Statistical Analysis Plan for "Imagery-enhanced versus verbally-based group cognitive behavioural therapy for social anxiety disorder: A randomised controlled trial"

Australia New Zealand Clinical Trials Registration Number: ACTRN12616000579493

Statistical Analysis Protocol Version Number: 1.01

Date: 18 October 2019

Prepared by:

David Erceg-Hurn (Senior Research Scientist, Centre for Clinical Interventions)

Andrew Johnson (Research Associate, Curtin University)

Peter McEvoy (Chief Investigator; Professor, Curtin University)

Funding: National Health and Medical Research Council Project Grant (APP1104007) and the Research Office at Curtin University.

Contents

Version History

Date	Version	Reason
04/10/2019	1.00	Initial upload
18/10/2019	1.01	Added assessment and handling of invalid cases (p. 10)

Background and Rationale

The objective of this trial is to evaluate whether imagery-enhanced cognitive behavioural group therapy (IE-CBGT) is superior to standard verbally-based cognitive behavioural group therapy (VB-CBGT) for the treatment of social anxiety disorder (SAD). The study protocol is described in McEvoy et al. (2017).

To improve the reproducibility, transparency, and validity of clinical trial analyses, it has been recommended that clinical trialists create a statistical analysis plan (SAP; Gamble et al., 2017; Yuan et al., 2019). The SAP is intended to complement the trial protocol. According to Gamble et al. (2017), a SAP should contain a more technical and detailed elaboration of the main features of the trial analyses. We have followed this recommendation in preparing this SAP, and further adhered to guidelines about the content of SAPs described in Gamble et al. (2017) and Yuan et al. (2019).

The current plan has been prepared during August and October 2019, prior to the data being made available. The final follow-up data for the last treatment cohort will be collected in November 2019. All trial data was blinded and not collated at the time of writing this analytic plan. Treatment condition will remain masked to the data analyst until after the analyses described in this SAP have been conducted. The present analysis plan will be time-stamped and made publicly-available.

Some of the proposed analyses deviate from those described in the protocol paper (McEvoy et al., 2017). The main reason is to better comply with best-practices in the analysis of clinical trial data, some of which have changed since the trial was registered and protocol published (Mallinckrodt & Lipkovich, 2017). We will also be taking advantage of improved analytic techniques and software that have been developed during the study's data collection period. Recent reanalysis of data collected in pilot studies (prior to the RCT) has also informed the feasibility of different planned analytic approaches (Erceg-Hurn & McEvoy, 2018; McEvoy, Erceg-Hurn, Saulsman, & Thibodeau, 2015; McEvoy & Saulsman, 2014). The analyses in this SAP supersede those described in McEvoy et al. (2017).

Trial Design, Methods, and Selection of Outcomes

A comprehensive overview of the trial design, eligibility criteria, randomisation schedule, sample size, outcomes, and measurement schedule can be found in McEvoy et al. (2017). Therefore, that information is not repeated in depth here.

In short, the study is a superiority trial that aimed to randomise at least 96 individuals with social anxiety-disorder to one of two 12-week group treatments for SAD (IE-GCBT and VB-GCBT). There is also a one-month follow up group review session. Over 20 outcome measures have been collected at various times before, during, and after treatment. The primary endpoint is the one month follow up (i.e., coinciding with the follow up group review session). This was selected on the basis of pilot data which showed differences between the treatments are larger at one-month follow up than at the end of weekly treatment sessions (McEvoy et al., 2015). No interim analyses or stopping rules were planned, nor have any been used.

The primary hypotheses are that, at the one-month follow up, IE-GCBT will be superior to VB-GCBT in terms of (a) reduction in severity of self-reported social anxiety symptoms (b) percentage of individuals with a SAD diagnosis (c) reduction in clinician-rated severity of social anxiety symptoms. McEvoy et al. (2017) also describe a wide range of additional hypotheses concerning secondary outcome measures, mechanisms of change, and moderation. Given the numerous outcomes and possible research questions, it is not possible to comprehensively report them all in a single paper. Therefore, they will be addressed across several publications.

In the first paper, we will report measures related to the primary hypotheses and key secondary hypotheses (see Table 1). The first three outcomes (self-reported SAD severity, diagnostic status, and clinician-rated severity) were selected in order to test the primary hypotheses. The remaining measures were selected because we judged them to be the most important secondary outcomes. Fear of negative evaluation is a mechanism proposed to underly SAD, thus it is relevant to evaluate whether IE-GCBT is superior to VB-GCBT in changing this mechanism. Similarly, in the treatment clinic where the trial is being conducted, SAD is often comorbid with major depression and generalised anxiety disorder. Therefore, it is meaningful to evaluate whether the treatments differ in their impact on the severity of symptoms of these comorbid conditions. Additional research questions and outcomes (e.g., moderators and predictors of change, health economic outcomes, and psychophysiological parameters) will be reported in subsequent papers. The remainder of this SAP describes how the variables that will be reported in the main outcomes paper will be analysed.

Table 1

Outcomes to be Reporting in the First Paper

Construct	**Measure**
Primary Outcomes	
Severity of social interaction anxiety (self-report)	Social Interaction Anxiety Scale (SIAS)
Presence or absence of SAD diagnosis (clinician-rated)	SCID-5 SAD diagnostic status
Severity of SAD diagnosis (clinician-rated)	Clinician-rated severity scale
Secondary outcomes	
Severity of social performance anxiety	Social Phobia Scale (SPS)
Severity of fear of negative evaluation	Brief Fear of Negative Evaluation - Straightforwardly Worded (BFNE-S)
Severity of fear of positive evaluation	Fear of Positive Evaluation Scale (FPE)
Severity of general anxiety	PROMIS Anxiety - Form 8A
Severity of depression	PROMIS Depression - Form 8A
Reliable & clinically significant change in social interaction anxiety	SIAS (categorised in accord with Jacobson & Truax, 1991)

Planned Reporting of Patient Characteristics and Measures

CONSORT Flow Chart

We will include a CONSORT flow chart showing the number of participants:

- Screened for the study
- Assessed
- Randomised
- Starting treatment
- Providing follow-up data

The chart will also show the reasons (where known) why participants did not meet trial eligibility criteria, discontinued treatment, and so on.

Number of Sessions Completed

As part of the CONSORT flow chart, we will report the number of randomised patients who commenced treatment. We will also report for each treatment arm: the mean number of treatment sessions completed and standard deviation. We will estimate the mean difference in the number of sessions completed, a 95% confidence interval for the difference, and a p-value computed using a Welch (unequal-variance) t-test. To ensure comparability with prior research (McEvoy et al., 2015), the number of sessions completed will be a number out of 13 (comprising the 12 weekly treatment sessions, and the one-month group follow-up session) and the denominator used for calculating the means will be the number of patients who started each treatment (as opposed to the number randomised).

Reporting Baseline Patient Characteristics

For each treatment arm (IE-GCBT and VB-GCBT) we will report the following characteristics at baseline:

Demographic characteristics

- Age
- Gender
- Highest level of education completed
- Employment status
- Relationship status
- % born in Australia

- Cultural background

Diagnostic Information

- Age at onset of SAD diagnosis
- Duration of current social anxiety episode
- Clinician-rated severity of SAD diagnosis
- % of cases diagnosed with performance-only SAD subtype
- Number of comorbid diagnoses
- Prevalence of the most common comorbid diagnoses

Baseline Scores on Self-Report Outcome Variables

- SIAS, SPS, BFNE, FPE, PROMIS Anxiety, PROMIS Depression

Other Clinical Features

- % who have previously received psychiatric treatment
- % hospitalised for psychiatric problems
- Concurrent use of psychotropic medications (defined as antidepressants, anxiolytics, antipsychotics, and mood stabilisers)

For continuous variables with symmetric distributions we will report means and standard deviations. For skewed continuous variables we will report the 50^{th} percentile (median) as an estimate of central tendency and the 25^{th} and 75^{th} percentiles to indicate variability. For categorical variables we will report the percentage of clients falling into each category.

We will not include significance testing for differences between treatment arms in each baseline characteristic. Tests for significance differences at baseline when patients have been randomly allocated to groups is not substantively meaningful, and statisticians have strongly recommended against their use. For more details see Harvey (2018) and Senn (1995).

Data Validity Checks

Prior to the exploration and analysis of the trial data, a series of checks will be performed to assess whether there are any systematic issues with the entry of data and subsequent calculation of composite scores for each scale. Firstly, a subset of the data entry for the SIAS at pre-treatment and at the 1-month follow-up will be independently audited to ensure that the ratings have been accurately entered. Secondly, the composite scores for each scale will be calculated independently by two members of the research team and assessed for

consistency. By taking these measures, we can ensure that neither the data entry or data processing introduces bias into the analysis.

Two clients will be excluded from analyses due to intentionally providing invalid responses to questionnaires. The implausible responding was identified by the treating clinicians, who observed that the clients repeatedly provided invalid responses throughout the course of therapy. Examples include repeatedly rating outside the maximum of a scale, continually using the same response option to rate every item in a scale; and wildly inconsistent responses from week to week to measures of the same construct. The likely implausible responses were reviewed by a member of the research team (DEH) blind to treatment condition, who confirmed that they were consistent with invalid responding. The scores will therefore be removed to prevent biasing both the imputation process and analysis.

Reliability of Outcome Measures

Diagnostic Reliability

In order to assess the reliability of diagnoses, the SCID was re-evaluated by a second diagnostician in a random sample of approximately 20% of assessments across the baseline, 1-month, and 6-month follow-ups. We will report the % of agreement and two indices of inter-rater reliability: Cohen's Kappa and Gwet's AC1. Two indices of inter-rater reliability will be used as Cohen's Kappa can falsely return low values at very high levels of agreement (Gwet, 2008).

Composite Reliability of Self-Report Outcome Measures

We will report the reliability of multi-item self-report outcomes (SIAS, SPS, BFNE-S, FPE, PROMIS Depression and Anxiety) at the first assessment session. The index of reliability will be McDonald's omega, which will be computed using the MBESS package in R (Kelley, 2019). Omega is a generalisation of Cronbach's alpha that is a more theoretically appropriate index of reliability to use in the present study (Dunn, Baguley, & Brunsden, 2014). This is because Cronbach's alpha assumes all items in a scale have identical factor loadings while omega allows them to vary. In more technical terms, omega uses a congeneric measurement model whereas alpha assumes essential tau-equivalence.

Omega reliability coefficients are appropriate for congeneric measures that were developed using classical test theory (CTT) and that are scored by summing or averaging the items. The PROMIS anxiety and depression scales were developed using item-response

theory (IRT) and use an IRT scoring scheme. We will nevertheless report omega reliabilities for these measures as an indicator of reliability under CTT assumptions. We will supplement the PROMIS reliability coefficients with additional information about the typical standard error of PROMIS scores taken from the PROMIS scoring manuals, which are available from http://www.healthmeasures.net/promis-scoring-manuals

Planned Exploration of Outcome Data Prior to Longitudinal Modelling

Texts on longitudinal data analysis for clinical trials (e.g., Diggle, Heagerty, Liang, & Zeger, 2002) strongly recommend that data be explored using numeric and graphical summaries prior to conducting any longitudinal modelling (e.g., linear mixed-models). This helps to understand trends in the data, patterns of missing data, and identify potential problems or unanticipated effects.

For each outcome measure, we will explore:

- Marginal distributions (e.g., boxplots and histograms of the scores at each measurement occasion)
- Marginal treatment trajectories (e.g., plot of the mean trajectories, computed by pooling data across patients at each measurement occasion)
- Subject-specific trends (e.g., spaghetti plots)
- Correlational structure (e.g., correlation matrix of scores at measurement occasion, with missing data handled using pairwise and/or listwise deletion)
- Variance structure (e.g., estimate variance and SDs at each time point)
- Missing data patterns (e.g., marginal trajectories stratified by dropout status)

Trial Estimand and Missing Data

Estimand

In clinical trials, it is common that not all patients fully adhere to the prespecified treatment protocol (e.g., some patients may miss some treatment sessions). As a result, several different types of treatment effect can theoretically be estimated. These include the expected effect if all randomised participants had adhered or the treatment effect amongst only completers (among others, Mallinckrodt, Lin, Lipkovich, & Molenberghs, 2012). The *estimand* of an analysis refers to the type of conceptual effect that is being estimated (Mallinckrodt & Lipkovich, 2017). The present analyses will target the ‘de jure’ estimand,

which is the estimated treatment effect if all randomised individuals adhered to the treatment protocol. Mallinckrodt and Lipkovich (2017) recommend that analyses of the de jure estimand include (a) data from all randomised subjects, consistent with the intent-to-treat principle, and (b) all outcome data collected up to the point that a patient became non-adherent. In the present analyses, non-adherence will be defined as missing more than two consecutive treatment sessions. This means that analyses will utilise all data collected from all randomised subjects up until the point they became non-adherent (in contrast to a completer analysis, which uses data from only a subset of the randomised participants). Other estimands (e.g., de facto estimands) will be reported in subsequent papers. For more about estimands and their implications for data analysis, see Mallinckrodt et al. (2012).

Missing Data

Traditionally, there are three primary approaches to accounting for missing data in clinical trial analyses: listwise deletion, multiple imputation, and full-information maximum likelihood (also referred to as 'direct maximum-likelihood'; National Research Council Panel on Handling Missing Data in Clinical Trials, 2011). Listwise deletion of cases with missing responses can strongly reduce power and risk introducing bias (National Research Council Panel on Handling Missing Data in Clinical Trials, 2011). Multiple imputation (MI) and full information maximum-likelihood (FIML) will give asymptotically equivalent results with missingness (i.e. results will be the same as the sample size tends to infinity; McNeish, 2017). However, FIML can return biased estimates with small sample sizes that depart from normality (McNeish, 2017). Given the sample size available in the present study, multiple imputation will be used. Multiple imputation may also be advantageous in the present study because there are many measures that could be used as auxiliary variables (i.e., additional predictors of the missing data). This may help to reduce bias and increase the efficiency of the analysis, compared to using FIML.

Multiple Imputation

The key steps in generating a multiply imputed dataset that can be used for analysis are: (a) exploring the missing data (b) determining what variables to select as predictors in the imputation model (c) imputing the data, and (d) checking the imputations (White, Royston, & Wood, 2011). We will employ current best-practices in implementing these steps, as discussed in van Buuren (2018), White et al. (2011), and Templ and Filzmoser (2008).

Imputation Model

In order to fill in the missing values (step c above), the analyst must specify an imputation model comprising variables in the dataset that are used to help generate the imputed values. At a minimum, the imputation model will include all variables in the analysis model (e.g., treatment condition, scores on the same variable at prior time points, and interactions between them). The imputation model may also include predictors of dropout, and auxiliary variables that help improve the precision of the imputations, which is consistent with best practice in the imputation literature (White et al., 2011). The exact form of the imputation model will depend on which variables in the dataset (if any) are predictive the presence of missing data, and the strength of associations between the outcome(s) being imputed and the other variables in the dataset.

Imputation Software and Settings

We will use the multiple imputation by chained equations approach to multiple imputation, as implemented in the R package mice (van Buuren, 2018). The default mice imputation algorithm, predictive mean matching (PMM), will be used to multiply impute the missing data. PMM was selected because of its performance in simulation studies with both normal and skewed data (Marshall, Altman, & Holder, 2010; Marshall, Altman, Royston, & Holder, 2010). If problems are encountered with convergence, or the plausibility of the imputations, we will attempt to address these by either (i) increasing the number of iterations, (ii) adjusting the imputation model, or (iii) using an alternative imputation method.

We will use 100 imputations, as using a relatively large number of imputations (e.g., 100 rather than 5) reduces between-imputation variance (van Buuren, 2018). In practice, this means parameter estimates are more precise.

Pooling Results

The primary end point in the study is the one-month group follow up. We will therefore construct a contrast where we generate an estimate of the difference between groups (e.g., difference in means) and standard error at the one-month follow-up for each outcome under study. For each outcome, the 100 estimates and 100 standard errors will then be combined using the standard multiple imputation pooling procedure, known as Rubin’s Rules (van Buuren, 2018). These pooled estimates and standard errors will then be used to assess the magnitude (and significance) of the difference in outcomes between treatments.

Confirmatory Longitudinal Analyses

We will now describe how each imputed dataset will be analysed. The choice of analytical approach was determined by both the type of measurement of the outcome (e.g., binary or continuous) and the number of repeated assessments (i.e., number times that these outcomes were measured). These attributes for the outcomes are summarised in Table 1. See the trial protocol paper for more information about outcomes in the trial and their measurement properties (McEvoy et al., 2017).

Table 1

Measurement Properties and Repeat Assessments for Outcomes under Analysis

Outcome	Type of Measurement	Number of Measurements	
		Baseline	Post-Baseline
Primary			
SIAS	Continuous	2	5
SCID-5 Diagnostic Status	Binary	1	2
Clinician Rated Severity	Ordinal (Ranges 0 – 8)	1	2
Secondary			
SPS	Continuous	1	5
BFNE-S	Continuous	1	13
FPE	Continuous	1	13
PROMIS Anxiety	Continuous	2	13
PROMIS Depression	Continuous	2	13

All outcomes aside from SCID-5 diagnostic status, clinician-rated severity, and reliable and clinically-significant change will be analysed using Linear Mixed Models (LMM), as implemented in the R package nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2019). There are some aspects of these analyses that will be the same for all outcomes and some that will differ depending on the number of measurement times for each outcome. The analytical decisions that are the same for all outcomes will be described first, followed by separate sections detailing the decisions individual to each analysis. A summary of the analyses and parameterizations of each is presented at the end.

Accounting for Nesting Within Group

As treatments are delivered in a group setting, a random intercept for 'group' will be included in all models. The random intercept allows for the responses of individuals within the same group to be correlated with each other. However, in our analyses of pilot data we have found that the variance of the random intercept is often close to zero (indicating little difference between groups). This may be because the treatment protocol is standardised to minimise differences between groups, and outcome largely depends on whether participants complete homework tasks between sessions, which do not overlap with other participants. Some outcomes in the present trial will also be measured after patients have been apart for a long time (i.e., the 1 and 6-month follow ups). The random intercept variance has been so small in analyses of some pilot data that it has caused problems with model convergence. Therefore, if in analyses of trial data, the model with a random intercept for group demonstrates issues with convergence or the random effect has a variance near zero, the random intercept will be removed.

Modelling of Baseline Measurements

Scores collected at baseline could be modelled as a covariate, or as part of the response vector (Mallinckrodt & Lipkovich, 2017). Clinical trial statisticians (e.g., Harrell, 2015; Senn, 2008) often recommend that the baselines be included as a covariate, as this can increase power. Baseline scores may also have a different distribution to responses at other times, which can make joint modelling of the baseline and post-baseline outcomes problematic (Mallinckrodt & Lipkovich, 2017). In the present RCT, baseline score and its interaction with time will be included as a covariate in the fixed effects part of the LMM.

For some outcomes (SIAS, PROMIS-Anxiety, PROMIS-Depression) there are two baseline measurements. These were taken at the initial assessment session, and again immediately prior to the first treatment session. The amount of time between the two baseline measurements is usually short. In analyses of our pilot data we have found the two baseline measurements are strongly correlated, and the mean change between the two assessments is near zero. In these situations, Senn (2008) recommends that the mean of the two baseline measurements be used as a covariate. This is because the mean baseline value is subject to less measurement error than either the individual baselines, the mean baseline is likely to be more strongly correlated with the post-baseline measurements, and it takes fewer parameters to model the mean baseline compared to including both baselines in the model. For the

measures with two baselines, we will follow Senn (2008)'s recommendation to use the mean of the multiple baselines as a fixed covariate.

Continuous Outcomes with 5 Post-Baseline Measurements (SIAS & SPS)

The SIAS and SPS have five post-baseline measurements (weeks 4, 8, and 12, and at the 1-month and 6-month follow-ups). Modelling time as a continuous linear covariate would not be appropriate for the present trial as this assumes the amount of change between each week is the same (Mallinckrodt & Lipkovich, 2017). This is implausible as our pilot data indicate that during weekly treatment there is a rapid improvement in symptoms whereas change over the one-month follow up period is more modest. We also expect that the mean change per week between the one- and six-month follow-ups will be different to the mean change per week during weekly treatment, or over the one-month follow up period. Therefore, time will be modelled as a categorical variable. Entering time into the model as a categorical covariate allows the slope of the mean trajectory between each measurement occasion to vary.

To account for the correlations between the repeated observations of the same individual an unstructured covariance matrix for the residuals will be used. Using an unstructured matrix allows for both the variance in residuals at each assessment time, and the covariance in residuals between times, to be different. This approach is recommended as it makes minimal assumptions about the structure of the covariance matrix, which helps prevent Type I errors (Lu & Mehrotra, 2010; Mallinckrodt & Lipkovich, 2017).

Continuous Outcomes with 13 Post-Baseline Measurements (BFNE, FPE, PROMIS Anxiety & Depression)

Treating time as a categorical variable and using an unstructured covariance matrix is a highly flexible approach that does not make any assumptions about the 'form' of change or the structure of the residuals throughout the trial. However, this approach is not feasible when the sample size is modest and there are many repeated assessments, as the sample size cannot support the large number of parameters that need to be estimated.

An alternative to modelling time as a categorical variable is to model time as a piecewise linear spline. Where categorical time estimates the amount of change between the first post-baseline observation and each subsequent measurement (and how this differs between groups), piecewise time instead separates the trial into distinct 'blocks' of

assessments and estimates the change within those blocks. The piecewise approach has the benefit of not assuming that the amount of change is the same between every assessment in the trial while also requiring the estimation of fewer parameters than categorical time. In other words, it represents a compromise that sits between the two extremes of modelling time: very rigidly (with a single, continuous linear variable) or very flexibly (categorically).

Whereas categorical time makes no assumptions about the form of change throughout the study, piecewise time requires the specification of the form of change within each of the assessment blocks. For the present analysis, the assessment period will be divided into three blocks: Session 2 to Session 12, Session 12 to 1-Month follow-up, 1-Month follow-up to 6-Month follow-up. The analysis will be assuming that the changes within each of these blocks will be linear. However, this may not be appropriate for the first block (Session 1 to Session 12) as individuals may change by different amounts throughout the trial. For example, individuals may see small amounts of benefit at the start of the trial, but larger amounts of benefit as treatment continues. To assess the appropriateness of a linear model of change for this block of assessments, plots of the model residuals at each assessment point will be inspected for systematic variation. If there is a systematic structure to the residuals (e.g., consistently larger at a given assessment or there appears to be some curvature), this would indicate that a linear model was not accurately capturing the changes in severity throughout the treatment sessions. If this is the case, the first block (Sessions 1-12) will be fitted with restricted cubic splines to represent the effects of time (Harrell, 2015).

Similar to the categorical treatment of time, the use of an unstructured covariance matrix for the residuals allows for a great deal of flexibility at the cost of estimating more parameters. An unstructured matrix estimates the residual variance at every assessment time, as well as the covariance in residuals between every assessment time. This residual structure is not feasible for the outcomes with 13 post-baseline measurements. An alternative for these outcomes is the continuous first-order autoregressive structure (CAR1) with heterogenous variances (i.e., different variances for each timepoint). The CAR1 residual structure is used to account for the temporally-correlated nature of repeated measures data. The CAR1 structure assumes that the assessments completed closer together in time will be more strongly correlated than those conducted further apart. This is completed using a ‘decay’ parameter, which determines the strength of correlation between two assessments as a function of the distance in time between them. By using a CAR1 structure with heterogenous variances, the

analyses with outcomes measured at 13 assessments can account for the autocorrelated nature of the residuals without requiring the estimation of as many parameters as would be required with an unstructured matrix. Analyses of pilot data suggest using a heterogeneous CAR1 structure to model the residual errors fits the data better than any of the other correlation structures available in R's nlme package (Pinheiro et al., 2019).

Binary Outcome with 2 Post-Baseline Measurements (SCID-5 Diagnosis)

As all individuals will have a diagnosis of SAD at the baseline assessment, no additional information is gained by including baseline diagnosis in the model. As such, the analysis of change in diagnostic status will only include the diagnoses at the 1-Month and 6-Month follow-ups. To determine whether the proportion of individuals with a SAD diagnosis at the 1-Month and 6-Month follow-ups is significantly different between the treatment groups, a pooled test of the difference in proportions will be carried out.

For each of the imputed datasets, the proportion of individuals in each treatment arm (IE-CBGT & VB-CBGT) with a diagnosis of SAD, and the standard error of that proportion, will be estimated. The difference in proportions, as well as the standard error of that difference, will also be estimated. These estimated proportions and standard errors for each imputed dataset will then be pooled using Rubin's Rules (van Buuren, 2018). These pooled results will give the percentage of individuals in each arm meeting diagnostic criteria at the 1-Month and 6-Month follow-ups, the difference between the arms, confidence intervals for difference, and the significance of the difference.

Ordinal Outcome with 2 Post-Baseline Measurements (Clinician-Rated Severity)

Because clinician-rated severity is an ordinal outcome (rated by category and the categories are ordered in severity) measured on multiple occasions, the analysis will use a cumulative-link mixed-effects model (CLMM) as implemented in the R package 'ordinal' (Christensen, 2019). The CLMM is an extension of the generalised linear mixed-model for use with ordinal data. Baseline severity will be included as a covariate and time will be entered as categorical. A random intercept for each group will be included, but as with the previous analyses will be removed if it shows little variance. To account for the correlation between one- and six-month follow up assessments, a random slope for patient will be used.

A key concern with the analysis of ordinal data is whether there are sufficient numbers of individuals at each of the rating categories. If there are only a small number of

individuals within a given severity rating, the estimated relationships could be easily influenced by outlying, or simply unrepresentative, responses. This is likely to be the case for the largest and smallest ratings of severity. If severity categories have low numbers of endorsements (< 5), those categories will be collapsed.

It should be noted that the CLMM relies on the assumption of proportional odds. This assumption implies that the effect of the covariates (e.g., treatment) are the same for every category. In other words, regardless of whether the individual has a high or low severity rating, their probability of changing category after receiving treatment is the same. If this is not the case, the resulting estimates of treatment effect are likely to be biased. An alternative is to estimate a different effect for each category. While this is more flexible, and does not assume proportional odds, it also requires estimating more parameters. The analysis of clinician-rated severity will first be carried out assuming proportional odds. The proportional odds assumption will then be evaluated. If there is a substantial departure from proportional odds for a given covariate, the analysis will be refitted with non-proportional odds for that covariate.

Reliable and Clinically Significant Change (SIAS)

Beyond simply testing whether the amount of change in social anxiety symptoms is statistically significant, it is also of interest to examine whether those changes are: a) beyond what would be expected due to measurement error, and b) clinically meaningful. This is the examination of reliable and clinically significant change (RC and CSC, respectively) described by Jacobson and Truax (1991). The analyses will examine the RC and CSC of the SIAS using the multiply-imputed datasets.

The calculation of RC is defined as:

$$RC = \frac{\left(\text{SIAS}_{\text{pos}} - \text{SIAS}_{\text{pre}}\right)}{\sqrt{2 * SE^2}}$$

Where:

$$SE = SD_{pre} * \sqrt{1 - r}$$

Where r is the reliability of the SIAS. For this analysis, the previously described McDonald's Omega coefficient will be used as the estimate of the reliability of the SIAS. If

RC > 1.96, this indicates that the amount of change is greater than can be attributed to measurement error.

The criteria for having achieved CSC were the same as described in previous pilot research (McEvoy et al., 2015). Individuals were required to have scored above a severity cut-off before treatment, achieved RC, and subsequently scored below the cut-off after treatment. The severity cut-off was defined as the mid-point between the clinical and normative means reported by Carleton et al. (2014), which was 40.56 for the SIAS.

To determine whether the proportions of individuals achieving reliable and/or clinically significant change significantly differed between treatment arms, a pooled test of the difference in proportions was carried out (the same procedure described for the analysis of diagnostic status).

Contrasts for Continuous and Ordinal Outcomes

The primary hypotheses are concerned with whether there are differences between the treatments at one month follow up. In order to determine this, for each continuous and ordinal outcome, the data from all the measurement occasions are jointly modelled using a LMM (rather than say, only using the one-month follow-up data in the analysis; see Mallinckrodt and Lipkovich (2017). We will then use contrast coefficients to (i) estimate the marginal means for each treatment at each measurement occasion, and (ii) compute the mean difference between treatments and a standard error at the one-month follow up time point. For continuous and ordinal outcomes, we will compute the contrasts using R's emmeans package. For the ordinal outcome this contrast is an odds ratio, representing the difference in probability of belonging to a higher severity rating between the two treatment arms.

Evolution of Outcomes Over Time

A secondary goal of this paper is to understand how differences between the groups evolve over time. Analyses of pilot data suggest these tend to be smaller at post-treatment than at one-month follow up. We have no pilot data for the six-month follow up. Therefore, we will use additional contrasts to examine differences between the treatments at week 12 (end of weekly treatment sessions) and the 6-month follow up. This will provide valuable information about the evolution of treatment outcomes before and after the primary endpoint.

Reporting of Results

Analysis Code and Output

All analysis code and output will be uploaded to the Open Science Foundation. This will ensure that all results are replicable, and that other scholars can easily extend our work. In the paper itself, we will summarise the key findings.

Standardized Effect Size

For continuous outcomes, we will convert the mean difference between groups into a standardised effect size by dividing the difference by the pooled pre-treatment standard deviation. We will also compute standardized within-group effect sizes, by subtracting the mean baseline score from each treatment's estimated marginal mean at the one-month follow up and dividing by the pooled pre-treatment standard deviation. For binary outcomes, the difference in proportion is an effect size. For ordinal outcomes, the odds ratio is an effect size.

Confidence Intervals

For each outcome, we will report a 95% confidence interval for the difference between groups at one-month follow up.

P-values

There have recently been calls to reform or abandon significance testing (for more details, see a recent special issue of the American Statistical Association's *American Statistician* journal here). In the present paper, we will report p-values continuously (e.g., $p = .078$) rather than just reporting whether they are less than or greater than .05.

Multiplicity

Due to the moderate sample size, there is a greater risk of Type II errors than Type I errors in the current trial. All hypotheses are also prespecified and the number of outcomes being analysed is limited. Therefore, no adjustment for multiplicity will be used.

Diagnostic Checks, Sensitivity, and Alternative Analyses

Model Assumptions

The graphical 'lineup' procedure will be used to assess the model assumptions of residual normality and normality of random effects for the mixed-model analyses of continuous and

ordinal outcomes. The lineup procedure involves creating a plot assessing a given model assumption (e.g., normality of residuals) from both the analytic model and from several simulated models where the assumption is met. If a blinded observer cannot distinguish the analytic model plot from the simulated model plots, then the assumption has been met (Loy, Hofmann, & Cook, 2017). Given that 100 multiply-imputed datasets were used in the analysis, it is not practical to check model assumptions using all of them. Rather, a single dataset will be randomly selected and used. The lineup plots will be generated by the data analyst and inspected by members of the research team that are blinded to the location of the analytic plot.

The mixed-effect models for continuous outcomes will initially be estimated with a Gaussian distribution and an identity link (i.e., assuming a linear relationship). If the diagnostic lineup plots identify issues with residual non-normality, alternative robust methods will be evaluated. These methods include the use of a sandwich estimator with the LMM, such as that implemented in the R package clubSandwich (Pustejovsky, 2019). Alternatively, a generalised LMM with an appropriate distribution (e.g., gamma or inverse gaussian) could be used, such as those implemented in the R package glmmTMB (Brooks et al., 2017). A weighted generalised estimating equations (wGEE) model could be used, as implemented in the R package wgeesel (Xu, Li, & Wang, 2018). Finally, a two-stage analysis could also be used, as this has shown to be very robust to violations of distributional assumptions (Overall & Tonidandel, 2010; Senn, Stevens, & Chaturvedi, 2000).

Model Convergence

A common difficulty with LMM analyses is model convergence (Bates, Kliegl, Vasishth, & Baayen, 2015). Model convergence occurs when the iterative maximum-likelihood procedure has arrived at a stable set of estimates (i.e., there is minimal change in estimates with further iteration). However, when models have large numbers of random effects this optimisation process can then take a large number of iterations before it converges (Bates et al., 2015). Problems with convergence can also be exacerbated by the presence of unnecessary random effects in the model, that is, if there are random effects for parameters that have very little variance between individuals or groups. (Bates et al., 2015).

If issues with model convergence are encountered in the present analysis, the maximum number of iterations for the optimisation procedure will first be increased. If

increasing the number of iterations does not improve convergence, more parsimonious random effect structures will be used (e.g., removing the random intercept for group).

Model Parsimoniousness

The analytic approaches described have been chosen for their increased flexibility and minimal assumptions about the direction or nature of relationships between variables. However, given the additional number of parameters that need to be estimated, this flexibility may come at the cost of precision in the estimates. To assess whether the flexible modelling of time is needed, plots of the observed and model-estimated values at each measurement occasion will be inspected. If simpler model structures (e.g., linear splines) appear appropriate, the model will be re-fitted, and the estimates inspected. Similarly, the residual variance-covariance matrix of the residuals will be inspected to identify whether a more parsimonious structure would be appropriate (e.g., CAR1).

Multiple Imputation

To assess whether the imputation model was consistent with the observed data, the analysis results will be compared with the results of applying each model to the observed (partially missing) data. The estimates from the multiply-imputed and the observed data should be broadly consistent. If the estimates markedly differ from each other, this could indicate issues with either the imputation model or algorithm, or the presence of bias in estimates with partially missing data that has been corrected by the imputation model.

A key sensitivity analyses will be testing the sensitivity of the results to violations of the missing at random (MAR) assumption. As multiple imputation assumes that the data are MAR, the imputed values could be biased if this is not the case. In other words, if there is a systematic reason for why some individuals did not adhere to treatment, ignoring this when imputing their missing responses could bias the results away from the ‘true’ effect. A common means of testing this assumption is called a ‘tipping-point’ analysis (Mallinckrodt & Lipkovich, 2017). The tipping-point approach to sensitivity analysis is implemented by modifying the imputed data by increasing amounts and identifying the magnitude of changes in imputed data that would be required for the results to longer be significant (Mallinckrodt & Lipkovich, 2017). This is achieved through the repeated use of what is called a ‘delta-adjustment’ of the imputed values (van Buuren, 2018). A delta-adjustment is simply the addition of a fixed value (delta) to every imputed response. In other words, this approach is

artificially raising or lowering the mean response of all non-adherent individuals. By using increasingly large values of delta (the fixed value), a sensitivity analysis can identify the extent to which the non-adherent individuals would need to differ in their response before there is no longer a significant difference between the treatment groups.

Summary of Confirmatory Analyses

Outcome	Analysis	Fixed Effects	Random Effects	Residual Covariance Structure
SPS & SIAS	Linear Mixed Model	• Baseline • Time (Categorical) • Treatment • Baseline*Time • Treatment*Time	• Intercept (Group)	• Unstructured • Heterogenous Variances
BFNES, PROMIS Anxiety & Depression	Linear Mixed Model	• Baseline • Time (Piecewise) • Treatment • Baseline*Time • Treatment*Time	• Intercept (Group)	• CAR1 • Heterogenous Variances
Diagnostic Status	Pooled Test of Difference in Proportions	N/A	N/A	N/A
Clinician-Rated Severity	Cumulative-Link Mixed Model	• Baseline • Time (Categorical) • Treatment • Baseline*Time • Treatment*Time	• Intercept (Group) • Intercept • Time	N/A

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., . . . Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*, 400. doi:10.3929/ethz-b-000240890

Carleton, R. N., Thibodeau, M. A., Weeks, J. W., Teale Sapach, M. J. N., McEvoy, P. M., Horswill, S. C., & Heimberg, R. G. (2014). Comparing short forms of the Social Interaction Anxiety Scale and the Social Phobia Scale. *Psychological Assessment, 26*(4), 1116-1126. doi:10.1037/a0037063

Christensen, R. H. B. (2019). ordinal - Regression Models for Ordinal Data (Version 2019.4-25) [R package]. Retrieved from http://www.cran.r-project.org/package=ordinal/

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis, 72*, 92-104. doi:10.1016/j.csda.2013.10.025

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. doi:10.1111/bjop.12046

Erceg-Hurn, D. M., & McEvoy, P. M. (2018). Bigger is better: Full-length versions of the Social Interaction Anxiety Scale and Social Phobia Scale outperform short forms at assessing treatment outcome. *Psychological Assessment, 30*(11), 1512-1526. doi:10.1037/pas0000601

Gamble, C., Krishan, A., Stocken, D., Lewis, S., Juszczak, E., Doré, C., . . . Loder, E. (2017). Guidelines for the content of statistical analysis plans in clinical trials. *JAMA, 318*(23), 2337-2343. doi:10.1001/jama.2017.18556

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29-48. doi:10.1348/000711006x126600

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival Analysis* (2nd ed.). New York: Springer.

Harvey, L. (2018). Statistical testing for baseline differences between randomised groups is not meaningful. *Spinal cord, 56*(10), 919.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19. doi:10.1037/0022-006X.59.1.12

Kelley, K. (2019). MBESS: The MBESS R Package (Version 4.6.0) [R package]. Retrieved from https://CRAN.R-project.org/package=MBESS

Loy, A., Hofmann, H., & Cook, D. (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics, 26*(3), 478-492. doi:10.1080/10618600.2017.1330207

Lu, K., & Mehrotra, D. V. (2010). Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Statistics in Medicine, 29*(4), 474-488. doi:10.1002/sim.3820

Mallinckrodt, C. H., Lin, Q., Lipkovich, I., & Molenberghs, G. (2012). A structured approach to choosing estimands and estimators in longitudinal clinical trials. *Pharmaceutical Statistics, 11*(6), 456-461.

Mallinckrodt, C. H., & Lipkovich, I. (2017). *Analyzing longitudinal clinical trial data: A practical guide*.

Marshall, A., Altman, D. G., & Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Research Methodology, 10*, 112-112. doi:10.1186/1471-2288-10-112

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology, 10*(1), 7. doi:10.1186/1471-2288-10-7

McEvoy, P. M., Erceg-Hurn, D. M., Saulsman, L. M., & Thibodeau, M. A. (2015). Imagery enhancements increase the effectiveness of cognitive behavioural group therapy for social anxiety disorder: A benchmarking study. *Behaviour Research and Therapy, 65*, 42-51. doi:10.1016/j.brat.2014.12.011

McEvoy, P. M., Moulds, M. L., Grisham, J. R., Holmes, E. A., Moscovitch, D. A., Hendrie, D., . . . Erceg-Hurn, D. M. (2017). Assessing the efficacy of imagery-enhanced cognitive behavioral group therapy for social anxiety disorder: Study protocol for a randomized controlled trial. *Contemporary Clinical Trials, 60*, 34-41. doi:10.1016/j.cct.2017.06.010

McEvoy, P. M., & Saulsman, L. M. (2014). Imagery-enhanced cognitive behavioural group therapy for social anxiety disorder: A pilot study. *Behaviour Research and Therapy, 55*, 1-6. doi:10.1016/j.brat.2014.01.006

McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics, 44*(1), 24-39. doi:10.1080/02664763.2016.1158246

National Research Council Panel on Handling Missing Data in Clinical Trials. (2011). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington D.C.: National Academies Press.

Overall, J. E., & Tonidandel, S. (2010). The case for use of simple difference scores to test the significance of differences in mean rates of change in controlled repeated measurements designs. *Multivariate Behavioral Research, 45*(5), 806-827. doi:10.1080/00273171.2010.519266

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). nlme: Linear and Nonlinear Mixed Effects Models (Version 3.1-140) [R package]. Retrieved from https://CRAN.R-project.org/package=nlme

Pustejovsky, J. (2019). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections (Version 0.3.5) [R package]. Retrieved from https://CRAN.R-project.org/package=clubSandwich

Senn, S. J. (1995). Base logic: Tests of baseline balance in randomized clinical trials. *Clinical Research and Regulatory Affairs, 12*(3), 171-182. doi:10.3109/10601339509019426

Senn, S. J. (2008). *Statistical issues in drug development* (2nd ed. Vol. 69). West Sussex, England: John Wiley & Sons.

Senn, S. J., Stevens, L., & Chaturvedi, N. (2000). Repeated measures in clinical trials: Simple strategies for analysis using summary measures. *Statistics in Medicine, 19*(6), 861-877. doi:10.1002/(sici)1097-0258(20000330)19:6<861::Aid-sim407>3.0.Co;2-f

Templ, M., & Filzmoser, P. (2008). Visualization of missing values using the R-package VIM. *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology*.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). FL: Boca Raton: Chapman and Hall/CRC.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. doi:10.1002/sim.4067

Xu, C., Li, Z., & Wang, M. (2018). wgeesel: Weighted Generalized Estimating Equations and Model Selection (Version 1.5) [R package]. Retrieved from https://CRAN.R-project.org/package=wgeesel

Yuan, I., Topjian, A. A., Kurth, C. D., Kirschen, M. P., Ward, C. G., Zhang, B., & Mensinger, J. L. (2019). Guide to the statistical analysis plan. *Pediatric Anesthesia, 29*(3), 237-242. doi:10.1111/pan.13576