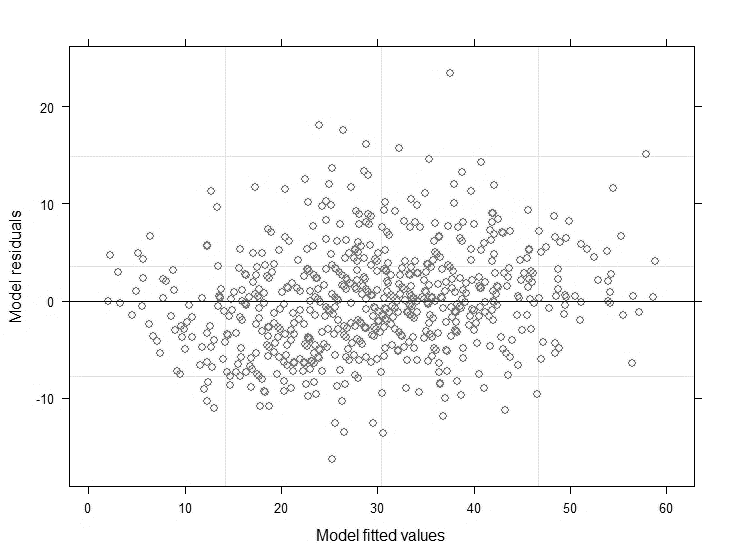
Table 1b *Observed means (SD) for depressive symptom severity (IDS-SR)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Baseline | Post-EMA | FU1 | FU2 | FU3 | FU4 |
| Control | 36.2 (10.2) | 32.6 (13.4) | 29.0 (13.2) | 28.4 (15.2) | 26.5 (15.0) | 27.8 (15.4) |
| Do-module | 35.7 (11.4) | 29.2 (11.3) | 26.5 (13.9) | 27.4 (16.6) | 26.2 (14.5) | 25.2 (16.2) |
| Think-module | 35.9 (10.5) | 30.8 (10.7) | 27.9 (10.8) | 28.3 (13.4) | 27.5 (13.6) | 25.8 (13.8) |

*Note.* Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

**Assumption checks**

The assumptions of normality and independence of residuals were satisfied, as shown in Figure 1 by the equal spread around a mean of zero and the random pattern across the model fitted values. Individual plots of the residuals against each predictor showed linear relationships and no seasonal pattern for the residuals, satisfying all linear mixed model assumptions.



*Figure 1.* Scatterplot of model residuals plotted against model fitted values for depressive symptoms (IDS-SR total score).

**Intention-to-treat analysis**

Our main analysis was based on the intention-to-treat principle: participants were compared within the groups to which they were initially randomized, independently of having received the allocated treatment, having dropped out of the study or having violated the initial protocol (for whatever reason).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 2. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 2.

**Random effects.** The random effects, representing the differences between participants in baseline levels and changes over time across groups, indicated a large variance in IDS-SR starting levels (*S2* IDS-SR intercept = 95.0) and changes over time (*S2* IDS-SR slope *=* 2.7). This heterogeneity is further illustrated by the spread of the individual predicted regression lines in Figure 3. The intraclass correlation (ICC) indicating the proportion of variance accounted for by participants (level 2) was .76. This effectively means that three quarters of the residual variance is explained by between-person differences.

Table 2 *Intention-to-treat analysis for depressive symptom severity (IDS-SR)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *p* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 35.8 (1.6) | 22.6 | 188 | < .001 | 32.7 | 38.9 |
| Time | -3.8 (0.7) | -5.4 | 539 | < .001 | -5.2 | -2.4 |
| Time2 | 0.3 (0.1) | 3.5 | 470 | < .001 | 0.1 | 0.5 |
| Do (vs control) | -0.9 (2.2) | -0.4 | 187 | .67 | -5.3 | 3.4 |
| Think (vs control) | -0.8 (2.2) | -0.4 | 188 | .71 | -5.1 | 3.5 |
| Do x Time | 0.1 (1.0) | 0.1 | 538 | .94 | -1.8 | 2.0 |
| Think x Time | 0.7 (1.0) | 0.8 | 540 | .44 | -1.2 | 2.7 |
| Do x Time2 | 0.0 (0.1) | 0.3 | 470 | .79 | -0.2 | 0.3 |
| Think x Time2 | -0.1 (0.1) | -0.5 | 470 | .59 | -0.3 | 0.2 |

*Note*. The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -0.1, *SE =* 2.2, *t*(187) = -0.1, *p* = .95; time *B* = -0.7, *SE* = 0.9, *t*(539) = -0.7, *p* = .48; time2 *B* = 0.1, *SE* = 0.1, *t*(471) = 0.8, *p* = .41.

*Figure 2.*Mean scores for depressive symptom severity (IDS-SR) and predicted lines plotted across time (intention-to-treat analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.



*Figure 3.*Individual predicted lines for depressive symptom severity (IDS-SR) within each group. The bold line represents the mean predicted line for the group. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

**Per-protocol analysis**

In addition to the intention-to-treat analysis, we examined the value of the add-on tool exclusively among participants who did not drop out of the study (Do-module: *n ­=* 48, Think-module: *n* = 42).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 3. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 4.

**Random effects.** The random effects indicated a large variance in IDS-SR starting levels (*S2* IDS-SR intercept = 89.2) and changes over time (*S2* IDS-SR slope *=* 2.7). The intraclass correlation (ICC) was .76.

Table 3 *Fixed effects of the per-protocol analysis for depressive symptom severity (IDS-SR)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 35.8 (1.5) | 23.1 | 172 | < .001 | 32.8 | 38.9 |
| Time | -3.8 (0.7) | -5.4 | 524 | < .001 | -5.2 | -2.4 |
| Time2 | 0.3 (0.1) | 3.4 | 461 | < .001 | 0.1 | 0.5 |
| Do (vs control) | -1.7 (2.2) | -0.8 | 170 | .44 | -6.1 | 2.6 |
| Think (vs control) | -1.7 (2.3) | -0.7 | 170 | .47 | -6.2 | 2.8 |
| Do x Time | 0.1 (1.0) | 0.1 | 520 | .88 | -1.8 | 2.0 |
| Think x Time | 1.1 (1.0) | 1.1 | 521 | .29 | -0.9 | 3.0 |
| Do x Time2 | 0.0 (0.1) | 0.3 | 458 | .79 | -0.2 | 0.3 |
| Think x Time2 | -0.1 (0.1) | -0.8 | 453 | .45 | -0.4 | 0.2 |

*Note.* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -0.1, *SE =* 2.3, *t*(168) = -0.0, *p* = .98; time: *B* = -0.9, *SE* = 1.0, *t*(515) = -0.9, *p* = .34; time2: *B* = 0.1, *SE* = 0.1, *t*(450) = 1.0, *p* = .37.

*Figure 4.*Mean scores for depressive symptom severity (IDS-SR) and predicted lines plotted across time (per-protocol analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

# Appendix E

Statistical analyses for disturbances in interpersonal relations

**Descriptive statistics**

The number of completed OQ-45 assessments per measurement wave and corresponding means (SD) per group are provided in Table 1a and Table 1b, respectively.

Table 1a*Number of completed assessments for OQ-45*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | Do-module | Think-module | Control |
| Baseline | 161 | 55 | 55 | 51 |
| Post-EMA | 132 | 47 | 44 | 41 |
| FU1 | 108 | 38 | 39 | 31 |
| FU2 | 96 | 31 | 31 | 34 |
| FU3 | 93 | 36 | 29 | 28 |
| FU4 | 88 | 31 | 29 | 28 |

Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

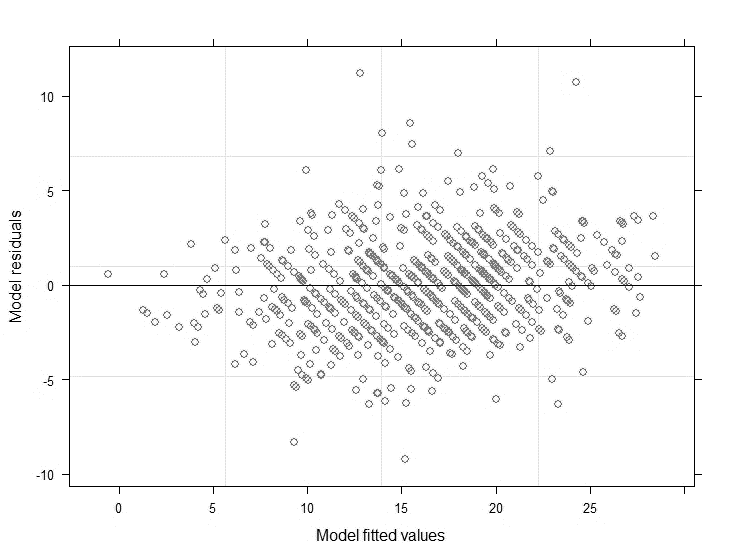
Table 1b *Observed means (SD) for disturbances in interpersonal relations (OQ-45 IR scale)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Baseline | Post-EMA | FU1 | FU2 | FU3 | FU4 |
| Control | 18.3 (6.5) | 17.3 (6.8) | 16.7 (6.8) | 16.4 (7.4) | 15.9 (6.5) | 16.1 (7.1) |
| Do-module | 16.5 (4.5) | 15.9 (6.1) | 15.8 (7.7) | 15.4 (7.1) | 15.2 (6.5) | 14.8 (6.0) |
| Think-module | 17.2 (5.2) | 16.9 (6.2) | 16.4 (5.7) | 16.5 (6.0) | 15.0 (5.3) | 14.0 (5.2) |

*Note.* Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

**Assumption checks**

The assumptions of normality and independence of residuals were satisfied, as shown in Figure 1 by the equal spread around a mean of zero and the random pattern across the model fitted values. Individual plots of the residuals against each predictor showed linear relationships and no seasonal pattern for the residuals, satisfying all linear mixed model assumptions.



*Figure 1.* Scatterplot of model residuals plotted against model fitted values for disturbances in interpersonal relations (OQ-45 IR scale).

**Intention-to-treat analysis**

Our main analysis was based on the intention-to-treat principle: participants were compared within the groups to which they were initially randomized, independently of having received the allocated treatment, having dropped out of the study or having violated the initial protocol (for whatever reason).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 2. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 2.

**Random effects.** The random effects, representing the differences between participants in baseline levels and changes over time across groups, indicated a large variance in IR starting levels (*S2* IR intercept = 26.4) and changes over time (*S2* IR slope *=* 0.3). This heterogeneity is further illustrated by the spread of the individual predicted regression lines in Figure 3. The intraclass correlation (ICC) indicating the proportion of variance accounted for by participants (level 2) was .76. This effectively means that three quarters of the residual variance is explained by between-person differences.

Table 2 *Intention-to-treat analysis* *for disturbances in interpersonal relations (OQ-45 IR scale)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 17.7 (0.8) | 22.5 | 161 | < .001 | 16.2 | 19.2 |
| Time | -0.4 (0.1) | -3.0 | 84 | .003 | -0.7 | -0.2 |
| Do (vs control) | -1.2 (1.1) | -1.1 | 160 | .26 | -3.4 | 0.9 |
| Think (vs control) | -0.4 (1.1) | -0.4 | 161 | .69 | -2.6 | 1.7 |
| Do x Time | 0.2 (0.2) | 0.9 | 85 | .36 | -0.2 | 0.6 |
| Think x Time | -0.1 (0.2) | -0.3 | 84 | .73 | -0.5 | 0.3 |

*Note* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -0.8, *SE =* 1.1, *t*(161) = -0.8, *p* = .45; time: *B* = 0.2, *SE* = 0.2, *t*(86) = 1.3, *p* = .20.



*Figure 2.*Mean scores for disturbances in interpersonal relations (OQ-45 IR scale) and predicted lines plotted across time (intention-to-treat analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

*Figure 3.*Individual predicted lines for disturbances in interpersonal relations (OQ-45 IR scale) within each group. The bold line represents the mean predicted line for the group. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

**Per-protocol analysis**

In addition to the intention-to-treat analysis, we examined the value of the add-on tool exclusively among participants who did not drop out of the study (Do-module: *n ­=* 48, Think-module: *n* = 42).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 3. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 4.

**Random effects.** The random effects indicated a large variance in IR starting levels (*S2* IR intercept = 26.8) and changes over time (*S2* IR slope *=* 0.3). The intraclass correlation (ICC) was .77.

Table 3 *Fixed effects of the per-protocol analysis for disturbances in interpersonal relations (OQ-45 IR scale)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 17.7 (0.8) | 22.4 | 145 | < .001 | 16.2 | 19.3 |
| Time | -0.4 (0.1) | -3.0 | 85 | < .01 | -0.7 | -0.2 |
| Do (vs control) | -1.6 (1.1) | -1.4 | 142 | .15 | -3.8 | 0.6 |
| Think (vs control) | -0.3 (1.2) | -0.2 | 142 | .82 | -2.6 | 2.0 |
| Do x Time | 0.2 (0.2) | 0.9 | 86 | .39 | -0.2 | 0.5 |
| Think x Time | -0.1 (0.2) | -0.5 | 85 | .61 | -0.5 | 0.3 |

*Note* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -1.4, *SE =* 1.2, *t*(140) = -1.1, *p* = .25; time: *B* = 0.3, *SE* = 0.2, *t*(86) = 1.4, *p* = .17.

**

*Figure 4.*Mean scores for disturbances in interpersonal relations (OQ-45 IR scale) and predicted lines plotted across time (per-protocol analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

# Appendix F

Statistical analyses for disturbances in social role functioning

**Descriptive statistics**

The number of completed OQ-45 assessments per measurement wave and corresponding means (SD) per group are provided in Table 1a and Table 1b, respectively.

Table 1a*Number of completed assessments for OQ-45*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | Do-module | Think-module | Control |
| Baseline | 161 | 55 | 55 | 51 |
| Post-EMA | 132 | 47 | 44 | 41 |
| FU1 | 108 | 38 | 39 | 31 |
| FU2 | 96 | 31 | 31 | 34 |
| FU3 | 93 | 36 | 29 | 28 |
| FU4 | 88 | 31 | 29 | 28 |

Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

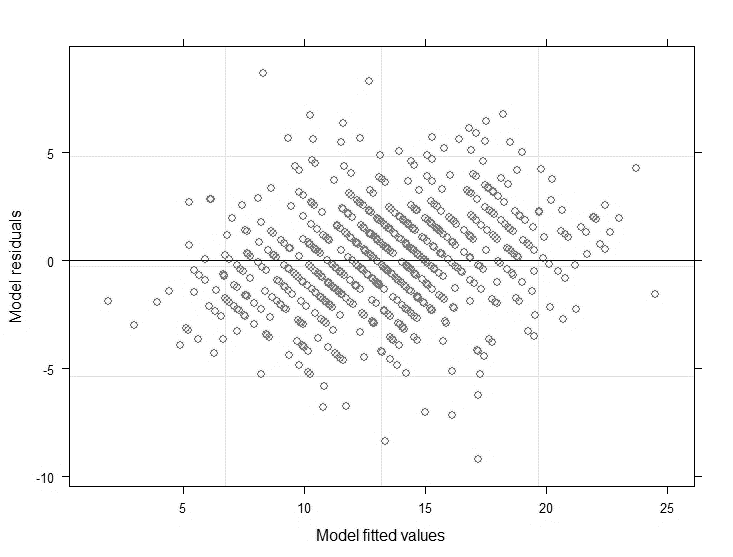
Table 1b *Observed means (SD) for disturbances in social role functioning (OQ-45 SR scale)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Baseline | Post-EMA | FU1 | FU2 | FU3 | FU4 |
| Control | 14.9 (5.2) | 14.8 (4.9) | 14.7 (4.4) | 14.0 (5.2) | 12.3 (5.7) | 14.0 (5.2) |
| Do-module | 13.7 (5.0) | 12.3 (4.7) | 12.0 (5.5) | 12.0 (5.3) | 11.2 (4.6) | 12.5 (5.0) |
| Think-module | 15.2 (4.4) | 13.3 (5.0) | 12.9 (4.2) | 12.2 (4.5) | 13.0 (4.3) | 12.8 (4.6) |

*Note.* Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

**Assumption checks**

The assumptions of normality and independence of residuals were satisfied, as shown in Figure 1 by the equal spread around a mean of zero and the random pattern across the model fitted values. Individual plots of the residuals against each predictor showed linear relationships and no seasonal pattern for the residuals, satisfying all linear mixed model assumptions.



*Figure 1.* Scatterplot of model residuals against model fitted values for disturbances in social roles (OQ-45 SR scale).

**Intention-to-treat analysis**

Our main analysis was based on the intention-to-treat principle: participants were compared within the groups to which they were initially randomized, independently of having received the allocated treatment, having dropped out of the study or having violated the initial protocol (for whatever reason).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 2. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 2.

**Random effects.** The random effects, representing the differences between participants in baseline levels and changes over time across groups, indicated a large variance in SR starting levels (*S2* SR intercept = 15.3) and (to a lesser extent) changes over time (*S2* SR slope *=* 0.2). This heterogeneity is further illustrated by the spread of the individual predicted regression lines in Figure 3. The intraclass correlation (ICC) indicating the proportion of variance accounted for by participants (level 2) was .66. This effectively means that two thirds of the residual variance is explained by between-person differences.

Table 2

*Intention-to-treat analysis* *for disturbances in social role functioning (OQ-45 SR scale)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 15.1 (0.7) | 23.1 | 193 | < .001 | 16.8 | 16.4 |
| Time | -0.6 (0.3) | -2.2 | 537 | .03 | -1.2 | -0.1 |
| Time2 | 0.1 (0.0) | 1.4 | 476 | .16 | -0.0 | 0.1 |
| Do (vs control) | -1.6 (0.9) | -1.7 | 193 | .09 | -3.3 | 0.2 |
| Think (vs control) | -0.2 (0.9) | -0.3 | 194 | .80 | -2.0 | 1.6 |
| Do x Time | -0.5 (0.4) | -1.2 | 533 | .2s5 | -1.3 | 0.3 |
| Think x Time | -0.3 (0.4) | -0.8 | 535 | .41 | -1.2 | 0.5 |
| Do x Time2 | 0.1 (0.1) | 1.5 | 476 | .14 | -0.0 | 0.2 |
| Think x Time2 | 0.0 (0.1) | 0.8 | 477 | .42 | -0.1 | 0.2 |

*Note* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -1.3, *SE =* 0.9, *t*(193) = -1.5, *p* = .14; time: *B* = -0.1, *SE* = 0.4, *t*(531) = -0.3, *p* = .74; time2: *B* = 0.0, *SE* = 0.1, *t*(477) = 0.7, *p* = .51

*Figure 2.*Mean scores for disturbances in social role functioning (OQ-45 SR scale) and predicted lines plotted across time (intention-to-treat analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.



*Figure 3.*Individual predicted lines for disturbances in social role functioning (OQ-45 SR-scale) within each group. The bold line represents the mean predicted line for the group. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

**Per-protocol analysis**

In addition to the intention-to-treat analysis, we examined the value of the add-on tool exclusively among participants who did not drop out of the study (Do-module: *n ­=* 48, Think-module: *n* = 42).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 3. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 4.

**Random effects.** The random effects, indicated a large variance in SR starting levels (*S2* SR intercept = 15.6) and (to a lesser extent) changes over time (*S2* SR slope *=* 0.2). The intraclass correlation (ICC) was .65.

Table 3

*Fixed effects of the per-protocol analysis for disturbances in social role functioning (OQ-45 SR-scale)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 15.1 (0.7) | 23.0 | 174 | < .001 | 13.8 | 16.4 |
| Time | -0.6 (0.3) | -2.2 | 521 | .03 | -1.2 | -0.1 |
| Time2 | 0.1 (0.0) | 1.4 | 463 | .16 | -0.0 | 0.1 |
| Do (vs control) | -1.3 (0.9) | -1.4 | 172 | .17 | -3.2 | 0.5 |
| Think (vs control) | -0.6 (1.0) | -0.7 | 172 | .51 | -2.6 | 1.3 |
| Do x Time | -0.5 (0.4) | -1.3 | 513 | .19 | -1.3 | 0.3 |
| Think x Time | -0.2 (0.4) | -0.5 | 511 | .59 | -1.1 | 0.6 |
| Do x Time2 | 0.1 (0.1) | 1.5 | 459 | .13 | -0.0 | 0.2 |
| Think x Time2 | 0.0 (0.1) | 0.6 | 456 | .53 | -0.1 | 0.2 |

*Note.* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -0.7, *SE =* 1.0, *t*(170) = -0.7, *p* = .50; time: *B* = -0.3, *SE* = 0.4, *t*(503) = -0.7, *p* = .46; time2: *B* = 0.0, *SE* = 0.1, *t*(451) = 0.9, *p* = .38.

*Figure 4.*Mean scores for disturbances in social role functioning (OQ45, SR-scale) and predicted lines plotted across time (per-protocol analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

# Appendix G

Statistical analyses for empowerment

**Descriptive statistics**

The number of completed NEL assessments per measurement wave and corresponding means (SD) per group are provided in Table 1a and Table 1b, respectively.

Table 1*Number of completed assessments for NEL*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | Do-module | Think-module | Control |
| Baseline | 161 | 55 | 55 | 51 |
| Post-EMA | 132 | 47 | 44 | 41 |
| FU1 | 107 | 38 | 39 | 30 |
| FU2 | 95 | 31 | 30 | 34 |
| FU3 | 93 | 36 | 29 | 28 |
| FU4 | 87 | 31 | 29 | 27 |

Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

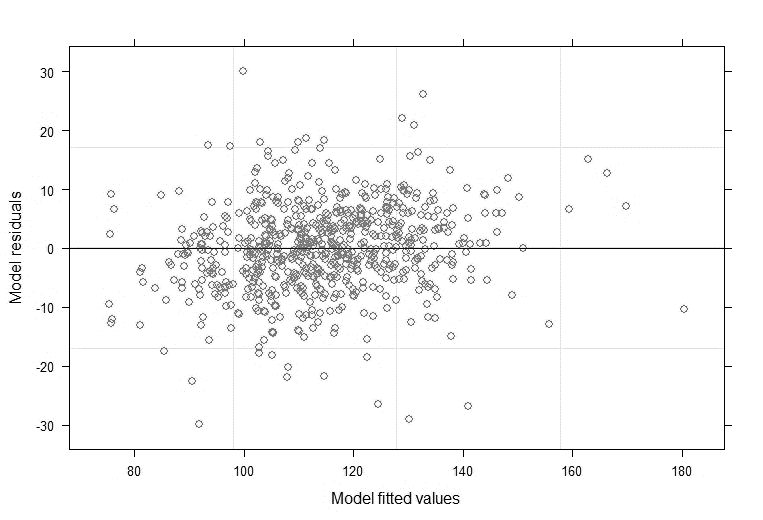
Table 1b *Observed means (SD) for empowerment (NEL)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Baseline | Post-EMA | FU1 | FU2 | FU3 | FU4 |
| Control | 109.9 (12.6) | 110.9 (14.2) | 112.7 (16.4) | 116.3 (15.8) | 114.0 (15.2) | 117.8 (15.5) |
| Do-module | 110.1 (15.4) | 116.4 (18.6) | 114.3 (24.1) | 118.4 (21.7) | 117.1 (20.5) | 119.4 (20.5) |
| Think-module | 110.5 (13.7) | 114.3 (16.5) | 115.2 (18.7) | 116.4 (18.1) | 115.5 (23.0) | 119.2 (20.7) |

*Note.* Means (SD) per group for the NEL total score without the professional help scale. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment.

**Assumption checks**

The assumptions of normality and independence of residuals were satisfied, as shown in Figure 1 by the equal spread around a mean of zero and the random pattern across the model fitted values. Individual plots of the residuals against each predictor showed linear relationships and no seasonal pattern for the residuals, satisfying all linear mixed model assumptions.



*Figure 1.* Scatterplot of model residuals plotted against model fitted values for empowerment (NEL).

**Intention-to-treat analysis**

Our main analysis was based on the intention-to-treat principle: participants were compared within the groups to which they were initially randomized, independently of having received the allocated treatment, having dropped out of the study or having violated the initial protocol (for whatever reason).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 2. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 2.

**Random effects.** The random effects, representing the differences between participants in baseline levels and changes over time across groups, indicated a large variance in empowerment starting levels (*S2* NEL intercept = 171.4) and changes over time (*S2* NEL slope *=* 2.1). This heterogeneity is further illustrated by the spread of the individual predicted regression lines in Figure 3. The intraclass correlation (ICC) indicating the proportion of variance accounted for by participants (level 2) was .74. This effectively means that three quarters of the residual variance is explained by between-person differences.

Table 2 *Intention-to-treat analysis* *for empowerment (NEL)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *Df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 110.4 (2.1) | 53.8 | 164 | < .001 | 106.4 | 114.5 |
| Time | 1.5 (0.4) | 3.8 | 105 | < .001 | 0.7 | 2.3 |
| Do (vs control) | 1.7 (2.8) | 0.6 | 163 | .56 | -3.9 | 7.2 |
| Think (vs control) | 0.8 (2.9) | 0.3 | 164 | .78 | -4.8 | 6.4 |
| Do x Time | -0.4 (0.5) | -0.7 | 106 | .46 | -1.4 | 0.7 |
| Think x Time | -0.1 (0.5) | -0.2 | 105 | .83 | -1.2 | 1.0 |

*Note.* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = 0.9, *SE =* 2.8, *t*(163) = 0.3, *p* = .76; time: *B* = -0.3, *SE* = 0.5, *t*(106) = -0.5, *p* = .60.

**

*Figure 2.*Mean scores for disturbances in empowerment (NEL) and predicted lines plotted across time (intention-to-treat analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.



*Figure 3.*Individual predicted lines for empowerment (NEL) within each group. The bold line represents the mean predicted line for the group. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

**Per-protocol analysis**

In addition to the intention-to-treat analysis, we examined the value of the add-on tool exclusively among participants who did not drop out of the study (Do-module: *n ­=* 48, Think-module: *n* = 42).

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 3. The fixed effects show the results for the typical participant and are represented by the three regression lines in Figure 4.

**Random effects.** The random effects, indicated a large variance in NEL starting levels (*S2* NEL intercept =180.7) and changes over time (*S2* NEL slope *=* 2.2). The intraclass correlation (ICC) was .75.

Table 3 *Fixed effects of the per-protocol analysis for empowerment (NEL)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 110.5 (2.1) | 52.6 | 147 | < .001 | 106.4 | 114.6 |
| Time | 1.5 (0.4) | 3.8 | 102 | < .001 | 0.7 | 2.3 |
| Do (vs control) | 2.1 (3.0) | 0.7 | 144 | .49 | -3.8 | 8.0 |
| Think (vs control) | 0.9 (3.1) | 0.3 | 143 | .77 | -5.2 | 7.0 |
| Do x Time | -0.3 (0.5) | -0.6 | 103 | .56 | -1.4 | 0.8 |
| Think x Time | -0.1 (0.6) | -0.2 | 101 | .86 | -1.2 | 1.0 |

*Note.* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = 1.2, *SE =* 3.1, *t*(140) = 0.4), *p* = .71; time: *B* = -0.2, *SE* = 0.5, *t*(102) = -0.4, *p* = .69.

**

*Figure 4.*Mean scores for empowerment (NEL) and predicted lines plotted across time (per-protocol analysis). Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

**Secondary analysis on alternative total empowerment score**

To compare the empowerment analysis with a previous trial (Simons et al., 2015), we recomputed the total empowerment score by averaging all available items (*i.e.,* either 36 or 40 depending on whether the professional help scale was applicable) and multiplying by 40.

**Fixed effects.** The results of the multilevel regression analysis are presented in Table 4. In contrast to the original empowerment score, the quadratic trend improved model fit.

**Random effects.** The random effects, indicated a large variance in NEL starting levels (*S2* NEL intercept =202.2) and changes over time (*S2* NEL slope *=* 2.3). The intraclass correlation (ICC) was .74.

Table 4 *Fixed effects of the per-protocol analysis for empowerment (adjusted NEL)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Intercept | 123.1 (2.3) | 53.3 | 193 | < .001 | 118.6 | 127.7 |
| Time | 3.4 (1.0) | 3.4 | 530 | < .001 | 1.4 | 5.3 |
| Time2 | -0.2 (0.1) | -1.8 | 478 | .07 | -0.5 | 0.0 |
| Do (vs control) | 1.5 (3.2) | 0.5 | 192 | .64 | -4.8 | 7.8 |
| Think (vs control) | 1.6 (3.2) | 0.5 | 193 | .62 | -4.7 | 7.9 |
| Do x Time | -0.1 (1.4) | -0.1 | 528 | .91 | -2.8 | 2.5 |
| Think x Time | -1.2 (1.4) | -0.8 | 528 | .40 | -3.9 | 1.6 |
| Do x Time2 | -0.0 (0.2) | -0.2 | 477 | .84 | -0.4 | -0.3 |
| Think x Time2 | 0.2 (0.2) | 0.8 | 476 | .40 | -0.2 | 0.5 |

*Note.* The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in baseline *B* = -0.1, *SE =* 3.1, *t*(189) = 0.0, *p* = .97; time: *B* = 1.0, *SE* = 1.3, *t*(525) = 0.7, *p* = .45, time2: *B* = -0.2, *SE* = 0.2, *t*(475) = 1.1, *p* = .28.

**Reference**

Simons, C. J. P., Hartmann, J. A., Kramer, I., Menne-Lothmann, C., Höhn, P., van Bemmel, A. L., … Wichers, M. (2015). Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. *European Psychiatry*, *30*, 900–906. doi: 10.1016/j.eurpsy.2015.09.004

# Appendix H

Post-hoc analyses

We explored depressive symptom trajectories in subgroups based on two treatment characteristics: compliance with the intervention (H1), and time to the start of treatment (*i.e.*, the first psychotherapy session, H2). Furthermore, we merged the figure comprising depressive outcome data of the experimental group and control group of the previous RCT (Kramer et al., 2014) with our outcome data (H3). We originally did not run any post-hoc analyses including covariates to study the impact on efficacy estimates, but added such analyses upon request by one of the reviewers (H4). For similar reasons, we added comparisons between baseline and all five individual follow-up measurements (H5).

1. **Compliance**

Regarding the self-assessments, approximately half of the participants in the experimental groups were highly compliant (*i.e.,* ≥ 75%). Figure 1 shows depressive symptom declines in the highly compliant participants (*n* = 54) and less compliant participants (*n* = 56) compared to the control participants. We observed a more favourable course in the highly compliant group compared to the less compliant group, but the intermediate trajectory of the control group (who did not complete any repeated self-assessments) suggests that compliance is rather a marker for a favourable course than its cause.

**

*Figure 1.*Mean scores for depressive symptom severity (IDS-SR) for highly compliant (≥ 75%), less compliant, and control participants plotted across time. Note that the two intervention groups were combined to create compliance subgroups. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

1. **Time to treatment start**

Most participants started psychotherapy before the end of the intervention period (*i.e.,* six weeks after study intake); a quarter did not start psychotherapy until afterwards (or not at all). Figure 2 shows that participants who engaged in the EMI while waiting for psychotherapy (late TAU intervention: *n* = 27) showed early symptom declines, but similar declines were seen in controls waiting for psychotherapy (late TAU controls: *n* = 11). Similar early declines were also seen in participants that did start psychotherapy early on (early TAU control: *n* = 39, early TAU intervention: *n* = 76). Subgroups were small, however, and the waitlist condition was not ‘clean’: before the start of psychotherapy patients turned out to often have had other appointments (*e.g.*, diagnostic testing, medication consults), which could already have set the recovery process in motion. Hence, the hypothesis that the EMI might have added value in the absence of TAU could not be adequately tested.



*Figure 2.*Mean scores for depressive symptom severity (IDS-SR) for early TAU (TAU start before the end of the intervention period) and late TAU (TAU start after intervention period) participants. For the intervention group the Do-module and Think-module were combined. Post-EMA = assessment in the week after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5-6 weeks rather than one month.

1. **ZELF-i and REMOD-ID study**

Figure 3 shows average depressive symptom scores of the experimental and control groups for both the ZELF-i and REMOD-ID study (Kramer et al., 2014). The experimental group of the REMOD-ID study had a slightly lower starting point and end point. However, the decline in the experimental groups was rather comparable across studies: in both studies the average decline in depression severity from baseline to the 6-month follow-up measurement was approximately 11 points on the IDS-SR.

*Figure 3.*Mean scores for depressive symptom severity (IDS-SR) for the experimental and control groups of the REMOD-ID and ZELF-i study plotted across time in weeks. Mean scores were connected by lines to improve readability. REMOD-ID: baseline, post-EMA (week 7), and five assessments during a 6-month follow-up (FU) period (after 1 week, 1, 2, 3 and 6 months). ZELF-i: baseline, post-EMA (week 5), and four FU assessments at 1, 2, 3 and 6 months. REMOD-ID figure adapted from Kramer *et al*. (2014). Copyright 2014, World Psychiatric Association.

1. **Covariates**

Upon reviewer request, we ran post-hoc analyses to study the impact of covariates on efficacy estimates, namely educational level (low: no/primary/low secondary, middle: high school/low vocational, high: higher vocational/university), daytime activity (yes (paid work, study, household) or no), change in antidepressant use between study intake and post-EMA assessment (yes/no), early or late start of TAU (*i.e.,* start of psychotherapy before or after debriefing), and whether participants received cognitive behavioral therapy as part of TAU (yes/no). Note that for some variables we reduced the number of categories to maintain sufficient power. Type of TAU, for instance, would otherwise comprise as many categories as there are participants (the combination of treatments is quite diverse). Tables 1-4 show estimates for the two-way interactions between time and group for each of the four outcome measures based on an intention-to-treat analysis with covariates. Results were essentially the same as those of the original model, with none of the interaction terms reaching significance. This indicates that the simple comparison of treatment groups was reasonable.

Table 1 *Two-way interaction between time and group for depressive symptom severity (IDS-SR) with covariates*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Do x Time | 0.1 (1.0) | 0.1 | 537 | .94 | -1.8 | 2.0 |
| Think x Time | 0.7 (1.0) | 0.7 | 540 | .47 | -1.2 | 2.6 |
| Do x Time2 | 0.0 (0.1) | 0.3 | 469 | .79 | -0.2 | 0.3 |
| Think x Time2 | -0.1 (0.1) | -0.5 | 469 | .61 | -0.3 | 0.2 |

*Note*. Numbers represent the unstandardized beta (B) with its standard error (SE) in parentheses and an asterisk if *p* < .05. The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in time: *B* = -0.6, *t*(538) = -0,7 , *p* = .50, time2: *B* = 0.1, *t*(470) = 0.8, *p* = .42.

Table 2 *Two-way interaction between time and group for disturbances in interpersonal relations (OQ45 IR scale) with covariates*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Do x Time | 0.2 (0.2) | 0.9 | 85 | .37 | -0.2 | 0.5 |
| Think x Time | -0.1 (0.2) | -0.4 | 84 | .72 | -0.5 | 0.3 |

*Note*. Numbers represent the unstandardized beta (B) with its standard error (SE) in parentheses and an asterisk if *p* < .05. The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in time: *B* = 0.2, *t*(86) = 1.3 , *p* = .50.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Do x Time | -0.5 (0.4) | -1.2 | 533 | .23 | -1.3 | 0.3 |
| Think x Time | -0.3 (0.4) | -0.8 | 535 | .41 | -1.2 | 0.5 |
| Do x Time2 | 0.1 (0.1) | 1.5 | 476 | .13 | -0.0 | 0.2 |
| Think x Time2 | 0.0 (0.1) | 0.8 | 477 | .42 | -0.1 | 0.2 |

Table 3 *Two-way interaction between time and group for disturbances in social role functioning (OQ45 SR scale) with covariates*

*Note*. Numbers represent the unstandardized beta (B) with its standard error (SE) in parentheses and an asterisk if *p* < .05. The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in time: *B* = -0.1, *t*(531) = -0.4, *p* = .71, time2: *B* = 0.0, t(477) = 0.7, *p* = .49.

Table 4 *Two-way interaction between time and group for empowerment (NEL) with covariates*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fixed effects | *B* (*SE*) | *t* | *df* | *P* | *CI95* | |
| Lower bound | Upper bound |
| Do x Time | -0.4 (0.5) | -0.7 | 106 | .46 | -1.4 | 0.7 |
| Think x Time | -0.1 (0.5) | -0.2 | 105 | .84 | -1.2 | 1.0 |

*Note*. Numbers represent the unstandardized beta (B) with its standard error (SE) in parentheses and an asterisk if *p* < .05. The control group is the reference group. Estimates of Think vs Do (obtained by recoding Think as reference group): difference in time: *B* = -0.3, *t*(106) = -0,5 , *p* = .59.

1. **Comparisons between baseline and separate follow-up measurements**

Upon reviewer request, we added comparisons between baseline and each of the five follow-up measurements (post-EMA, FU1-4). Means (SD) per group for each measurement wave can be found in Table 1b of Appendices D-H. For each outcome measure, each individual measurement (*e.g.,* FU1) was regressed onto the baseline measure with group as added predictor. were used to compare group differences in changes between two time points (*e.g.*, baseline vs post-EMA). Table 5 shows that baseline was a highly significant predictor for each of the follow-up measurements (for all outcome measures). More importantly, Table 5 shows that group differences were small for each of the individual comparisons (for all outcome measures); none of the group comparisons reached an uncorrected significance level of .05, let alone after Bonferroni adjustment (*p* < .01).

Table 5 *Direct comparisons between baseline and each of the post intervention measurements*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Post-EMA | FU1 | FU2 | FU3 | FU4 |
| *Depressive symptom severity (IDS-SR)* | | | | | |
| Baseline *F* (df1, df2) | 77.9\* (1,129) | 49.0\* (1,106) | 35.5\* (1,95) | 30.6\* (1,92) | 15.0\* (1,85) |
| Group *F* (df1, df2) | 0.3 (2,129) | 0.0 (2,106) | 0.0 (2,95) | 0.4 (2,92) | 0.1 (2,85) |
| *Disturbances in interpersonal relations (OQ-45 IR scale)* | | | | | |
| Baseline *F* (df1, df2) | 154.5\* (1,128) | 63.6\* (1,104) | 64.3\* (1,92) | 42.6\* (1,89) | 35.4\* (1,84) |
| Group *F* (df1, df2) | 0.6 (2,128) | 0.5 (2,104) | 0.5 (2,92) | 1.2 (2,89) | 0.9 (2,84) |
| *Disturbances in social role functioning (OQ-45 SR scale)* | | | | | |
| Baseline *F* (df1, df2) | 62.6\* (1,128) | 67.5\* (1,104) | 33.6\* (1,92) | 36.2\* (1,89) | 18.7\* (1,84) |
| Group *F* (df1, df2) | 2.3 (2,128) | 3.2 (2,104) | 2.0 (2,92) | 0.4 (2,89) | 0.5 (2,84) |
| *Empowerment (NEL)* |  |  |  |  |  |
| Baseline *F* (df1, df2) | 96.0\* (1,128) | 81.4\* (1,103) | 58.4\* (1,91) | 74.2\* (1,89) | 32.5\* (1,83) |
| Group *F* (df1, df2) | 1.1 (2,128) | 0.0 (2,103) | 0.0 (2,91) | 0.1 (2.89) | 0.1 (2,83) |

*Note.* Post-EMA = assessment after the 28-day intervention period. FU = follow-up assessment 1, 2, 3 and 6 months following the post-EMA assessment. *\* = p* < .0001

**Reference**

Kramer I., Simons C. J. P., Hartmann J. A., Menne-Lothmann C., Viechtbauer W., Peeters F., … Wichers, M. (2014). A therapeutic application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry,* *13*, 68–77. doi: 10.1002/wps.20090