Brief Report for *Walking with the Wounded*: GAD7 & PHQ9 analysis

Completed by Curtis Sinclair and Graham Durcan

Contents

Summary of findings

The results provide supporting evidence to the effectiveness of the intervention to achieve significant improvement in self-reported measures of anxiety and depression from baseline scores to final session:

- The effect size can be considered as very large for both scores of anxiety (1.52) and depression (1.43).

- Nearly half of all patients starting above the thresholds for caseness moved below the threshold for either anxiety or depression by their final session (48.2% for anxiety and 49.7% for depression).
- Reliable improvement (a reduction in score of four or more for GAD and six or more for the PHQ) was experienced by around 70 per cent of clients for anxiety and 63 per cent for depression.
- 40 per cent of clients achieved reliable recovery (below threshold for caseness in both domains, reliable improvement in at least one domain and without deterioration in the other) and around 37 per cent achieved reliable improvement (reliable improvement in at least one domain and without deterioration in the other).

However, there is evidence that the size of the effect decreases over time. Sample sizes for patients with follow-up scores are currently limited so caution must be taken but results tentatively support the notion that:

- Some level of improvement persists at 6 months, with around 42% below caseness for anxiety and 38% for depression.
- Average impacts remain statistically significant at 6 months although have declined to 1.02 for anxiety and 0.83 for depression.
- At 12 months 43% remain below the caseness threshold for anxiety and 33% for depression.
- Improvements remain statistically significant at 12 months for anxiety with an effect size of 0.73.
- However, improvements at 12 months were not significant for depression. This is a similar finding to that of Dunne and colleagues (2019), in that there was significant change in anxiety and depression scores by the end of therapy but the changes in depression "*…were not sustained at follow up…*" (Page 3), however, participants informed Dunne et al that although symptoms of depression increased, improvements in work and social functioning were maintained.

Collection of additional data and further analysis (e.g., analysis of variance) and/or investigation (e.g., sampling same clients at both time points) is warranted to strengthen the evidence relating to longer term impacts of the intervention.

Analysis

Descriptive Statistics

There were 340 clients who completed pre and post Generalised Anxiety Disorder-7 (GHQ) and Public Health Questionnaire-9 (PHQ) measures. There were 26 clients and 21 clients who completed follow-up GAD and PHQ measures at 6-month (6M) and 12-month (12M) respectively.

The mean scores for GAD were 15.5 (SD=3.8) at the first session and 8.7 (SD=5.0) at the final session. The mean scores for PHQ were 17.8 (SD=4.4) at the first session and 10.2 (SD=6.1) at the final session (see Appendix A for distribution of scores for first and last sessions).

Clinical caseness [1]

At the final session, slightly less than half of cases were below the threshold for CC for either the GAD (48.2%) and PHQ (49.7%) (see Appendix B for further details). At follow-up, the proportion of cases below threshold at 6 months was 42% for GAD and 38% for PHQ. At 12 months the figures were 38% and 33% respectively.

The proportion of all clients at 6M (N=26) and 12M (N=21) that moved from below the threshold for caseness back to being at caseness for GAD was 15.4%% (N=4) at 6M and 9.5% (N=2) at 12M follow-up. In relation to PHQ scores, the proportion of all clients that moved from no caseness back to being at caseness was 23.1% (N=6) at 6M and 19.0% (N=4) at 12M follow-up.

Conversely, the proportion of all clients that went from being at caseness at final session to scoring below the threshold for caseness at follow-up for GAD was 3.8% (N=1) at 6M and 15.4% (N=4) at 12M. In terms of PHQ scores, the proportion of clients that went from being at caseness at final session to scoring below the threshold for caseness was 7.7% (N=2) at 6M and 9.5% (N=2) at 12M at follow-up.

As different clients were asked to complete follow-up measures at each time point these figures cannot be directly compared (see Appendix C for further breakdown of figures).

Table 1: Descriptive statistics for GAD and PHQ scores.

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
1st GAD	340	23	8	31	15.5	3.8	14.3
Final GAD	340	20	1	21	8.7	5.0	25.3
6M GAD	26	26	0	26	10.7	6.8	46.5
12M GAD	21	24	2	26	12.3	7.2	51.2
1st PHQ	340	17	10	27	17.8	4.4	19.0
Final PHQ	340	31	1	32	10.2	6.1	37.3
6M PHQ	26	29	1	30	12.9	8.7	75.2
12M PHQ	21	38	0	38	15.3	9.0	81.1

In terms of CC, 40 per cent of clients no longer met the threshold for caseness in GAD and PHQ scores, almost 18 per cent remained at caseness for at least one domain, and just over 42 per cent remained at caseness for GAD and PHQ (see Appendix D).

[1] Guidance on caseness for both GAD 7 & PHQ 9 is provided in National Collaborating Centre for Mental Health (2019).

Reliable improvement and deterioration[2]

Reliable improvement in GAD and PHQ scores, between the first and final session, were observed for 242 (71.2%) and 215 (63.2%) clients respectively. No reliable change in scores were seen for GAD and PHQ in 92 (27.1%) and 118 (34.7%) clients. Lastly, reliable deterioration in GAD and PHQ scores were observed for 6 (1.8%) and 7 (2.1%) clients (see Appendix E for further details).

Reliable recovery

Reliable recovery is defined as a client showing reliable improvement in one or both scores for GAD and PHQ, whilst both scores must fall below the threshold for CC. Reliable recovery was calculated by using the first and final session scores only. In total, there were 134 (39.4%) clients who achieved reliable recovery, 127 (37.4%) clients who achieved reliable improvement, 69 (20.3%) who experienced no reliable change in scores, and 10 (2.9%) clients who reported a reliable deterioration (see Appendix F for further details).

Analysis – t-tests and effect sizes

A t-test was conducted on the GAD and PHQ scores to determine whether a difference was observed from baseline to final session, 6M, and 12M follow-up.

Comparison of baselines GAD scores to subsequent scores

Overall, results indicated that there were significant improvements in GAD scores from baseline (see Appendix G for further details). Clients achieved a significant improvement in their self-reported level of anxiety from first (*M* = 15.45, *SE* = .21) to last (*M* = 8.70, *SE* = .273) session, *t*(339) = 23.96, *p* < .01, *d* = 1.52; first session (*M* = 16.31, *SE* = .72) and 6M follow-up (*M* = 10.69, *SE* = 1.34), *t*(339) = 4.85, *p* < .01, *d* = 1.02; first session (*M* = 16.33, *SE* = .67) and 12M follow-up (*M* = 12.29, *SE* = .1.56) session, *t*(339) = 2.63, *p* < .05, *d* = .73.

Effect sizes for first and final, and first and 6M scores were extremely large, whilst first and 12M score was medium-to-large effect size.

Comparison of baseline scores to subsequent scores

There were two significant results and one non-significant result for changes in PHQ scores (see Appendix G for further details). Clients achieved a significant improvement in their self-reported level of depression from first (*M* = 17.78, *SE* = .24) to final (*M* = 10.20, *SE* = .33) session, *t*(339) = 21.90, *p* < .01, *d* = 1.43; first session (*M* = 18.50, *SE* = .82) and 6M follow-up (*M* = 12.85, *SE* = 1.70), *t*(339) = 3.81, < .01, *d* = .83.

There was a non-significant improvement in scores between first session (*M* = 18.33, *SE* = 5.02) and 6M follow-up (*M* = 15.33, *SE* = 1.97), *t*(339) = 1.41, *p* = .17 , *d* = .41.

Effect sizes for first and last, and first and 6M scores were extremely large and large respectively. The effect size for first and 12M was small-to-medium size.

Conclusion

The results provide support to the effectiveness of the intervention in significantly improving self-report scores of anxiety and depression compared to baseline, with an effect that can be considered extremely large. Notably, the effect size does decrease over time and is

[2] Guidance on reliable change scores for both GAD 7 & PHQ 9 is provided in National Collaborating Centre for Mental Health (2019).

worthy of further investigation to better understand what factors might help to protect against an increase. Changes in depression scores are not maintained at the 12-month follow-up.

Some caution should be taken in interpreting the follow-up tests (i.e., first and 6M, first and 12M) since the risk of finding a significant difference where one does not exist is increased by conducting several t-tests in this manner (i.e., risk increased by roughly 5% for each t-test). Best practice should be to undertake analysis of variance to mitigate the likelihood of finding a false positive. Furthermore, a small number of clients who were at caseness for GAD or PHQ appear to improve during the period between final session and follow-up whereby they are no longer at caseness. This will have impacted several of the resultant statistical significance and effect sizes and adds further credence to the caution of interpreting the follow-up results. Therefore, it would be highly beneficial to collect follow-up scores, at 6M and 12M time points, from the same individual to further increase the robustness of future analysis.

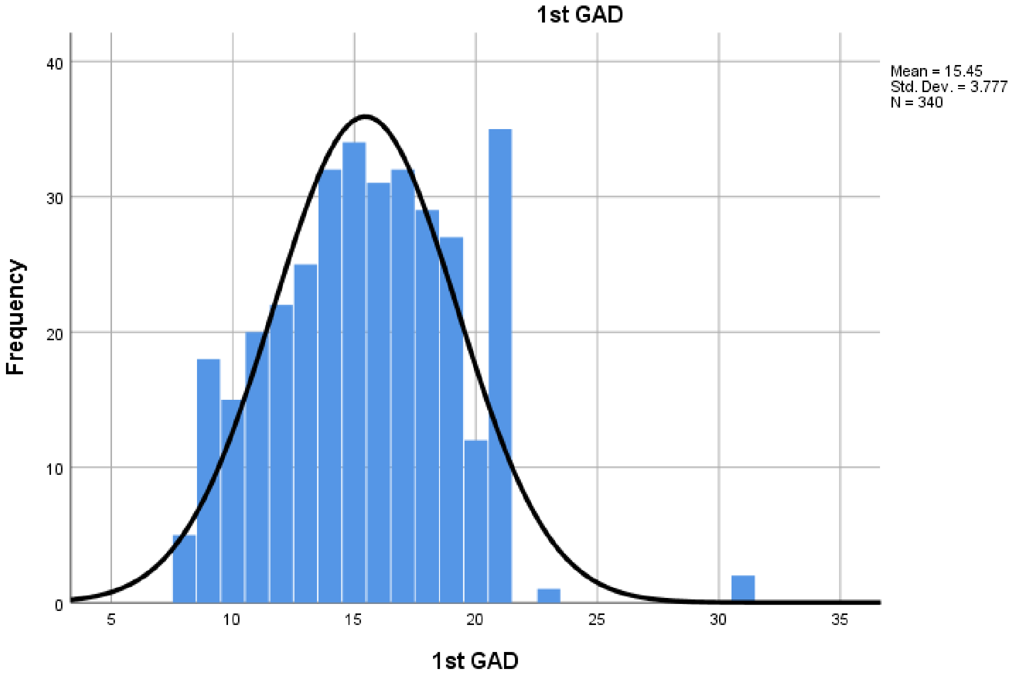
References

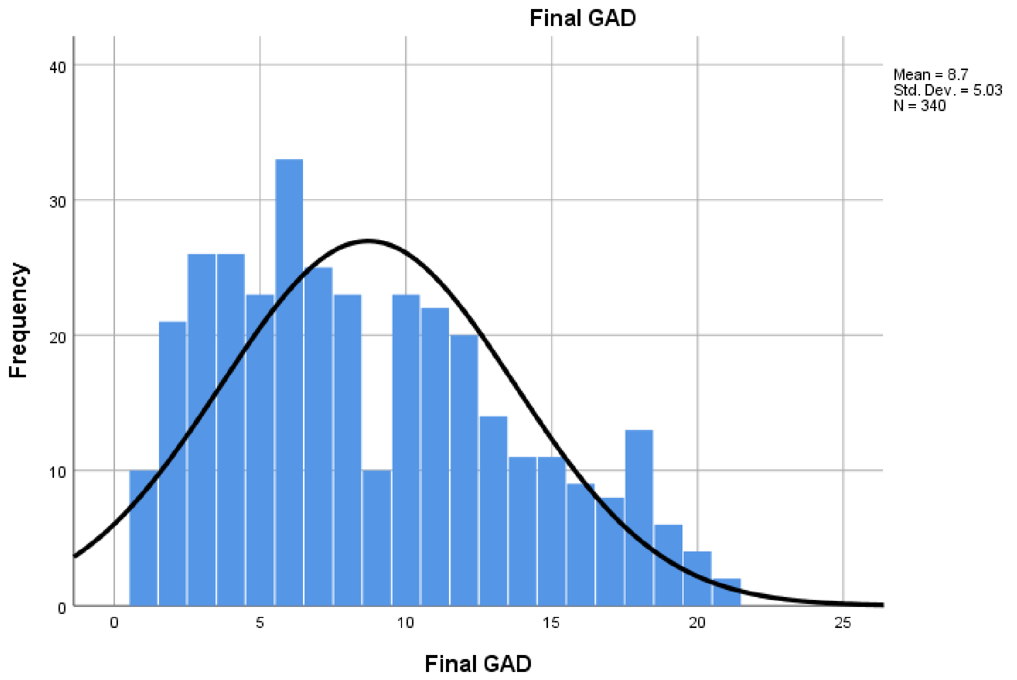
Dunne R, Goodwin L, Brooks S & Greenberg N (2019) *An Evaluation of the Walking With The Wounded Programmes: Final Report*. London. Kings College London.

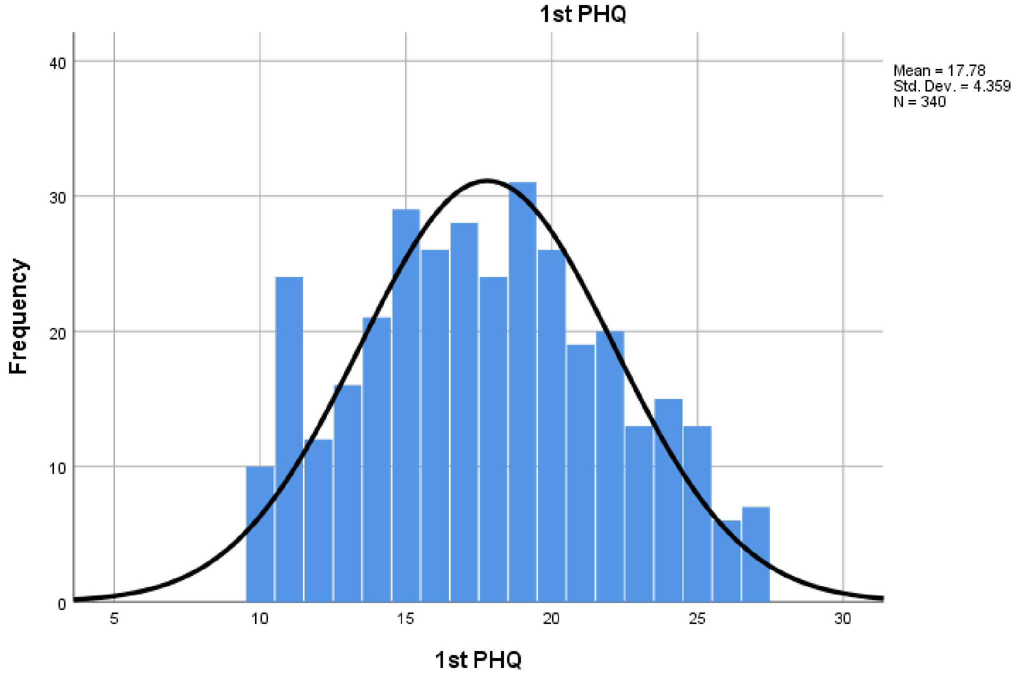
National Collaborating Centre for Mental Health (2018 – updated 2019) *The Improving Access to Psychological Therapies Manual: Appendices and helpful resources.* London. National Collaborating Centre for Mental Health.

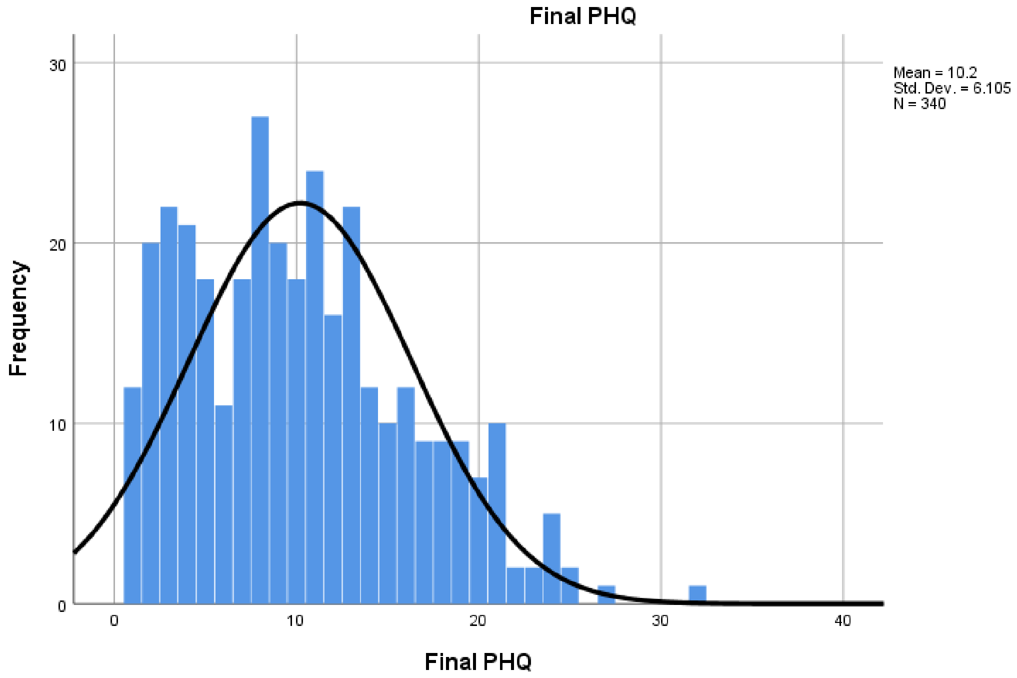
Appendices

Appendix A: Distribution of scores for GAD and PHQ at first and last sessions.




Final GAD
Frequency
0
10
20
30
40
0
5
10
15
20
25
Final GAD
Mean = 8.7
Std. Dev. = 5.03
N = 340


1st PHQ
Frequency
0
10
20
30
40
5
10
15
20
25
30
1st PHQ
Mean = 17.78
Std. Dev. = 4.359
N = 340


Final PHQ
Frequency
30
20
10
0
0
10
20
30
40
Final PHQ
Mean = 10.2
Std. Dev. = 6.105
N = 340

Appendix B: Frequency of GAD and PHQ meeting criteria for clinical caseness.

	At caseness (%)	Below threshold for caseness
GAD Final	176 (51.8)	164 (48.2)
GAD 6M	15 (57.7)	11 (42.3)
GAD 12M	12 (57.1)	9 (42.9)
PHQ Final	171 (50.3)	169 (49.7)
PHQ 6M	16 (61.5)	10 (38.5)
PHQ 12M	14 (66.7)	7 (33.3)

Appendix C: Clinical caseness at final session compared to 6M and 12M follow-up.

Comparison of final session and 6M follow-up GAD caseness

		At caseness (>=8)	Does not meet caseness (<8)	Total
GAD Final Session Caseness	At caseness (>=8)	11	1	12
	Does not meet caseness (<8)	4	10	14
Total		15	11	26

Comparison of final session and 12M follow-up GAD caseness

		At caseness (>=8)	Does not meet caseness (<8)	Total
GAD Final Session Caseness	At caseness (>=8)	10	4	14
	Does not meet caseness (<8)	2	5	7
Total		12	9	21

Comparison of final session and 6M follow-up PHQ caseness

		Meets caseness (>=10)	Does not meet caseness (<10)	Total
PHQ Final Session Caseness	Meets caseness (>=10)	10	2	12
	Does not meet caseness (<10)	6	8	14
Total		16	10	26

Comparison of final session and 12M follow-up PHQ caseness

		Meets caseness (>=10)	Does not meet caseness (<10)	Total
PHQ Final Session Caseness	Meets caseness (>=10)	10	2	12
	Does not meet caseness (<10)	4	5	9
Total		14	7	21

Appendix D: Clinical caseness at final session in GAD and/or PHQ

	Frequency	**Percent**	**Cumulative Percent**
No clinical caseness in GAD and PHQ	136	40.0	40.0
Clinical caseness in either GAD or PHQ	61	17.9	57.9
Clinical caseness in both GAD AND PHQ	143	42.1	100.0

Appendix E: Reliable improvement, deterioration or no change in GAD and PHQ scores.

Improvement in GAD scores between first and final session

	Frequency	Percent	Cumulative Percent
Reliable deterioration (<=-4)	6	1.8	1.8
No reliable change	92	27.1	28.8
Reliable improvement (>=4)	242	71.2	100.0

Improvement in PHQ scores between first and final session

	Frequency	Percent	Cumulative Percent
Reliable deterioration (<=-6)	7	2.1	2.1
No reliable change	118	34.7	36.8
Reliable improvement (>=6)	215	63.2	100.0

Appendix F: Proportion achieving reliable recovery.

Improvement, deterioration, recovery or no change

	Frequency	Percent	Cumulative Percent
Reliable deterioration	10	2.9	2.9
No reliable change	69	20.3	23.2
Reliable improvement	127	37.4	60.6
Reliable Recovery	134	39.4	100.0

Appendix G: SPSS output for paired t-test comparing scores on GAD at first session, final session, and 6- and 12-month follow-up.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	1st GAD	15.45	340	3.777	.205
	Final GAD	8.70	340	5.030	.273
Pair 2	1st GAD	16.31	26	3.653	.716
	6M GAD	10.69	26	6.822	1.338
Pair 3	1st GAD	16.33	21	3.071	.670
	12M GAD	12.29	21	7.156	1.562

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	1st GAD & Final GAD	340	.331	.000
Pair 2	1st GAD & 6M GAD	26	.503	.009
Pair 3	1st GAD & 12M GAD	21	.250	.274

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	1st GAD - Final GAD	6.750	5.194	.282	6.196	7.304	23.963	339	.000
Pair 2	1st GAD - 6M GAD	5.615	5.900	1.157	3.232	7.998	4.853	25	.000
Pair 3	1st GAD - 12M GAD	4.048	7.046	1.538	.840	7.255	2.632	20	.016

Appendix G: SPSS output for paired t-test comparing scores on PHQ at first session, final session, and 6- and 12-month follow-up.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	1st PHQ	17.78	340	4.359	.236
	Final PHQ	10.20	340	6.105	.331
Pair 2	1st PHQ	18.50	26	4.159	.816

	6M PHQ	12.85	26	8.670	1.700
Pair 3	1st PHQ	18.33	21	5.023	1.096
	12M PHQ	15.33	21	9.007	1.966

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	1st PHQ & Final PHQ	340	.291	.000
Pair 2	1st PHQ & 6M PHQ	26	.487	.012
Pair 3	1st PHQ & 12M PHQ	21	.131	.571

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	1st PHQ - Final PHQ	7.585	6.387	.346	6.904	8.267	21.897	339	.000
Pair 2	1st PHQ - 6M PHQ	5.654	7.573	1.485	2.595	8.713	3.807	25	.001
Pair 3	1st PHQ - 12M PHQ	3.000	9.721	2.121	-1.425	7.425	1.414	20	.173