Nottingham Trent University
Psychology

Analysis of Project-100 Closed Cases: October, 2024

Authors: Mark Andrews

Version: v2.0_3.10.24-0-g7666dc0

Date: 3 October, 2024

Abstract

The reports describes a set of analyses of Project-100 closed cases using data collected between 1/4/2022 and 1/7/2024 (using raw dataset `PTSDRes-1.4.2022-1.7.2024-V1.csv`). As with the previous analysis published in October 2023, we analysed clients' improvement while in therapy in terms of the GAD-7, PHQ-9, PCL-5, and CORE-10 outcome variables, and analysed how clients' scores on these outcomes changed when measured at follow-up after the therapy concluded. We also analysed which variables are reliable predictors of improvements on the outcome variables or of drop-out from therapy, the effectiveness of the rewind intervention method, client and therapist feedback scores

1 Client details

There were 468 clients with closed cases. For some or all clients, we have information about their gender, age, ethnicity, employment status, military service history, living arrangements, medication use, and how they were referred to their treatment. For their employment status, military service history, living arrangements, medication use, and referral status, their responses, if any, were in free form. This made it difficult or impossible to have a limited and meaningful number of values for each variable and as such, it was necessary to simplify or aggregate their responses concerning these variables, as we explain below.

1.1 Gender

Of all the 468 clients, 356 (76.1%) were listed as male, 104 (22.2%) were listed as female, and for 8 (1.7%) others, their gender was not stated.

1.2 Age

Of the 468 clients, 462 provided their age. Their median age was 46, with 95% of cases being between 25 and 71.95. Their mean age was 46.84 with a standard deviation was 12.02. The minimum age was 18 and the maximum age was 81.

1.3 Ethnicity

Overall, 425 clients (90.8%) stated their ethnicity as *White British*. Of those remaining, 16 did not state their ethnicity, 8 stated their ethnicity as *Other White*, 5 as *Other Mixed*. The total breakdown by ethnicity category is as follows:

Ethnicity	n
White British	425 (90.8%)
Not known/stated	16 (3.4%)
Other White	8 (1.7%)
Other Mixed	5 (1.1%)
Black British	3 (0.6%)
Carribean	3 (0.6%)
Any Other	2 (0.4%)
White Irish	2 (0.4%)
White and Black Carribean	2 (0.4%)
African	1 (0.2%)
Other Asian	1 (0.2%)

1.4 Employment status

If clients stated their employment status as either “employed”, “self-employed”, or “student”, they were classified as “employed”. If they stated their employment status as “long-term sick” or “short-term sick”, their employment status was classified as “sick”. If they stated they were seeking work, their status was classified as “seeking”, and if they stated they were either retired or not seeking work, they were classified as “not seeking”. Any other responses concerning employment were classified as “other”. The breakdown of these categories is as follows:

Employment status	n
employed	272 (58.1%)
sick	89 (19%)
not seeking	49 (10.5%)
other	35 (7.5%)
seeking	23 (4.9%)

1.5 Military service

At screening, 453 clients provided a free text response to the question *Where did you serve?* Their responses were simplified into three categories: “combat”, meaning that they experienced active military duty in some military conflict theatre; “no combat”, meaning that they did not serve in a military conflict theatre; “other”,meaning any other response. Clients who mentioned in their free response terms like *Northern Ireland*, *NI*, *Afghan(istan)*, *Gulf*, *Iraq*, *Falklands*, *Yugoslavia*, *Bosnia*, or *Kosovo* were classified as “combat”. If they explicitly stated *no active* tours, they were classified as “no combat”. All other responses, were classified as “other”. The breakdown of these categories is as follows:

Military service	n
combat	301 (66.4%)
other	92 (20.3%)
no combat	60 (13.2%)

1.6 Living arrangements

Many clients stated that their living arrangements were one of the following: “with partner”, “house/flat share”, “with parents”, “alone”, “homeless”. The first three of these responses were relabelled as “partner”, “share”, “parents”, respectively, and the response “Wife and children” was also classified as “partner”. All other responses, each of which had only one response across clients, were classified as “other”. The breakdown is as follows:

Living arrangements	n
partner	226 (48.3%)
alone	146 (31.2%)
other	38 (8.1%)
parents	31 (6.6%)
shared	16 (3.4%)
homeless	11 (2.4%)

1.7 Medication use

In response to a question about medication use, 170 (36.3%) clients indicated they were taking anti-depressants, and 4 indicated they were taking anti-psychotics. All others indicated they were taking no medication, or "other" medication, or did not respond. Because of the absence of much detail about medication use, we treated this as just a binary variable that indicated whether the clients was taking anti-depressants or not. As mentioned, 170 (36.3%) stated they were taking anti-depressants and so 298 (63.7%) were recorded as not taking anti-depressants.

1.8 Referral status

In response to a question about how they were referred for treatment, most either explicitly indicated that they self-referred or else they did not respond at all. However, those case where there was no response, were treated as self-referrals. The remaining clients listed many different sources for their referral, such as *Care After Combat*, *RBLI*, *Red Poppy Factory*, but each of these other values usually had a very small number of respondents, often just one, each. To simplify, therefore, we treat referral status as a binary variable indicating whether a client self-referred or not. Overall, 265 (56.6%) self-referred, and so 203 (43.4%) did not.

2 Session details

In total, 452 (96.6%) clients did one or more (post-screening, prior to follow-up) therapy sessions with one of 73 different therapists. The median number of sessions done by these clients was 7, and 90% of clients did between 2 and 14 sessions. Of these 452 clients, 93 (20.6%) did not complete the therapy programme as planned, i.e., they dropped-out before completion.

3 Analysis of improvements in outcome measures

For the GAD-7, PHQ-9, PCL-5, and CORE-10 outcome variables, we summarized and analysed improvement while in therapy, i.e. from initial assessment to final assessment excluding assessment done at follow-up times, using an IAPT style approach. For each variable, we calculated the number of clients who can be included in the analysis, the number of initial clinical "cases", the number of recovered cases, the number of reliably recovered cases, the number of clients who reliably improve. We also analysed improvement using a paired-samples t-test and calculated the Cohen's d[1] and Hedges' g score. These results are summarized in Table 1 and are explained in more detail in the following subsections.

In addition, following standard IAPT guidelines (see NHS England, 2019, p. 53), we report the number of clients who are initially defined as cases by either GAD-7 or PHQ-9 definitions and who are then recovered cases on both measures at the end of therapy. In particular, we used the following definitions:

[1] In the context of paired initial and final scores, Cohen's d effect size is defined, see Cohen (1988), as follows:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 + s_2^2 - 2rs_1s_2}} \times \sqrt{2(1-r)},$$

where $\bar{x}_1$ is the mean of the initial scores, $\bar{x}_2$ is the mean of the final scores, s_1 is the standard deviation of the initial scores, s_2 is the standard deviation of the final scores, and r is the Pearson correlation coefficient between the initial and the final scores.

- A client is defined as a combined GAD-7 and PHQ-9 case if they are defined as a case on GAD-7 or PHQ-9 or both.
- A client is defined as a recovered combined GAD-7 and PHQ-9 case if at the end of therapy, they are not a case on GAD-7 and also not a case on PHQ-9, i.e. below the caseness threshold on both measures.
- A client is defined as reliably improved if at the end of therapy their GAD-7 score or their PHQ-9 or both have improved by more than the improvement threshold, and neither their GAD-7 nor their PHQ-9 scores have not reliably deteriorated[2].
- A client is defined as reliably recovered if they are both recovered and have reliably improved.

Likewise, we analyse the improvement scores in the PRN-14 outcome measures, discuss some of its psychometric properties of PRN-14, and analyse the degree of correlation between the improvements as measured by GAD-7, PHQ-9, PCL-5, CORE-10, and PRN-14.

3.1 GAD-7

- There were 422 clients who provided initial measurements of GAD-7. Of these, 379 (89.8%) clients were measured at least one more time and so can be included in this analysis.
- Of the clients for whom we have at least two measures, 337 (88.9%) were cases, defined as having a score of 8 or above, at the time of their first measurement.
- Of those clients who were initial cases, 183 (54.3%) had recovered, i.e. were no longer cases, by the time of their final measurement.
- Of these recovered cases, 175 (51.9%) had reliably recovered in that they had both recovered and their score had improved by at least 4 points.
- Of all the clients with two or more GAD-7 measures, 246 (64.9%) reliably improved in that their score improved by at least 4 points from the initial to the final measurement, regardless of whether their initial or final measurement was above or below the caseness threshold.
- The average (and standard deviation) improvement from first to final measure was 6.06 (5.9).
- A paired t-test comparing the initial and final scores shows that there was a highly significant improvement on average: $t(378) = 19.97, p < 0.001$.
- The Cohen's d effect size for the average improvement is d = 1.14. The Hedges' g correction of this Cohen's d effect size is g = 1.14.

3.2 PHQ-9

- There were 406 clients who provided initial measurements of PHQ-9. Of these, 362 (89.2%) clients were measured at least one more time and so can be included in this analysis.
- Of the clients for whom we have at least two measures, 291 (80.4%) were cases, defined as having a score of 10 or above, at the time of their first measurement.
- Of those clients who were initial cases, 154 (52.9%) had recovered, i.e. were no longer cases, by the time of their final measurement.
- Of these recovered cases, 142 (48.8%) had reliably recovered in that they had both recovered and their score had improved by at least 6 points.
- Of all the clients with two or more PHQ-9 measures, 191 (52.8%) reliably improved in that their score improved by at least 6 points from the initial to the final measurement, regardless of whether their initial or final measurement was above or below the caseness threshold.
- The average (and standard deviation) improvement from first to final measure was 6.69 (6.91).
- A paired t-test comparing the initial and final scores shows that there was a highly significant improvement on average: $t(361) = 18.41, p < 0.001$.

[2]From the IAPT manual, "Patients are considered reliably improved if their scores for depression and/or anxiety have reduced by a reliable amount and neither measure has shown a reliable increase."

Table 1: IAPT style analysis of GAD-7, PHQ-9, PLC-5, and CORE-10 outcomes.

	GAD-7	PHQ-9	PCL-5	CORE-10
Initial n	422	406	403	427
Included n	379 (89.8%)	362 (89.2%)	293 (72.7%)	383 (89.7%)
Initial cases	337 (88.9%)	291 (80.4%)	223 (76.1%)	357 (93.2%)
Recovered cases	183 (54.3%)	154 (52.9%)	139 (62.3%)	159 (44.5%)
Reliable recovery	175 (51.9%)	142 (48.8%)	134 (60.1%)	151 (42.3%)
Reliable improvement	246 (64.9%)	191 (52.8%)	200 (68.3%)	251 (65.5%)
Average improvement	6.06 (5.9)	6.69 (6.91)	19.07 (17.81)	8.49 (8.13)
Paired t-test	t(378) = 19.97, p < 0.001	t(361) = 18.41, p < 0.001	t(292) = 18.33, p < 0.001	t(382) = 20.46, p < 0.001
Cohen's d	1.14	0.98	1.05	1.09
Hedges' g	1.14	0.98	1.04	1.08

- The Cohen's d effect size for the average improvement is d = 0.98. The Hedges' g correction of this Cohen's d effect size is g = 0.98.

3.3 PCL-5

- There were 403 clients who provided initial measurements of PCL-5. Of these, 293 (72.7%) clients were measured at least one more time and so can be included in this analysis.
- Of the clients for whom we have at least two measures, 223 (76.1%) were cases, defined as having a score of 32 or above, at the time of their first measurement.
- Of those clients who were initial cases, 139 (62.3%) had recovered, i.e. were no longer cases, by the time of their final measurement.
- Of these recovered cases, 134 (60.1%) had reliably recovered in that they had both recovered and their score had improved by at least 10 points.
- Of all the clients with two or more PCL-5 measures, 200 (68.3%) reliably improved in that their score improved by at least 10 points from the initial to the final measurement, regardless of whether their initial or final measurement was above or below the caseness threshold.
- The average (and standard deviation) improvement from first to final measure was 19.07 (17.81).
- A paired t-test comparing the initial and final scores shows that there was a highly significant improvement on average: $t(292) = 18.33, p < 0.001$.
- The Cohen's d effect size for the average improvement is d = 1.05. The Hedges' g correction of this Cohen's d effect size is g = 1.04.

3.4 CORE-10

- There were 427 clients who provided initial measurements of CORE-10. Of these, 383 (89.7%) clients were measured at least one more time and so can be included in this analysis.
- Of the clients for whom we have at least two measures, 357 (93.2%) were cases, defined as having a score of 11 or above, at the time of their first measurement.
- Of those clients who were initial cases, 159 (44.5%) had recovered, i.e. were no longer cases, by the time of their final measurement.
- Of these recovered cases, 151 (42.3%) had reliably recovered in that they had both recovered and their score had improved by at least 6 points.
- Of all the clients with two or more CORE-10 measures, 251 (65.5%) reliably improved in that their score improved by at least 6 points from the initial to the final measurement, regardless of whether their initial or final measurement was above or below the caseness threshold.
- The average (and standard deviation) improvement from first to final measure was 8.49 (8.13).
- A paired t-test comparing the initial and final scores shows that there was a highly significant improvement on average: $t(382) = 20.46, p < 0.001$.

- The Cohen's d effect size for the average improvement is d = 1.09. The Hedges' g correction of this Cohen's d effect size is g = 1.08.

3.5 GAD-7 & PHQ-9 combined

- There were 405 clients who provided initial measurements of both GAD-7 and PHQ-9. Of these, 361 (89.1%) clients were measured at least one more time on both and so can be included in this analysis.
- Of the clients for whom we have at least two measures on both, 332 (92%) were cases, defined as being a case on either GAD-7 or PHQ-9 or on both, at the time of their initial measurements.
- Of those clients who were initial cases, 161 (48.5%) had recovered, i.e. they were not cases on both GAD-7 and PHQ-9, by the time of their final measurement.
- Of these recovered cases, 155 (46.7%) had reliably recovered in that they had recovered by the end of therapy, i.e. were not cases on both GAD-7 and PHQ-9, and either their GAD-7 or PHQ-9 scores or both had reliably improved and neither scores had reliably deteriorated.

These results are summarized in the following table:

	GAD-7 & PHQ-9
Initial n	405
Included n	361 (89.1%)
Initial cases	332 (92%)
Recovered cases	161 (48.5%)
Reliable recovery	155 (46.7%)

3.6 PRN-14

- There were 400 clients who provided initial measurements of PRN-14. Of these, 295 (73.8%) clients were measured at least one more time and so can be included in this analysis.
- The average (and standard deviation) improvement[3] from first to final measure was 29.78 (30.7).
- A paired t-test comparing the initial and final scores shows that there was a highly significant improvement on average: $t(294) = 16.66, p < 0.001$.
- The Cohen's d effect size for the average improvement is d = 0.966.
- The Hedges' g correction of this Cohen's d effect size is g = 0.963.
- A *parallel analysis* method to identify the optimal number of factors in an exploratory factor analysis of the 14 items suggest that the optimal number is 3, and so there are three underlying latent variables onto which all items load.
- The Cronbach's α score of internal consistency/reliability of the PRN-14 items is high: α = 0.97 (95% confidence interval: 0.963-0.968).

3.7 P-value adjustment

In Table 1 and the the preceding paragraphs, the p-values from the t-tests testing the average improvements in GAD-7, PHQ-9, PCL-5, CORE-10, and PRN-14 are all the original or unadjusted p-values. These should be adjusted for the fact that five independent t-tests were conducted.

[3] Unlike with GAD-7, PHQ-9, PCL-5, CORE-10, higher scores in the PRN-14 measure are better outcomes. To simplify comparisons with the other measures, improvements in PRN-14 are calculated by subtracting the initial score from the final score, while in the other measures, improvement scores are calculated from subtracting the final score from the initial score.

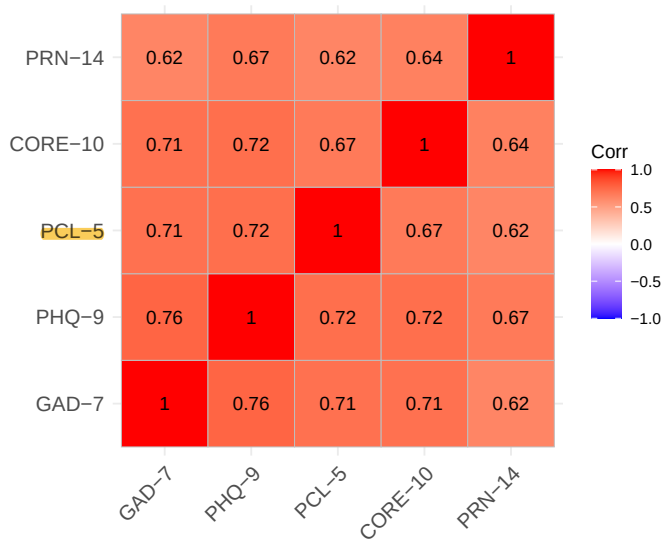

Figure 1: Intercorrelation of improvement scores in five outcome variables.

Applying the conservative Bonferroni correction, all p-values are increased by a factor of 5, the number of tests. Even after this correction, all p-values are far below the lowest conventional p-value thresholds: $\max p \ll 0.001$.

3.8 Inter-correlations of improvements across outcomes

There is a relatively high degree of correlation between the improvements on GAD-7, PCL-5, PHQ-9, CORE-10, PRN-14, as can be seen in the correlation matrix heatmap in Figure 1. All of these correlations are highly significant ($\max p \ll 0.001$).

4 Analysis of outcomes at follow-up

Some clients had some or all of the GAD-7, PHQ-9, PCL-5, CORE-10, and PRN-14 outcomes variables measured at a follow-up time after normal therapy ended. The first follow-up was normally approximately 3 months later. Here, we analyse the improvement or deterioration of scores of these outcomes variables between the last session of normal therapy and the first follow-up. Sometimes the first follow-up occurred 6 months or more after therapy ended, and so in this analysis, we will restrict ourselves to follow-up sessions no later than 7 months after therapy ended.

4.1 GAD-7

- Of the 379 clients for whom we measured GAD-7 on at least two occasions, we also have follow-up scores for 156 (41.2%) for them.
- The median number of months between the final measurement of GAD-7 and the first follow-up session was 3.19.
- Of the all the clients with follow-up scores, 143 (91.7%) had their first follow-up session no more than 7 months after their final normal therapy session.
- Restricting this analysis to just those clients whose first follow-up session occurred within 7 months, the average (and standard deviation) of their improvement scores is -1.66 (5.39), which is an average deterioration.

- A paired samples t-test comparing scores at the last normal therapy session and at the first follow-up showed that the average change in scores was highly significant: t(142) = -3.69, p <0.001.
- The Cohen's d effect-size of the average deterioration is 0.32, and the Hedges' g is 0.32.

4.2 PHQ-9

- Of the 362 clients for whom we measured PHQ-9 on at least two occasions, we also have follow-up scores for 153 (42.3%) for them.
- The median number of months between the final measurement of PHQ-9 and the first follow-up session was 3.23.
- Of the all the clients with follow-up scores, 140 (91.5%) had their first follow-up session no more than 7 months after their final normal therapy session.
- Restricting this analysis to just those clients whose first follow-up session occurred within 7 months, the average (and standard deviation) of their improvement scores is -1.86 (6.37), which is an average deterioration.
- A paired samples t-test comparing scores at the last normal therapy session and at the first follow-up showed that the average change in scores was highly significant: t(139) = -3.45, p <0.001.
- The Cohen's d effect-size of the average deterioration is 0.28, and the Hedges' g is 0.28.

4.3 PCL-5

- Of the 293 clients for whom we measured PCL-5 on at least two occasions, we also have follow-up scores for 155 (52.9%) for them.
- The median number of months between the final measurement of PCL-5 and the first follow-up session was 3.45.
- Of the all the clients with follow-up scores, 137 (88.4%) had their first follow-up session no more than 7 months after their final normal therapy session.
- Restricting this analysis to just those clients whose first follow-up session occurred within 7 months, the average (and standard deviation) of their improvement scores is -2.93 (18.36), which is an average deterioration.
- A paired samples t-test comparing scores at the last normal therapy session and at the first follow-up showed that the average change in scores was non-significant: t(136) = -1.87, p = 0.064.
- The Cohen's d effect-size of the average deterioration is 0.17, and the Hedges' g is 0.16.

4.4 CORE-10

- Of the 383 clients for whom we measured CORE-10 on at least two occasions, we also have follow-up scores for 160 (41.8%) for them.
- The median number of months between the final measurement of CORE-10 and the first follow-up session was 3.19.
- Of the all the clients with follow-up scores, 148 (92.5%) had their first follow-up session no more than 7 months after their final normal therapy session.
- Restricting this analysis to just those clients whose first follow-up session occurred within 7 months, the average (and standard deviation) of their improvement scores is -1.97 (7.51), which is an average deterioration.
- A paired samples t-test comparing scores at the last normal therapy session and at the first follow-up showed that the average change in scores was highly significant: t(147) = -3.2, p = 0.0017.
- The Cohen's d effect-size of the average deterioration is 0.24, and the Hedges' g is 0.24.

4.5 PRN-14

- Of the 295 clients for whom we measured PRN-14 on at least two occasions, we also have follow-up scores for 150 (50.8%) for them.
- The median number of months between the final measurement of PRN-14 and the first follow-up session was 3.42.
- Of the all the clients with follow-up scores, 134 (89.3%) had their first follow-up session no more than 7 months after their final normal therapy session.
- Restricting this analysis to just those clients whose first follow-up session occurred within 7 months, the average (and standard deviation) of their improvement scores is -9.66 (34.1), which is an average deterioration.
- A paired samples t-test comparing scores at the last normal therapy session and at the first follow-up showed that the average change in scores was highly significant: t(133) = -3.28, p = 0.0013.
- The Cohen's d effect-size of the average deterioration is 0.28, and the Hedges' g is 0.28.

4.6 P-value adjustment

As was done for the p-values for the t-tests of the outcome improvements in Section 3, the p-values for the five t-tests just reported should be likewise adjusted. Adjusting by the conservative Bonferroni adjustment, the lowest p-value is now a p-value of 0.002 for the GAD-7 t-test. Adjusting by the less conservative Benjamini-Hochberg False Discovery Rate (FDR) method, the lowest p-value is now a p-value of 0.002 for the GAD-7 t-test.

For each outcome, the adjusted p-values are as follows:

Measure	Bonferroni adjusted p-value	FDR adjusted p-value
GAD-7	0.002*	0.002*
PHQ-9	0.004*	0.002*
PCL-5	0.318	0.064
CORE-10	0.009*	0.002*
PRN-14	0.007*	0.002*

In this table, the $*$ indicates a p-value less than 0.05.

5 Analysis of outcomes from first to last follow-up

Some clients had some or all of the GAD-7, PHQ-9, PCL-5, CORE-10, and PRN-14 outcomes variables measured in more than one follow-up sessions. Here, we analyse the improvement or deterioration of these outcomes variables between the first and last follow-up session. We restricted the analysis to those clients whose first follow-up session was within 7 months of therapy ending and where the last follow-up session was at least 3 months after the first follow-up session.

5.1 GAD-7

- Of the 379 clients for whom we measured GAD-7 on at least two occasions, we also have more than one follow-up scores for 60 (15.8%) of them.
- The median number of months between the first and last follow-up session was 8.96.
- The average (and standard deviation) of their improvement scores from the first to last follow-up is -0.63 (4.13), which is an average deterioration.

- A paired samples t-test comparing scores at the first and last follow-up sessions showed that the average change in scores was non-significant: t(59) = -1.19, p = 0.24.
- The Cohen's d effect-size of the average deterioration is 0.11, and the Hedges' g is 0.11.

5.2 PHQ-9

- Of the 362 clients for whom we measured PHQ-9 on at least two occasions, we also have more than one follow-up scores for 57 (15.7%) of them.
- The median number of months between the first and last follow-up session was 8.97.
- The average (and standard deviation) of their improvement scores from the first to last follow-up is -0.46 (5.18), which is an average deterioration.
- A paired samples t-test comparing scores at the first and last follow-up sessions showed that the average change in scores was non-significant: t(56) = -0.66, p = 0.51.
- The Cohen's d effect-size of the average deterioration is 0.06, and the Hedges' g is 0.06.

5.3 PCL-5

- Of the 293 clients for whom we measured PCL-5 on at least two occasions, we also have more than one follow-up scores for 57 (19.5%) of them.
- The median number of months between the first and last follow-up session was 8.97.
- The average (and standard deviation) of their improvement scores from the first to last follow-up is -1.21 (12.56), which is an average deterioration.
- A paired samples t-test comparing scores at the first and last follow-up sessions showed that the average change in scores was non-significant: t(56) = -0.73, p = 0.47.
- The Cohen's d effect-size of the average deterioration is 0.07, and the Hedges' g is 0.06.

5.4 CORE-10

- Of the 383 clients for whom we measured CORE-10 on at least two occasions, we also have more than one follow-up scores for 61 (15.9%) of them.
- The median number of months between the first and last follow-up session was 8.96.
- The average (and standard deviation) of their improvement scores from the first to last follow-up is -1.23 (5.91), which is an average deterioration.
- A paired samples t-test comparing scores at the first and last follow-up sessions showed that the average change in scores was non-significant: t(60) = -1.63, p = 0.11.
- The Cohen's d effect-size of the average deterioration is 0.13, and the Hedges' g is 0.13.

5.5 PRN-14

- Of the 295 clients for whom we measured PRN-14 on at least two occasions, we also have more than one follow-up scores for 55 (18.6%) of them.
- The median number of months between the first and last follow-up session was 8.96.
- The average (and standard deviation) of their improvement scores from the first to last follow-up is -5.15 (23.89), which is an average deterioration.
- A paired samples t-test comparing scores at the first and last follow-up sessions showed that the average change in scores was non-significant: t(54) = 1.6, p = 0.12.
- The Cohen's d effect-size of the average deterioration is 0.14, and the Hedges' g is 0.14.

5.6 P-value adjustment

The p-values from the analyses just reported were all greater than 0.05 and so none were significant. Nonetheless, as was done previously, these can be adjusted, which will necessarily increase all p-values. Adjusting by the conservative Bonferroni adjustment, the lowest p-value is now a p-value of 0.547 for the CORE-10 t-test. Adjusting by the less conservative Benjamini-Hochberg False Discovery Rate (FDR) method, the lowest p-value is now a p-value of 0.29 for the CORE-10 t-test.

For each outcome, the adjusted p-values are as follows:

Measure	Bonferroni adjusted p-value	FDR adjusted p-value
GAD-7	1	0.398
PHQ-9	1	0.509
PCL-5	1	0.509
CORE-10	0.547	0.29
PRN-14	0.58	0.29

6 Predictors of outcome improvement or drop-out

6.1 Outcome improvement

As described above, the outcome variables GAD-7, PHQ-9, PCL-5, CORE-10, PRN-14 (reversed) are all highly positively intercorrelated. It is convenient therefore to create an average improvement score, which we will call the global improvement score, that is the average of these five variables after they have been normalized (re-scaled to have a mean of zero and a standard deviation of 1). This global score can itself be normalized to aid interpretation (e.g. when normalized, the average of the global improvement scores across all clients is zero, and a change in one unit is a one standard deviation change).

We performed a linear regression model modelling global improvement scores as a function of the therapist, the client's age, their gender, their employment status, their military service history, their living arrangements, whether they self-referred to the therapy programme, and whether they take anti-depressants. These client based predictor variables were those introduced and described in Section 1. Using an AIC based forward/backward stepwise regression to select which of these variables are predictive of the outcome, no variables were selected. In other words, none of the client variables significantly predicted whether there was a change in average improvement score.

6.2 Drop-out

We performed a binary logistic regression model modelling the probability of drop-out of the treatment programme as a function of the same predictor variables used in the previous linear regression: therapist, the client's age, their gender, their employment status, their military service history, their living arrangements, whether they self-referred to the therapy programme, and whether they take anti-depressants. Using an AIC based forward/backward stepwise regression to select which of these variables are predictive of the outcome, three variables were automatically selected: was age, gender, and whether client takes anti-depressant medication. However, of these selected variables, only the age variable was significant[4].

[4]More precisely, in the forward/backward stepwise regression, although the age, gender, and anti-depressant medication use variables were selected, the coefficients for gender and anti-depressant medication use were not significant. Furthermore, the model with the three selected variables had a Akaike Information Criterion (AIC) score

The following table shows the predicted probability of drop-out, and its 95% confidence interval, for a set of representative ages. Specifically, these ages represent the 10th, 25th, 50th, 75th, 90th age percentiles.

Age	Prob of drop-out	95% CI
32	0.29	0.22, 0.38
38	0.25	0.2, 0.32
46	0.21	0.16, 0.26
55	0.16	0.12, 0.22
63	0.13	0.08, 0.19

As can be seen, the older the clients, the lower the probability of drop-out. The youngest clients have a nearly one in four chance of drop-out, while the oldest clients have a one in ten chance of drop-out.

6.3 Inter-therapist differences in outcomes and drop-out

The previous two analysis showed that average global improvement scores and the probability of client drop-out do not vary significantly across therapists. In other words, different therapists do not lead to significantly better or worse outcomes on average, and different therapists do not lead to higher or lower probabilities of client drop-out.

7 Analysis of the *rewind* intervention

In each of the therapy sessions, a number of intervention methods were used. In total, there were 27 possible intervention techniques. In any session, if no intervention were recorded as having been used, we will treat this session as missing data with respect to the intervention. In other words, we will assume that if no intervention were recorded for a given session, this was an omission on the part of the therapist, and that there is therefore no record of which interventions were used in that session. In total, 20.9% of the therapy sessions were recorded as having had no interventions. After removing these missing cases, the median number of interventions used in any given therapy session (not including screening or follow-up sessions) was 7. Figure 2 shows each method and the number and proportion of clients who used that method at least once in their therapy sessions.

Of particular interest is the *rewind* technique. This was used by 235 (54%) clients. We analysed the effectiveness of this method in two distinct ways. First, we analysed if there was a difference in the average improvement of those clients who used the rewind method compared to those who did not use it. Second, we analysed if the client's session-by-session improvement on the outcome variables accelerated after they had used the rewind method.

For the first analysis, separately for the GAD-7, PHQ-9, PCL-5, CORE-10, and PRN-14, we performed an independent samples t-test on the improvement scores of the clients who used rewind compared to those who did not use it. The t-statistics and p-values, including adjusted p-values, for these tests are shown in the next table.

of 469.2, but a model with just the age variable as a predictor had an AIC score of 470. This is a negligible difference in the two AIC values, with differences in model AIC values of less than 3 being interpreted by conventional statistical standards as not indicating any practical difference in model fit. As such, the model with age, gender and anti-depressant usage as predictors have a negligible improvement over a model with age alone as the predictor.

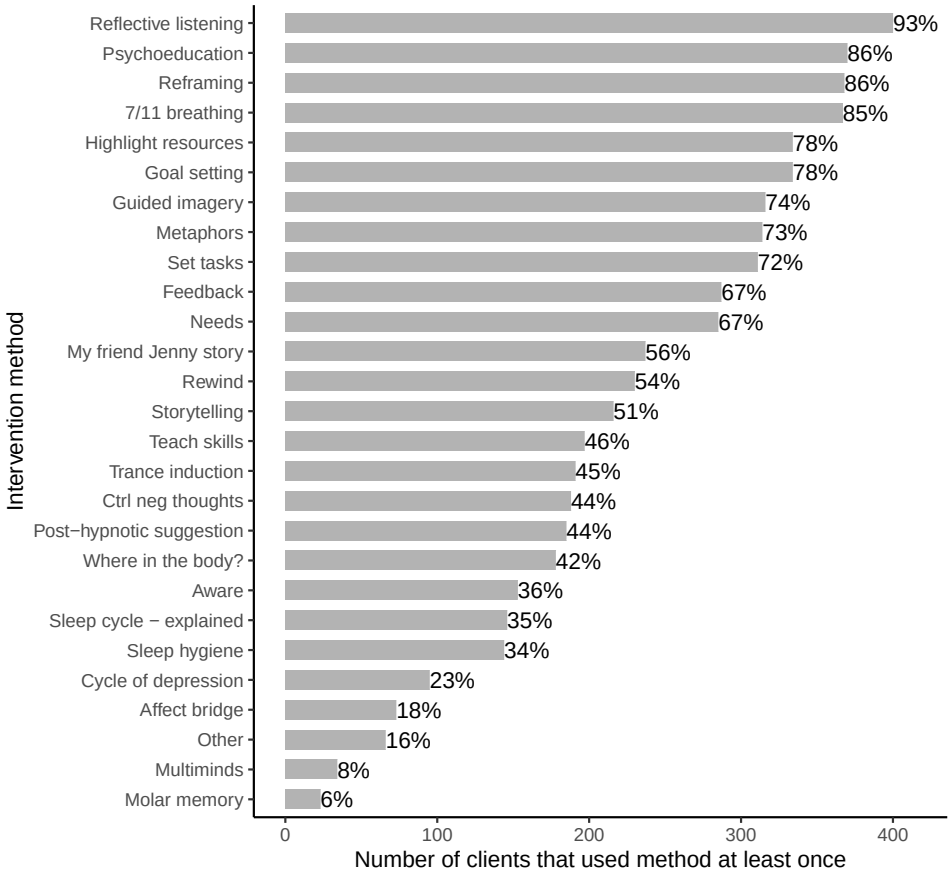

Figure 2: The set of therapeutic intervention methods used and the number of clients who used each method at least once. Percentages indicate the percentage of clients using each method at least once.

Measure	t-test	Bonferroni adjusted p-value	FDR adjusted p-value
GAD-7	t(376) = -2.46, p = 0.014	0.072	0.036
PHQ-9	t(359) = -1.92, p = 0.055	0.277	0.074
PCL-5	t(291) = -1.89, p = 0.059	0.297	0.074
CORE-10	t(380) = -3.34, p <0.001	0.005	0.005
PRN-14	t(293) = -1.29, p = 0.2	0.986	0.197

Note that for these analyses, the numerator of the t-statistic is the average improvement score[5] for those clients who did not use the rewind method minus the average improvement of those who did use it. This means that if the t-statistic is negative, the average improvement of those who did use the rewind technique is higher than those who did not use it. For all measures, we see that the there are higher improvement scores for those that used rewind, and in some cases, these are significant even after there is a p-value adjustment.

For the second analysis, and separately for each outcome variable, for those clients who used rewind, we calculated their average session-by-session score change before and after the technique was first used. We then compared these average changes using a paired samples t-test. The t-statistics and p-values, including adjusted p-values, for these tests are as follows.

[5]Here as before, the improvement score for all measures except PRN-14 is the initial score minus the final score. For PRN-14, it is the final score minus the initial score, or equivalently, the negative of the initial score minus the negative of the final score.

Measure	t-test	Bonferroni adjusted p-value	FDR adjusted p-value
GAD-7	t(134) = 0.17, p = 0.86	1	0.894
PHQ-9	t(128) = 0.66, p = 0.51	1	0.848
PCL-5	t(29) = -0.13, p = 0.89	1	0.894
CORE-10	t(137) = 0.91, p = 0.36	1	0.848
PRN-14	t(28) = 0.79, p = 0.44	1	0.848

Note that for these analyses, the numerator of the t-statistic is the average improvement before the intervention minus the average improvement after the intervention. From these results, for those clients who did use rewind, there is no evidence that their session-by-session improvement accelerates after they experienced rewind for the first time.

8 Analysis of client and therapy feedback

Clients provided feedback using the five item *Agnew Relationship Measure*. Each item is on a 7 point scale, with higher values indicating higher levels of client satisfaction with the therapist and the therapy session. The maximum total score is 35, which indicates maximum satisfaction overall. The median client feedback score across all clients and sessions is 35. This high level of satisfaction occurs from initial session and remains constant across all subsequent sessions.

Therapists indicate their satisfaction with the therapy session using the therapist version of *Agnew Relationship Measure*. This is also a five 7 point item scale, with higher values indicating higher satisfaction, and maximum score of 35 indicating maximum satisfaction. The median therapist feedback score across all clients and sessions is 33, and this near maximum score remains constant over the sessions.

Because of the extreme ceiling effect on both the client and therapist versions of this feedback scale, a meaningful correlation between the client and therapist scores is not possible. However, it is evident that both the client and the therapist are fully satisfied with the progress of therapy from the beginning to the end.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

NHS England. (2019). The improving access to psychological therapies manual: Appendices and helpful resources. *NHS Digital. Https://Www. Rcpsych. Ac. Uk/Doc s/Default-Source/Improving-Care/Nccmh/Iapt/Nccmh-Iapt-Manual-Appendices-Help Ful-Resources-V2. Pdf*.