Supplement 2. Statistical Analysis Plan

DO MORE, FEEL BETTER (DMFB)
STATISTICAL ANALYSIS PLAN (SAP)

Principal Investigators – Patrick Raue, PhD, Jo Anne Sirey, PhD, & Amber Gum, PhD
Authors – Brittany Blanchard, PhD, Julien Rouvere, MA, Patrick Raue, PhD

Grant Number: R01MH124975
National Clinical Trial (NCT) Identified Number: NCT04621877
Sponsor: National Institute of Mental Health (NIMH)

Modifications to the DMFB Statistical Analysis Plan

Change	Reason	Date
Scoring WHODAS 2.0 modified to use response scale from 0-4.	To align with NIMH reporting standards.	5/5/2026

Table of Contents

1. Introduction 2
2. Study Design 2
3. Outcomes 2
  3.1. Primary Outcomes 2
  3.2. Secondary Outcomes 2
  3.3. Moderators 3
4. Analyses 3
  4.1. Multiple Primary Outcomes 3
  4.2. Non-inferiority Margin 3
  4.3. Secondary Outcomes 4
  4.4. Mediation 4
  4.5. Exploratory Outcomes 4
5. Missing Data 4
6. Power Calculation 4
7. References 4

1. Introduction

This collaborative R01 (UW, Cornell, and USF) was in response to NIMH RFA-MH-18-700 to test interventions "for which there is already evidence of efficacy, for use in community and practice settings." We proposed to conduct a non-inferiority trial of "Do More, Feel Better" (DMFB), a simplified version of Behavioral Activation (BA) delivered by lay volunteers to depressed senior center clients, compared to traditional BA delivered by clinicians (clinician BA). This trial will test the effectiveness of lay-delivered DMFB compared to clinician-delivered BA among older (60+ years) senior center clients in increasing overall activity level and/or reducing depression symptoms.

2. Study Design

This randomized controlled trial uses a parallel arm design to test whether 9-week DMFB (intervention) is non-inferior in effectiveness to clinician-delivered BA (active control). Depressed clients will be individually randomized to DMFB or clinician BA using a 1:1 allocation ratio and blocked randomization with randomly selected block sizes. Randomization will be stratified by senior center. Clients will receive research assessments at baseline, 3 weeks, 6 weeks, and 9 weeks, as well as two longer-term follow-up periods (24 and 36 weeks). We will test the following hypotheses:

Hypothesis 1) DMFB is non-inferior to clinician BA in increasing overall activity level (proposed mechanism of action; primary outcome) over 9 weeks.

Hypothesis 2) DMFB is non-inferior to clinician BA in reducing depressive symptoms (target client outcome; primary outcome) over 9 weeks.

Hypothesis 3) Change in activity level (Behavioral Activation for Depression Scale; BADS[1,2]) at each assessment time (baseline-3 weeks, 3-6 weeks, 6-9 weeks) predicts severity of depression (Hamilton Rating Scale for Depression; HAM-D[3]) at next assessment time (3, 6, 9 weeks) across conditions.

We will also examine secondary outcomes, including disability (World Health Organization Disability Assessment Schedule; WHODAS 2.0[4]), satisfaction with treatment (Client Satisfaction Questionnaire; CSQ-3[5]), and client-level moderators.

3. Outcomes

3.1. Primary Outcomes. We have multiple primary outcomes: the proposed mechanism of action (activity level) and the client outcome of interest (depression symptom severity). Activity level is measured using the Behavioral Activation for Depression Scale (BADS; H1). The BADS includes 25 items on a 7-point scale ranging from 0 (not at all) to 6 (completely). The BADS has evidence of a 4-factor structure (Activation, Avoidance, School/Work Impairment, Social Impairment), and the total score ranges from 0 to 150. The BADS is scored such that negatively worded items are reverse-scored so that higher scores reflect higher activation. The BADS has evidence of construct validity, adequate test-retest reliability (r = .74), and good internal consistency (alpha = .87).[2]

Depression symptom severity is measured using the 24-item Hamilton Rating Scale for Depression (HAM-D; H2) with a Likert-type scale of options ranging from 0 to 2 or from 0 to 4. Total scores range from 0 to 72, with higher scores indicating more severe depressive symptoms. Item 3 assesses suicide risk. The HAM-D has evidence of good internal consistency (alpha = .79), inter-rater (r = .93), and high test-retest reliability (r = .87).[6]

3.2. Secondary Outcomes. We will measure disability using the 12-item World Health Organization Disability Assessment Schedule (WHODAS 2.0). This measure assesses past 30-day difficulties in communicating, getting around, engaging in self-care, social interactions, household/work activities, and participating in community. Response options range from 0 (none) to 4 (extreme/cannot do). Total scores range from 0 to 48, with higher scores indicating greater disability. The WHODAS 2.0 has evidence of robust internal consistency (alpha = .81 to .96).[7]

To assess satisfaction with treatment, we will use the 3-item Client Satisfaction Questionnaire.[8] The CSQ-3 is an empirically derived set of the CSQScales®[9] and uses items 3, 7, and 8 from the CSQ-8, which has evidence of excellent internal consistency (alpha = .93).[10] The CSQ-3 has been used across a range of services including inpatient and outpatient services.[10] The CSQ-3 uses a 4-point scale ranging from 0 (strongly disagree) to 4 (strongly agree). Total scores range from 0 to 12, with higher scores indicating higher satisfaction.

3.3. Moderators. We will examine multiple potential sociodemographic and clinical moderators of the effect of DMFB vs clinician BA. We will examine age (60-74 vs 75+ years old) and gender (women vs men). For education and race, we will examine groups that represent a minimum of 20% of the sample. At minimum, we will examine these moderators dichotomously (≤ high school and > high school education; People of Color and non-Hispanic White, respectively). Clinical moderators include diagnostic status (depressive disorder versus subthreshold as determined by the Structured Interview for DSM-5 [SCID][11]), previous psychiatric hospitalization (yes/no based on self-report on the SCID-5), and disability (defined as WHODAS 2.0 scores >10). See Table 1.

Table 1. Outcomes and Assessment Schedule

Construct	Instrument	Baseline	3-week Follow-Up	6-week Follow-Up	9-week Follow-Up
Multiple Primary Outcomes					
Activity Level	BADS	X	X	X	X
Depression Symptom Severity	24-item HAM-D	X	X	X	X
Secondary Outcomes					
Disability	12-item WHODAS 2.0	X	X	X	X
Client Satisfaction	CSQ-3				X
Exploratory Outcomes (Moderators)					
Gender (women vs men)	Demographics	X			
Age (60-74 vs 75+)	Demographics	X			
Educational attainment	Demographics	X			
Race (POC vs White)	Demographics	X			
Depressive disorder diagnosis (yes vs no)	SCID-5	X			
Previous psychiatric hospitalization (yes vs no)	SCID-5	X			

Note: BADS, Behavioral Activation in Depression Scale; CSQ-3, 3-item Client Satisfaction Questionnaire; HAM-D, Hamilton Depression Rating Scale; POC, People of Color; SCID-5, Structured Interview for DSM-5; WHODAS 2.0, World Health Organization Disability Assessment Schedule version 2.0.

4. Analyses

4.1. Multiple Primary Outcomes. We will use generalized linear mixed models to account for clustering of observations (level 1) within patients (level 2) within coaches (level 3). Both outcomes (activity level – H1; depression symptom severity – H2) will be modeled as continuous. Time will be modeled as continuous, and we will include a time x condition interaction. Covariates will include design variables (location: FL, NY, WA), statistically significant differences in baseline client characteristics across conditions, gender, self-reported antidepressant use at baseline (yes/no), psychotherapy receipt (yes/no), and previous psychiatric hospitalization (yes/no). To test the non-inferiority Hypotheses 1 (activity level) and 2 (depression symptom severity), we will compute the difference between model-estimated means divided by the pooled baseline standard deviations. We will report adjusted means, standard deviations, and Cohen *d* with 95% confidence intervals.

4.2. Non-inferiority Margin. We will interpret DMFB as non-inferior to clinician BA if the difference between conditions is below a non-inferiority margin that separates a clinically meaningful from a clinically negligible difference. To select this margin, we will use the Reliable Clinical Index (RCI) criterion, a metric of minimal clinical important difference which accounts for measurement error. We will estimate the RCI criterion for the HAM-D and BADS for this sample using the baseline pooled standard deviation. We will use the following test-retest estimates from the literature for the HAM-D (r = .87)[6] and the BADS (r = .74).[2] We will use the following formula:

$$RCI(.05) = SD_{baseline} * 1.96 * \sqrt{(1 - r_{xx})} * \sqrt{2}$$

$SD_{baseline}$ = pooled baseline SD from each condition; r_{xx} = test-retest reliability

Using SD estimates from the ENGAGE study (older adults with depression; N = 249; HAM-D: SD = 4.0; BADS: SD = 25.3) and the pilot DMFB trial (N = 56; HAM-D: SD = 3.2; BADS: SD = 21.5), we estimate the HAM-D RCI will range from 3.2-4.0, and the BADS RCI will range from 30.4-35.8. This equates to a Cohen *d* of .99 and 1.41, respectively. Cohen *d* effect sizes will be calculated and presented with 95% confidence intervals. Between-group effect sizes will be calculated as the difference in model-estimated means at 9 weeks divided by the pooled, raw baseline standard deviation.

4.3. Secondary Outcomes. For the secondary outcome of disability, we will use the same model, with disability modeled as continuous, and calculate non-inferiority margin as described for the primary outcomes. We used a test-retest estimate (r = .79) based on estimates derived from older adult samples from the existing literature.[12,13] For client satisfaction, we will use a simplified version of the same model, with CSQ-3 scores at 9 weeks as the outcome. Consistent with New England Journal of Medicine recommendations, we are limiting use of *P* values. For secondary outcomes, we will report descriptives, including means, standard deviations, effect sizes, and 95% confidence intervals.

4.4. Mediation. We will conduct a mediation analysis to determine whether increases in activity level (proposed mechanism of action) predict subsequent decreases in depression symptom severity. If our non-inferiority hypothesis is confirmed, we will not proceed with a mediation analysis. If our hypotheses are not confirmed, we will examine whether change in BADS scores predicts HAM-D scores, adjusting for design effects. We will modify our primary outcome models and test for mediation using the product of coefficients approach with bias-corrected bootstrapping to estimate asymmetric confidence intervals for the indirect effect.

4.5. Exploratory Outcomes. Exploratory analyses include potential moderators of the treatment effect (i.e., client baseline demographics and clinical profiles). We will build on the primary outcome models and include prespecified moderators in separate models (gender, age, education, race, previous psychiatric hospitalization, disability status) as time-invariant covariates using responses at baseline. Potential treatment moderators will be modeled at level 2 (client level). The models will include a condition x moderator interaction, and the primary outcome of interest will be the condition x moderator x time interaction. We will use the Holm-Bonferroni method to adjust for multiple comparisons. We will report descriptives, including means, standard deviations, effect sizes, and 95% confidence intervals.

5. Missing Data

To handle missing data, we will use maximum likelihood. As a sensitivity analysis, we will compare estimates to those obtained from complete case analysis.

6. Power Calculation

With 2 clinicians and 2 coaches per clinic and 6 clinics per site (n = 3), clients will be nested in 36 clinicians/coaches per arm. We are assuming an intraclass correlation coefficient (ICC) of .05 to account for coach- and site-level effects. Each clinician/coach is expected to deliver the intervention to 4 clients (n = 288), and we estimate an attrition rate of 20%. Assuming only 3 clients from each clinician/coach completes the 9-week assessment, and an individual autocorrelation of r = .70, the minimal detectable effect size at 80% power is Cohen *d* = 0.29 for a one-sided test with an alpha level set to .025 to account for the two primary outcomes. We are therefore adequately powered to test for non-inferiority at the FDA margin recommended for this design (i.e., 1/2 of the effect size of the comparator intervention against placebo/inert control; SMD = .70 -.72 / 2 = .35 -.36). However, we will estimate reliable change index (RCI)[14] for our sample to inform our non-inferiority margin.[15]

7. References

1. Kanter JW, Rusch LC, Busch AM, Sedivy SK. Validation of the Behavioral Activation for Depression Scale (BADS) in a community sample with elevated depressive symptoms. *J Psychopathol Behav Assess*. 2009;31(1):36-42. doi:10.1007/s10862-008-9088-y
2. Kanter JW, Mulick PS, Busch AM, Berlin KS, Martell CR. The Behavioral Activation for Depression Scale (BADS): psychometric properties and factor structure. *J Psychopathol Behav Assess*. 2007;29(3):191-202. doi:10.1007/s10862-006-9038-5

3. Hamilton M. A Rating Scale for Depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56-62. doi:10.1136/jnnp.23.1.56
4. Üstün TB, Chatterji S, Kostanjsek N, et al. Developing the World Health Organization Disability Assessment Schedule 2.0. *Bull World Health Organ*. 2010;88(11):815-823. doi:10.2471/BLT.09.067231
5. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD. Assessment of client/patient satisfaction: development of a general scale. *Eval Program Plann*. 1979;2(3):197-207. doi:10.1016/0149-7189(79)90094-6
6. Trajković G, Starčević V, Latas M, et al. Reliability of the Hamilton Rating Scale for Depression: a meta-analysis over a period of 49 years. *Psychiatry Res*. 2011;189(1):1-9. doi:10.1016/j.psychres.2010.12.007
7. Saltychev M, Katajapuu N, Bärlund E, Laimi K. Psychometric properties of 12-item self-administered World Health Organization disability assessment schedule 2.0 (WHODAS 2.0) among general population and people with non-acute physical causes of disability - systematic review. *Disabil Rehabil*. 2021;43(6):789-794. doi:10.1080/09638288.2019.1643416
8. Nguyen TD, Attkisson CC, Stegner BL. Assessment of patient satisfaction: development and refinement of a service evaluation questionnaire. *Eval Program Plann*. 1983;6(3):299-313. doi:10.1016/0149-7189(83)90010-1
9. CSQ Scales. https://www.csqscales.com. Accessed August 6, 2026.
10. Miglietta E, Belessiotis-Richards C, Ruggeri M, Priebe S. Scales for assessing patient satisfaction with mental health care: a systematic review. *J Psychiatr Res*. 2018;100:33-46. doi:10.1016/j.jpsychires.2018.02.014
11. First M, Williams J, Karg R, Spitzer R. Structured Clinical Interview for DSM-5, Research Version (SCID-5-RV). Published online 2015. https://www.appi.org/50044
12. Serrano-Dueñas M, Serrano M, Mafla D, Martínez-Martín P. Evaluation of the metric properties of the WHODAS 2.0, WHODAS-S, and RADS in the assessment of disability in Parkinsonian patients. *Clin Neurol Neurosurg*. 2020;194:105872. doi:10.1016/j.clineuro.2020.105872
13. Moreira A, Alvarelhão J, Silva AG, Costa R, Queirós A. Validation of a Portuguese version of WHODAS 2.0 - 12 items in people aged 55 or more. *Rev Port Saúde Pública*. 2015;33:179–182. doi:10.1016/j.rpsp.2015.06.003
14. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991;59(1):12-19. doi:10.1037/0022-006X.59.1.12
15. Saver JL, Mistry E. Time to recognize three classes of non-inferiority trial margins. *J Neurointerv Surg*. 2023;15(1):2-4. doi:10.1136/neurintsurg-2022-018879