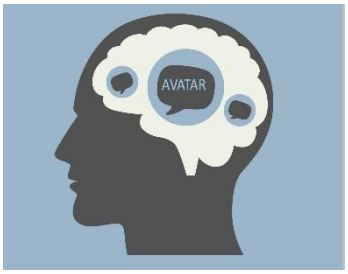


# Optimising Avatar Therapy for Distressing Voices: A Multi-Centre Randomised Controlled Trial

## Statistical Analysis Plan

Version 1.1

11/05/2023

ISRCTN: ISRCTN55682735

IRAS project ID: 277118

REC reference: 20/LO/0657

SAP based on Protocol Version: 1.4

**SAP Sign off**

**CI signature:**

Print name: Prof Philippa Garety

Date: 20/06/23

**Trial statistician signature:**

Print name: Prof Richard Emsley

Date: 16/06/23

**Chair of the TSC signature:**

Print name: Prof Matthew Broome

Date: 13th July 2023

**Independent DMEC statistician signature:**

Print name: Dr Reuben Ogollah

Date: 21/06/2023

Version history:

| Version Number | Date of sign-off | Changes from previous version |
|---|---|---|
| 1.0 | | First signed version |
| 1.1 | | • Added categorisation of ethnicity to be used in trial reporting<br>• Added socio-economic analysis |

# Contents

# AVATAR2 Statistical Analysis Plan

This document details the presentation and analysis strategy for the primary paper reporting results from the AVATAR2 trial. It is intended that the results reported in these papers will follow the strategy set out herein; subsequent papers of a more exploratory nature will not be bound by this analysis plan but will be expected to follow the broad principles laid down for the primary paper(s). The principles are not intended to curtail exploratory analysis or to prohibit sensible statistical and reporting practices, but they are intended to establish the strategy that will be followed as closely as possible, when analysing and reporting the trial.

Chief Investigator: Professor Philippa Garety

Trial Coordinator: Dr Clementine Edwards

Senior Trial Statistician: Prof Richard Emsley

Trial statistician: Dr Hassan Jafari

# 1. Description of AVATAR2 trial

Voices are the most commonly reported form of auditory hallucinations (Mawson et al., 2010) heard by as many as 70% of people suffering from schizophrenia (Waters et al., 2012). They are often very distressing experiences involving threats, denigrating commentary and commands to self-harm or assault others. Voices frequently persist for many years despite pharmacotherapy, and the currently recommended psychological therapy for voices, cognitive behavioural therapy for psychosis (CBTp) has a modest impact (van der Gaag et al., 2014). Consequently, there is considerable interest in the development of novel therapies that build on the CBTp experience, but which are both shorter and capable of being delivered by a wider workforce. AVATAR therapy is one such development. It involves the use of a digital representation (avatar) of the entity the person believes is the source of the voice. During sessions, this avatar is voiced by the therapist who uses a console to switch between speaking in the transformed avatar voice and speaking in their own voice (as therapist). Therapy proceeds as a trialogue between participant, avatar and therapist, with the avatar changing to be less intimidating and persecutory in response to changes in the participant's responses, guided by the therapist.

## 1.1. Research objectives

The trial will address questions of treatment efficacy, therapy delivery, and its implementation in NHS settings, including testing the provision of the integrated and enhanced software platform for the delivery of therapy employed in the clinical trial, together with the operational and therapy manuals.

### 1.1.1. Primary objective

Are both brief and extended AVATAR therapy effective in reducing distress associated with voices over 16 and 28 weeks when added to treatment as usual (TAU) in comparison to TAU alone?

### 1.1.2. Secondary objectives:

- Does AVATAR therapy reduce the frequency and distress associated with voices by reducing the perceived omnipotence and malevolence of voices?
- Does the extent to which the participant experiences a more highly characterised voice (i.e. has a clear and detailed belief about the entity behind the voice) determine response to treatment?
- Does treatment effect vary by other clinical and demographic characteristics of the participant?

### 1.1.3. Hypotheses

1. AVATAR-brief will be more effective in reducing voice-related distress, total voice severity and voice frequency than TAU at post-treatment (16 weeks) and follow up (28 weeks)
2. AVATAR-extended will be more effective in reducing voice related distress, total voice severity and voice frequency than TAU, at post-treatment (16 weeks) and follow up (28 weeks)
3. AVATAR-extended will reduce perceived omnipotence and malevolence (BAVQ-R) compared to TAU and these improvements will mediate change in the primary outcome.
4. In both AVATAR-brief and AVATAR-extended treatment effects on voice-related distress will be mediated by anxiety reduction, as measured by ESM in daily life.
5. Greater baseline complexity of voice characterisation will moderate the treatment effects of AVATAR-brief and AVATAR-extended compared to TAU. Other clinical and demographic characteristics will be explored as potential moderators.
6. AVATAR-brief and AVATAR-extended will both have favourable incremental cost-effectiveness ratios compared to routine care.

## 1.2. Trial design including blinding.

A three-arm randomised controlled trial, with 1:1:1 allocation and blinded assessors, to test the efficacy of brief and extended AVATAR therapy in people diagnosed with schizophrenia spectrum disorders in reducing the distress and frequency of voices when added to standardised Treatment As Usual (TAU) compared to standardised TAU alone. The RCT will be carried out in 4 main sites, each with at least two NHS Trust partners. The trial will compare the effect of two forms of AVATAR therapy plus treatment as usual (TAU) to TAU alone.

*Description of Intervention:* AVATAR therapy is a brief therapy using digital technology to enable the participant to have a dialogue with the entity they believe to be responsible for the voices they hear. The first meeting with the voice-hearer involves a detailed assessment of the voice (including the key aspects of verbatim content, voice characterisation, the nature of voice-hearer relationship and developmental history (including trauma). A digital representation of the entity is then built and voiced by the therapist who uses a console to switch between speaking in the transformed avatar voice and their own voice (as the therapist). Therapy proceeds as a trialogue between participant,

avatar and therapist, with the avatar changing to be less intimidating and persecutory in response to changes in the participant's responses, guided by the therapist. The therapy may proceed in two phases: In the phase 1, (Exposure and Assertiveness) the avatar delivers verbatim voice content (including threats and abuse) and the person practices assertive responding. Over time the avatar becomes less hostile as the person develops increased power and control within the dialogue. This cues a second phase (phase 2) with formulation-driven therapeutic targets, which can include work on beliefs about voices, self-concept and trauma. Participants will be allocated to receive either of two levels of AVATAR therapy (Brief therapy or Extended Therapy), in addition to standardized TAU. The AVATAR therapy itself will be delivered according to the clinical manual, developed following the previous trial, but in two levels. In the extended therapy, delivery occurs over 12 sessions and includes both phases of treatment as described earlier and detailed in the AVATAR clinical manual. The brief therapy includes a focus on the existing phase1 exposure/assertiveness component, over 6 sessions. Trial therapists will be provided with training and ongoing supervision.

*Fidelity to the intervention:* Fidelity to the clinical manual will be assessed by the therapist completing a session-by-session checklist of specified components (these will also be used during training and ongoing supervision). Each session will be audio recorded with consent. After completion of training, therapist competence will be assessed by an expert in AVATAR therapy for general/clinical and AVATAR-specific skills using ratings adapted from the first AVATAR therapy trial and allowing for differing skill requirements for each level of therapy. Each trial therapist will be rated for competence based on the review of early, mid and late session therapy delivery for at least one completed intervention. Ratings will be conducted for two cases for therapists who deliver completed therapy with more than five participants.

*Intervention stopping guidance:* It will be made clear to each participant that, should they find any aspect of the research distressing, and/or no longer wish to continue, they will be able to withdraw from the therapy without this impacting on their usual clinical care in any way. Should the therapy prove aversive or distressing to many participants, we would consider an elective stop, however, this was not a concern in the previous trial and we think this unlikely. We have the oversight of the DMEC committee who will be reviewing trial progress and the occurrence of adverse events.

The trial may be prematurely discontinued by the Wellcome Trust based on new safety information or for other reasons given by the Data Monitoring & Ethics Committee or Trial Steering Committee. If the study is prematurely discontinued, active participants will be informed and no further randomisations will be performed.

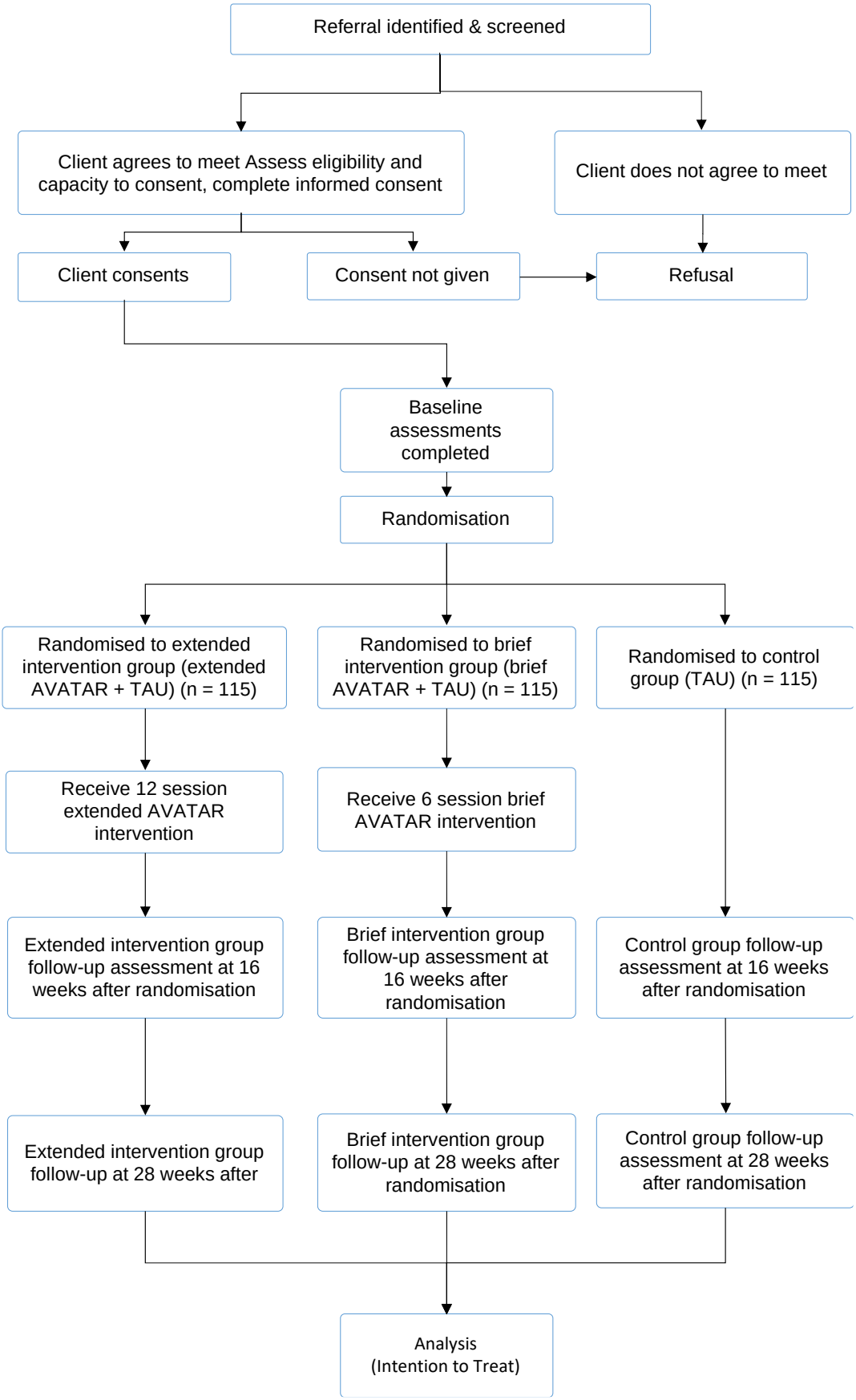


**Figure 1**. Trial design flow diagram

## 1.3. Method of allocation and blinding of groups

Once baseline assessments are complete, the individuals will be randomised to one of the treatment arms. Randomisation will be done in a 1:1:1 ratio. Randomisation is at the patient level and is performed using an online randomisation system King's Clinical Trials Unit (KCTU). Randomisation is stratified by *recruitment centre* and *baseline voice entity characterisation* with dynamically generated permuted blocks of random size.

The procedure is as follows: On completion of the baseline assessment, the researcher electronically submits details of each participant to the CTU. This includes: participant ID number, site, initials and date of birth. The system immediately notifies unblinded members of the research team; the Trial Coordinator, who records the outcome, the site PI and trial therapist, who notifies the participant.

*Blinding*: Research staff will be blind to treatment allocation. The blinding procedure will be explained to participants and they will be reminded not to inform research workers of therapy allocation. We will be using a system of web-based data entry that ensures assessors do not have access to information in the database that might reveal allocation. Breaks in blindness will be monitored and recorded

## 1.4. Duration of the treatment period

The AVATAR therapy will be delivered in two forms (brief and extended, one in each arm of the trial).

The participation of each person within the trial will be 7 months from assessment/randomisation to the 28 weeks follow up.

## 1.5. Frequency and duration of follow-up

Baseline assessments will be administered by a research assistant following consent. These assessments consist of a range of self-report and interview measures involving questions about voice frequency, content and associated distress as well as mood, self-esteem and contact with health services. Assessments will be conducted at baseline, post treatment at 16 weeks, and at 28 weeks follow-up.

*Intervention/therapy records*: Assessments and therapy sessions will be digitally recorded (after first establishing consent) to allow for assessment of adherence to the research protocol and assessment ratings. sessions attended.

## 1.6. Scheduled visit windows

At the 16- and 28-week assessments, data is scheduled to be collected within a visit window of one week prior to 16- or 28-week due date, and up to four weeks following this date. Data may be collected after the four week time period in exceptional circumstances, with the reason recorded. The exact date of each completed assessment will be recorded.

## 1.7. Data collection

### 1.7.1. Eligibility screening

*Inclusion Criteria*

- Aged 18+ years.
- Currently under the care of a specialist mental health team (inpatient and outpatient settings).
- Have current frequent and distressing voices, (as measured by a score of at least 2 on the sum of the intensity of distress and frequency items of the PSYRATS (Voices) scale), persisting for at least 6 months and spoken in English.
- Speak and read English to a sufficient level to provide consent and complete the assessment procedures.
- A clinical diagnosis of Schizophrenia spectrum disorder (ICD10 F20-29) or affective disorder with psychotic symptoms (ICD-10 F30–39, subcategories with psychotic symptoms) as determined through clinical records and additional consultation with clinical team if unclear.

*Exclusion criteria*

- Primary diagnosis of substance disorder, personality disorder or learning disability.
- Lacking capacity to consent.
- Profound visual/hearing impairment or insufficient comprehension of English to be able to engage in assessment or therapy.
- Currently undertaking individual psychological therapy for voices.
- Currently experiencing an acute mental health crisis.

### 1.7.2. Baseline

Baseline assessments will be administered by a research assistant following consent. These assessments consist of a range of self-report and interview measures involving questions about voice frequency, content and associated distress as well as mood, self-esteem and contact with health services.

In addition to the primary and secondary outcomes, the following measures are collected at baseline only:

- Demographics and Clinical Information
- Complexity of voice characterisation (more or less characterisation)
- Scale for the Assessment of Positive Symptoms
- Clinical Assessment Interview for Negative Symptoms (CAINS)
- Mini-Trauma and Life Events (TALE) Checklist
- The Relationships Questionnaire (RQ) (Partial)

### 1.7.3. Primary outcome measures

Distress associated with voices as measured by the distress dimension of the Psychotic Symptoms Rating Scale- Auditory Hallucinations (Haddock, 1999) over 16 and 28 weeks. The distress dimension has five items in total: voice distress: negative content (2 items amount and degree), voice distress (2 items amount and intensity), and control item.

### 1.7.4. Secondary outcome measures

- Frequency of voices as measured by the Psychotic Symptoms Rating Scale – Auditory Hallucinations (Haddock, 1999)
- Total score on Psychotic Symptoms Rating Scale – Auditory Hallucinations
- Remission of Voices (standalone item)
- Beliefs about Voices Revised (BAVQ-R) (Chadwick et al., 2000)
- Voices acceptance and action scale (VAAS) (Shawyer et al., 2007).
- First item (power) from the Voice Power Differential Scale (Birchwood et al., 2000).
- Measure of anxiety, using Experience Sampling Method at quasi-random occasions during the waking day over a 7-day period at baseline, 16 and 28 week assessment points (Myin-Germeys et al., 2018).
- Wellbeing and patient-led outcome measures: Warwick-Edinburgh Mental Well-being Scale (Tennant et al., 2007), Choice of Outcome in CBT for Psychoses (CHOICE) (Greenwood et al., 2010).
- Clinical characteristics: Beck Depression Inventory-II, Depression Anxiety and Stress Scales (DASS-21), Psychotic Symptoms Rating Scale (PSYRATS-DEL), International Trauma Questionnaire (ITQ).

### 1.7.5. Serious adverse events

Serious adverse events will also be categorised as the following:

1. Distress associated with completion of assessment measures
2. Significant distress during the avatar therapy
3. Admission to hospital for psychological health event
4. Admission to hospital for physical health event
5. Referral to crisis team
6. Violent incident necessitating police involvement (victim)
7. Violent incident necessitating police involvement (accused)
8. Deliberate self-harm
9. Other psychological health event
10. Other physical health event

**Table 1.** List of measures

| | | | Time Points | | | |
|---|---|---|---|---|---|---|
| | **Variable** | **Primary function** | **Baseline** | **W16** | **W28** | **ON** |
| 1 | Eligibility | Screening | ● | | | |
| 2 | Demographics and Clinical Information | Baseline | ● | | | |
| 3 | Randomisation | Baseline/CONSORT | ● | | | |
| 4 | PSYRATS – Auditory Hallucinations | Primary outcome | ● | ● | ● | |
| 5 | Complexity of Voice Characterisation | Baseline/Moderator | ● | | | |
| 6 | Hallucinations Remission Score | Secondary outcome | ● | ● | ● | |
| 7 | Beliefs about Voices Revised (BAVQ-R) | Secondary outcome | ● | ● | ● | |
| 8 | Voices acceptance and action scale (VAAS) | Secondary outcome | ● | ● | ● | |
| 9 | Voice Power Differential Scale (Partial) | Secondary outcome | ● | ● | ● | |
| 10 | Experience Sampling Methodology | Secondary outcome | ● | ● | ● | |
| 11 | Beck Depression Inventory-II | Secondary outcome | ● | ● | ● | |
| 12 | Depression Anxiety and Stress Scales (DASS-21) | Secondary outcome | ● | ● | ● | |
| 13 | Warwick-Edinburgh Mental Well-being Scale | Secondary outcome | ● | ● | ● | |
| 14 | Scale for the Assessment of Positive Symptoms (SAPS) | Descriptor | ● | | | |
| 15 | PSYRATS – Delusions | Secondary outcome | ● | ● | ● | |
| 16 | CAINS | Descriptor | ● | | | |
| 17 | Adapted CHOICE Short Form | Secondary outcome | ● | ● | ● | |
| 18 | Mini-Trauma and Life Events (TALE) Checklist | Moderator | ● | | | |
| 19 | International Trauma Questionnaire | Secondary outcome | ● | ● | | |
| 20 | TVAQ (Partial) | Moderator | ● | | | |
| 21 | The Relationships Questionnaire (RQ) (Partial) | Moderator | ● | | | |
| 22 | CSRI | Secondary outcome | ● | ● | ● | |
| 23 | EQ-5D-5L | Secondary outcome | ● | ● | ● | |
| 24 | COVID-19 Context Questionnaire | Descriptor | ● | ● | ● | |
| 25 | Psychological and Psychosocial Interventions Log | Descriptor | | | | ● |
| 26 | Antipsychotic Medications Log | Descriptor | | | | ● |
| 27 | Withdrawal Form | CONSORT | | | | ● |
| 28 | Working Alliance Inventory - Short Form Revised * | Descriptor | | | | |
| 29 | Therapy Attendance Log | Descriptor | | | | ● |
| 30 | Adverse Events Log | Safety | | | | ● |

*Data collection at session 4 and 10 of therapy, ON: Ongoing

## 1.8. Sample size estimation (including clinical significance)

Summary: Total N=345 (87 per site), with n=115 per treatment arm.

For the current proposal, we are interested in three comparisons in a hierarchical order, with plausible effect sizes for the first two of these, which we will test formally, based on the findings of the previous AVATAR therapy trial (Craig et al, 2018). There we found a clinically meaningful reduction in PSYRATS-AH distress of 4.8 points, with an effect size of approximately d=0.8, but we have conservatively reduced this for the current trial, to take into consideration the increase in number of centres, the follow-up comparison (not only end of treatment) and a more pragmatic trial design. The third and final comparison will be exploratory.

1. Extended (phase 1+phase 2: 12 sessions) AT vs. TAU – plausible effect size 0.6

2. Brief AT (phase 1: 6 sessions) vs. TAU – plausible effect size 0.5

3. Extended AT vs. brief AT (exploratory comparison)

The study will be powered for an overall treatment effect at an 5% significance level, accounting for 2 multiple comparisons in which the tests are correlated (at r=0.5), giving an alpha level for each test of 0.035. Accordingly, a sample size of 92 per group or 276 in total in the analysis set will have 90% power to detect a minimum clinically significant difference (effect size) of 0.5 standard deviations. We would seek to recruit 345 participants in total at baseline, allowing for conservative attrition rates of 20%.

# 2. Data analysis plan – Data description

## 2.1. Recruitment and representativeness of recruited patients

CONSORT flow chart will be constructed [1] and displayed as in **Figure 2**. This will include the number of eligible patients, number of patients agreeing to enter the trial, number of patients refusing, then by treatment arm: the number of patients who receive/do not receive the allocated interventions, the number continuing through the trial, the number withdrawing, the number lost to follow-up and the numbers excluded/analysed.

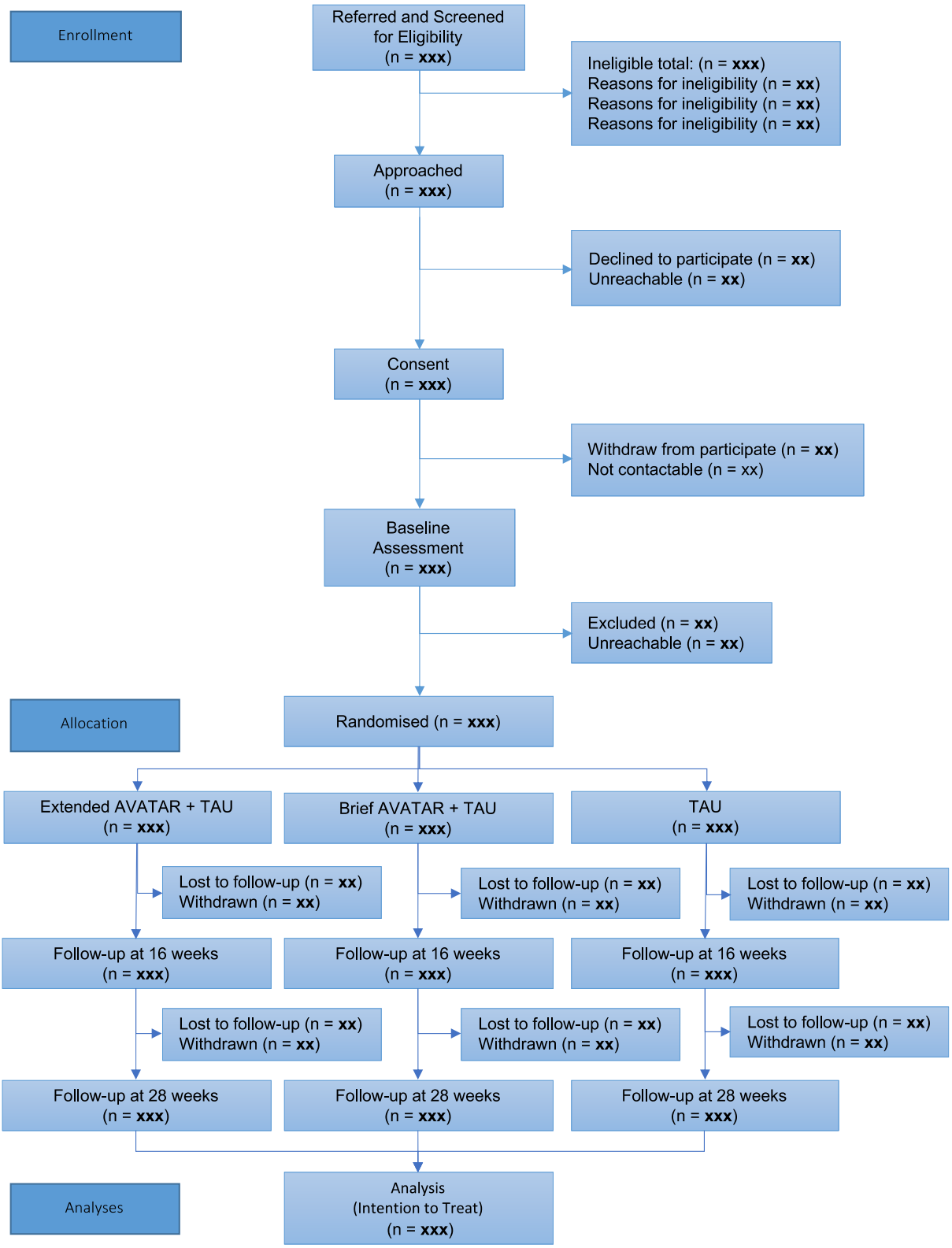


**Figure 2.** Template CONSORT diagram for AVATAR2 trial

## 2.2. Baseline comparability of randomised groups

Appropriate summary statistics will be applied to describe demographic and clinical measures: mean and standard deviation for all symmetric (non-skewed) distributed measures; median, 25th and 75th quartiles for skewed distributions. QQ plots and histograms will be used to assess data distributions of continuous measures. Categorical outcomes will be described using both numbers and proportions (percentage).

The baseline differences between the three randomised groups will be presented descriptively and will not be tested for between-group differences. The randomisation of participants to intervention groups means that any imbalance over all measured and unmeasured baseline characteristics is, by definition, due to chance. According to the CONSORT statement, significance testing of baseline differences in randomized controlled trials should not be performed. (Moher et al., 2010), unless otherwise specified.

## 2.3. Adherence to allocated treatment and treatment fidelity

Assessments and therapy sessions will be digitally recorded (after first establishing consent) to allow for assessment of adherence to the research protocol and assessment ratings.

*Treatment adherence:* The number of active sessions attended (i.e. session involving active avatar dialogue) by the participant will be measured to assess treatment adherence. Therapy Attendance Log will be used to capture the attended sessions.

*Treatment fidelity:* Fidelity to the clinical manual will be assessed by the therapist completing a session-by-session checklist of specified targets.

Adherence to treatment will be described by the median, 25th and 75th percentiles and range of the number of sessions offered and attended, length of sessions. The number/proportion of sessions a participant adhered to the treatment will be determined at the binary cuts of X sessions for each arm.

## 2.4. Loss to follow-up and other missing data

The numbers and proportions of participants with missing data for each baseline, primary and secondary variable will be summarised overall, and by arm and time point. The baseline characteristics of those missing follow up (at 16 and 28 weeks) will be compared to those with complete follow up using descriptive statistics and if possible, depending on how many cases, a logistic predictor of missingness model (27). The number and proportion actively withdrawing from the trial and reasons for withdrawal will be summarised overall and by treatment group separately from those that are passively lost to follow-up.

## 2.5. Assessment of outcome measures (unblinding)

Incidences of unblinding for blinded research staff will be reported, with the proportion (n (%)) of those affecting the primary outcome.

## 2.6. Descriptive statistics for outcome measures

For each primary and secondary outcome, a summary of results will be presented by trial arm. Appropriate summary statistics will be applied to describe demographic and clinical measures: mean and standard deviation for all symmetric (non-skewed) distributed measures; median, 25th and 75th quartiles for skewed distributions. QQ plots and histograms will be used to assess data distributions of continuous measures. Categorical outcomes will be described using both numbers and proportions (percentage). When necessary, we will summarise outcomes, by trial arm, by the time points, by site and by main demographic classifying variables.

## 2.7. Description of therapists

The number of participants per therapist will be summarised by numbers and proportions (percentage).

The self-report versions of the Working Alliance Inventory - Short Form Revised (WAI-SF-R) will be summarised by mean and standard deviation of the total score (can be tabulated by centre) which will form a descriptive of therapeutic alliance.

## 2.8. Covid-19

The Covid-19 context questionnaire (6 items) will be summarised at each timepoint. Additionally, we will summarise the delivery mode of the intervention (Face-to-face vs videoconferencing).

# 3. Data Analysis Plan – Inferential analysis

## 3.1. General analysis principles

This Statistical Analysis Plan will be agreed with the Trial Steering Committee and Data Monitoring and Ethics Committee before any inspection of post-randomisation data by the research team. Analyses will be carried out by the junior trial statistician under the supervision of the senior trial statistician.

Significance level (type 1 error) will be 0.035 for all analyses, and 96.5% confidence intervals will be reported. This accounts for multiple testing of the pairwise comparisons with the TAU only group. There will be no further adjustment for multiple testing for primary or secondary outcomes.

### 3.1.1. Analysis populations and estimands

The primary estimand will be the treatment policy estimand. The primary analyses will be carried out using the intention to treat sample: participants will be analysed in the group they are randomised to, and available data from all participants is included, including those who do not complete therapy. Every effort will be made to follow up all participants in both arms for research assessments.

### 3.1.2. Reporting guidelines

We will report data in line the most recent relevant Consolidated Standards of Reporting Trials (CONSORT) guidelines and include a CONSORT flow diagram (see Figure 2). These include:

- CONSORT extension for Social and Psychological Interventions (Grant et al, 2018)
- CONSORT extension for reporting harms outcomes (Ioannidis et al, 2004)
- CONSORT extension for Reporting of Multi-Arm Parallel-Group Randomized Trials (Juszczak, 2019)

### 3.1.3. Timing of analysis

● *Database completion and checking, dissemination and implementation plan*

All follow-ups 28 weeks post baseline (i.e. 12 weeks post end-of-treatment) completed; database fully checked, cleaned and locked (July 23)

● *Final analysis, dissemination and implementation, and writing up*

Analysis of primary outcome study data completed and final report with dissemination plan drafted. (August to October 23)

In order to ensure timely publication for the primary research question, analysis of the primary and secondary research objectives will take place following last patient last visit. The final analysis will therefore take place following last patient last visit at 28 weeks post-randomisation, where all available time points will be included in each model.

### 3.1.4. Outliers

Potential data outliers will be identified during the screening and cleaning process of dataset and queried as appropriate. Any data points that are identified as possible outliers, but are subsequently verified through the query process, will be treated as valid data, and analysed accordingly.

## 3.2. Main analysis of treatment differences

### 3.2.1. Analysis of primary outcome

The primary analysis (hypothesis 1&2 in Section 1.1.1) for between-group difference in the distress of auditory hallucinations as measured by PSYRATS-AH distress score will be analysed using a mixed (random) effects model at all post-randomisation time points (Week 16 and 28). Fixed effects will be centre, baseline assessment for the outcome under investigation, voice characterisation, treatment, time and time*treatment interactions. Participant and therapist will be included as random intercepts, with the participants in the TAU arm considered as being in individual clusters of size 1. Marginal treatment effects will be estimated for outcomes at each time point and reported separately as mean adjusted differences in scores between the randomised groups with 96.5% confidence intervals and two-sided p-values. For binary secondary outcomes, the same approach will be followed using logistic mixed models.

The random effect structure will account for repeated measures and clustering due to the nested design and allow estimates of separate ICCs in both randomised arms. All models will be estimated using maximum likelihood estimation, which allows for missing outcome data under the Missing At Random assumption; we may also use inverse probability weighting to adjust for non-adherence to allocated treatment and other intermediate outcomes as predictors of future loss to follow-up.

In addition, we will report estimates for Cohen's D effect sizes at 16 and 28 weeks as the adjusted mean difference of the outcome divided by the sample standard deviation of the outcome at baseline. Confidence intervals for Cohen's D will be calculated by dividing the confidence limits by the sample standard deviation of the outcome at baseline. These will be displayed in a Forest Plot with the primary outcome at the top, followed by secondary outcomes, with a separate plot for each time point.

### 3.2.2. Analysis of secondary outcomes

For continuous clinical secondary outcomes listed in section 1.7.4, we will follow the same model as the primary analysis; linear mixed models including the outcome measures at all post-randomisation time points with a time by treatment interaction to allow the estimation of the between arm difference at each time point.

For binary secondary outcomes listed in section 1.7.4, a logistic mixed effects model will be used, similarly with 3 levels including the outcome at all post-randomisation time points, and the reported treatment effect will be the odds ratio at each time point.

### 3.2.3. Interpretation of results

For each comparison of AVATAR-brief versus TAU, and extended AVATAR versus TAU, if the estimated between-group difference at 16-weeks is statistically significant we will conclude that there is a treatment effect on the outcome at the end of the intervention period. This will constitute partial support of our hypothesis.

If the estimated between-group difference at 28-weeks is statistically significant, we will conclude that there is a treatment effect on the outcome at follow-up. If there is a statistically significant between-group difference at 28-weeks but not at the earlier 16-week time point, this will constitute partial support of our hypothesis.

If there is a statistically significant between-group difference at both time points, we will conclude that the treatment effect is sustained and this will constitute full support of our hypothesis.

For the primary outcome of PSYRATS-distress, we will assess the magnitude of the between-group difference against the plausible effect sizes in the sample size calculations.

### 3.2.4. Adverse events

Adverse events (AEs) and serious adverse events (SAEs), listed by event type, will be summarised by time point and randomised group. Each table will detail the number of participants who were still in the trial at the time points by randomisation group. If a sufficiently large number of SAEs are reported, we will explore the use of graphical methods to present SAEs between groups, based on the methods described in Phillips, et al. (2020).

Serious adverse events will be summarised by type:

1. Distress associated with completion of assessment measures
2. Significant distress during the avatar therapy
3. Admission to hospital for psychological health event
4. Admission to hospital for physical health event
5. Referral to crisis team
6. Violent incident necessitating police involvement (victim)
7. Violent incident necessitating police involvement (accused)
8. Deliberate self-harm
9. Other psychological health event
10. Other physical health event

By category of relatedness to the trial:

1. Therapy-related (Yes, related - Possibly - No, unrelated)
2. Device-related (Yes, related - Possibly - No, unrelated)

3. Assessment-related (Yes, related - Possibly - No, unrelated)

By severity:

1. Not serious
2. Category A: Death Category
3. Category B: Incidents which acutely jeopardise the health or psychological wellbeing of...
4. Category C: Resulting in injury requiring immediate medical attention

### 3.2.5. Mediation analysis

For hypothesis 3&4 (Section 1.1.1), causal mediation analysis will be based on parametric regression models. For each mediator separately, this involves estimating a linear model for each mediator with random allocation, baseline outcome, baseline mediator, site and characterisation as covariates, and separately estimating a linear model for each outcome with the mediator, random allocation, baseline outcome, baseline mediator, site and characterisation as covariates. The effect of random allocation on the mediator is multiplied by the effect of mediator on the outcome to estimate the indirect effect, and the effect of random allocation on outcome in the model including mediator is an estimate of the direct effect. The indirect and direct effects sum to the total effect and bootstrapping with 1000 replications will be used to obtain valid standard errors for the causal effects.

### 3.2.6. Moderation analysis

For the moderation analyses (hypothesis 5), these will be conducted by adding interaction terms between random allocation, time and the respective moderators. These moderators will include voice-characterisation, gender, ethnicity, socio-economic status (see Appendix 1 for more details), trauma and attachment measures (TALE, TVAQ, RQ) and may include other baseline measures. The difference in treatment effect between unit levels of the moderator can be interpreted as the difference in the estimated treatment effect between a participant with a moderator value at baseline of a+ 1 and a participant with a moderator value at baseline of a.

### 3.2.7. Exploratory ESM analysis

The full ESM analysis is not covered by this Statistical Analysis Plan. Exploratory ESM analysis can only be conducted on the subsample of the trial participants who consent to take part in the ESM study. This will contain participants randomised to any of the three groups, but is different to the intention-to-treat sample for the primary analysis.

To analyse the ESM data at the two time points we will use multilevel models with an appropriate random effect structure. For analyses involving ESM variables as outcome, an additional level of nesting will be included, with multiple ESM observations (level 1) being nested within time points (level 2) and time points as nested within subjects (level 3).

For analysis to compare the treatment and control groups on ESM variables, this will be performed by adding a fixed effect for randomised group into the multilevel models and will follow an intention-to-treat principle in the subgroup of ESM participants.

### 3.2.8. Statistical considerations

*Missing data*

Measures will be taken to minimize missing outcome data, continuing to gather follow up data from participants who wish to drop out after randomisation (unless the participant is unwilling/withdraws consent). However, it is likely that there will be some missing data in post-randomisation variables as participants are lost to follow-up.

Although baseline data should be complete prior to randomisation, there may be some limited missing data. Descriptive baseline summaries will be presented as complete case. The proportion of missing data will be summarised by scale/assessment. If any of the baseline measures are found to relate to missing primary outcome at 16 or 28 weeks as per Section 2.4, we will consider adjusting for them in models for the primary outcome as a sensitivity analysis. To allow for this, any baseline measure considered as a covariate in the main model would best be imputed to a full single dataset. Missing baseline covariate data will therefore be imputed using the missing indicator method where a dummy variable for missingness will be included as a covariate in the model for binary data, and mean imputation for continuous data, as per the recommendations of White and Thompson (2005).

For questionnaire outcome measures where there are published methods for dealing with missing items, these will be applied. Otherwise, we will prorate missing items only when there are no more than 20% missing items (i.e. for a ten item questionnaire, prorate only where one or two items are missing) by replacing the missing item values with the mean value of the complete items for each individual. If after prorating there are still missing total questionnaire scores at baseline, these will be imputed as described in the paragraph just above.

Multiple imputation is not planned for the primary analysis. Where imputation can be required is if post-randomisation variables, such as adherence, are related to missing follow-up data. Post-randomisation variables cannot be included in analysis models being used to assess treatment differences based on the intention to treat principle; however, they can and should be included in a multiple imputation model if they predict missing data. In order to assess this, the analysis described in Section 2.4 will be repeated, with adherence as described in Section 2.3 as a predictor. If adherence is found to be associated with missing follow-up data, we will consider using multiple imputation as a sensitivity analysis, with an imputation model including all primary and secondary outcomes, all predictors and the adherence variable and potentially other auxiliary variables if needed. Multiple imputation with chained equations (MICE) will be used, which provides valid inference under the missing at random assumption.

*Model assumption checks*

The linear models assume normally distributed outcomes. Residual plots will be used to assess any departures from normality; if any exist, we will look for a more appropriate distribution shape.

### 3.2.9. Sensitivity analyses

There are no planned sensitivity analyses, beyond those specified above.

### 3.3. Exploratory analyses

There are currently no exploratory analyses planned as part of the primary analysis. After the publication of the CONSORT and SPIRIT Extension for RCTs reVised in Extenuating circumstances (CONSERVE) (Orkin et al, 2021), we will consider whether any additional exploratory analysis is required to assess the robustness of any findings due to changes caused by the COVID-19 pandemic. Any changes required will be documented in this SAP. This analysis plan does not cover any additional secondary exploratory analysis that may be requested by reviewers.

## 4. Software

*Data management*: An online data collection system for clinical trials (MACRO; InferMed Ltd) will be used for data entry and storage. This is hosted on a dedicated server at KCL and managed by the KCTU. The KCTU Data Manager will extract data periodically as requested using Data Extraction and Randomisation Extraction forms and provide these in comma separated (.csv) or Stata software format.

*Statistical analysis*: The latest version of Stata (currently v17.1) will be used for data description and the main inferential analysis.

## 5. Health Economic Analysis Plan

There is a separate Health Economic Analysis Plan prepared by the Health Economics team. Therefore the methods and plan for this cost-effectiveness analysis is reported elsewhere and not covered by this document.

## 6. References

Altman, D. G., & Dore, C. J. (1991, May). Baseline comparisons in randomized clinical trials. *Stat Med, 10*(5), 797-799. https://doi.org/10.1002/sim.4780100514

Birchwood, M., Gilbert, P., Gilbert, J., Trower, P., Meaden, A., Hay, J., Murray, E., & Miles, J. N. (2004, Nov). Interpersonal and role-related schema influence the relationship with the dominant 'voice' in schizophrenia: a comparison of three models. *Psychol Med, 34*(8), 1571-1580. https://doi.org/10.1017/s0033291704002636

Chandwick, P., Lees, S., & Birchwood, M. (2000, Sep). The revised Beliefs About Voices Questionnaire (BAVQ-R). *Br J Psychiatry, 177*, 229-232. https://doi.org/10.1192/bjp.177.3.229

Craig, T. K., Rus-Calafell, M., Ward, T., Leff, J. P., Huckvale, M., Howarth, E., Emsley, R., & Garety, P. A. (2018, Jan). AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *Lancet Psychiatry, 5*(1), 31-40. https://doi.org/10.1016/S2215-0366(17)30427-3

Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials* (2nd ed.). CRC Press.

Grant, S., Mayo-Wilson, E., Montgomery, P., Macdonald, G., Michie, S., Hopewell, S., Moher, D., & Group, o. b. o. t. C.-S. P. I. (2018, Jul 31). CONSORT-SPI 2018 Explanation and Elaboration: guidance for reporting social and psychological intervention trials. *Trials, 19*(1), 406. https://doi.org/10.1186/s13063-018-2735-z

Greenwood, K. E., Sweeney, A., Williams, S., Garety, P., Kuipers, E., Scott, J., & Peters, E. (2010, Jan). CHoice of Outcome In Cbt for psychosEs (CHOICE): the development of a new service user-led outcome measure of CBT for psychosis. *Schizophr Bull, 36*(1), 126-135. https://doi.org/10.1093/schbul/sbp117

Haddock, G., McCarron, J., Tarrier, N., & Faragher, E. B. (1999, Jul). Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). *Psychol Med, 29*(4), 879-889. https://doi.org/10.1017/s0033291799008661

ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. (1999, Aug 15). *Stat Med, 18*(15), 1905-1942. https://www.ncbi.nlm.nih.gov/pubmed/10532877

Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004; 141(10):781-788

Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of Multi-Arm Parallel-Group Randomized Trials: Extension of the CONSORT 2010 Statement. JAMA. 2019;321(16):1610–1620. doi:10.1001/jama.2019.3087

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010, Mar 23). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869. https://doi.org/10.1136/bmj.c869

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018, Jun). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry, 17*(2), 123-132. https://doi.org/10.1002/wps.20513

Orkin AM, Gill PJ, Ghersi D, Campbell L, Sugarman J, **Emsley R**, Steg PG, Weijer C, Simes J, Rombey T, Williams HC, Wittes J, Moher D, Richards DP, Kasamon Y, Getz K, Hopewell S, Dickersin K, Wu T, Ayala AP, Schulz KF, Calleja S, Boutron I, Ross JS, Golub RM, Khan KM, Mulrow C, Siegfried N, Heber J, Lee N, Kearney PR, Wanyenze RK, Hróbjartsson A, Williams R, Bhandari N, Jüni P, Chan AW; CONSERVE Group. (2021). Guidelines for Reporting Trial Protocols and Completed Trials Modified Due to the COVID-19 Pandemic and Other Extenuating Circumstances: The CONSERVE 2021 Statement. *JAMA* 326(3):257-265.

Phillips R, Sauzet O, Cornelius V. (2020). Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. *BMC Med Res Methodol.* 20(1):288. doi: 10.1186/s12874-020-01167-9

Shawyer, F., Ratcliff, K., Mackinnon, A., Farhall, J., Hayes, S. C., & Copolov, D. (2007, Jun). The voices acceptance and action scale (VAAS): Pilot data. *J Clin Psychol, 63*(6), 593-606. https://doi.org/10.1002/jclp.20366

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., & Stewart-Brown, S. (2007, Nov 27). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health Qual Life Outcomes, 5*, 63. https://doi.org/10.1186/1477-7525-5-63

White, I. R., & Thompson, S. G. (2005, Apr 15). Adjusting for partially missing baseline measurements in randomized trials. *Stat Med, 24*(7), 993-1007. https://doi.org/10.1002/sim.1981

# 7. Appendix 1: Table of measures and scoring rules

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| PSYRATS – Auditory Hallucinations<br><br>(AH or AHS) | The last week for the majority of the items | 11 items<br><br>The AHS items are Frequency, Duration, Location, Loudness, Origin, Negativity (Amount/Degree), Distress (Amount/ Intensity), Disruption, and Controllability. | For both the AHS and the DS, frequency-of-experience related items (eg, voices and/or thoughts are continuous) separated from distress related items in all studies.1,16–18 This suggests that separable phenomenological and etiological processes underlie duration and distress for both hallucinations and delusions. In other words, although duration and distress may share some underlying etiological processes (ie, the dimensions may be correlated), they are measurably distinct in some way | Each item scores between 0 to 4. Total Score (sum) of 11 items (range between 0 to 44. | Higher scores mean higher severity of the psychotic symptoms of hallucinations |
| PSYRATS – Auditory Hallucinations<br><br>Distress Dimension | The last week for the majority of the items | 5<br>negative content (2 items: amount and degree), voice distress (2 items: amount and intensity), and control item.<br><br>Items (6,7,8,9,11) | NA | Total Score (sum) of 5 items ranges between 0 to 20. | Higher score means higher distress associated with Voices (voice hallucination) |
| Characterisation of Voice Entity Complexity | | 11 Items in total.<br><br>10 binary items to measure the characterisation / personification of the voice.<br><br>The last items are a binary outcome endorsing the participants as more vs less highly characterised, based on the sum of the scores of the previous items (1 to 10). | NA | Total Score is sum of the items endorsed as present (1), and the range of the score is between 0 to 10.<br><br>Less than 7 = Less highly characterised<br><br>7 or higher = more highly characterised | Higher score means voices are more characterised<br><br>Less than 7 = Less highly characterised<br><br>7 or higher = more highly characterised |
| Hallucinations Remission Score | When was the last time you heard a distressing voice? | 1 | NA | 4. Today<br>3. In the last week<br>2. In the last two weeks<br>1. In the last month<br>0. Longer than one month ago | |
| Beliefs about Voices Revised (BAVQ-R) | In the past week | 35 | beliefs about voice power ('Omnipotence')<br>beliefs about voice intent ('Malevolence'/'Benevolence')<br><br>Beliefs:<br>1. malevolence (1, 4, 7, 10, 13, 16)<br>2. benevolence (2, 5, 8, 11, 14, 17) | Each item response choices are:<br>0. Disagree<br>1. Unsure<br>2. Agree slightly<br>3. Agree strongly<br><br>Score = subscale totals (sums) | Higher score means stronger beliefs about the positive or negative intent of the voice or belief about the voice power |

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| | | | 3. omnipotence (3, 6, 9, 12, 15, 18)<br><br>Emotional and Behavioural:<br>4. resistance (20, 22, 23, 25, 27, 28, 29, 30, 31)<br>5. engagement (19, 21, 24, 26, 32, 33, 34, 35) | Total score for each of the subscales of Malevolence, Benevolence, and omnipotence ranges between 0 to 18.<br><br>Resistance emotion ranges between 0 to 12<br><br>Resistance Behaviour ranges between 0 to 15<br><br>Engagement emotion score ranges between 0 to 12<br><br>Engagement behaviour Ranges between 0 to 12 | |
| Voices acceptance and action scale (VAAS) | | 31 items<br><br>Subscales:<br><br>Section A: items A1 to A12 (acceptance of, and disengagement from, auditory and command hallucinations).<br><br>Section B: items B1 to B19 (commitment to effective action rather than acting in relation to the voice). | The 31 items of VAAS are separated into Section A and B. This 31-item scale is divided into section A (i.e., 12 item stand-alone scale for general auditory hallucinations) and section B, referring specifically to command hallucinations.<br><br>Full scale (31 item)<br><br>Section A (as a stand-alone subscale) (A1 – A12)<br><br>Acceptance subscale (A1, A2, A4, A5, A6, A8, A9, A10, A11, B9, B11, B12, B15, B16, B17, B19)<br><br>Action subscale (A3, A7, A12, B1-B8, B10, B13, B14, B18) | Response choice:<br>1. Strongly Disagree<br>2. Disagree<br>3. Unsure Neutral<br>4. Agree<br>5. Strongly Agree<br><br>The full scale is sum of all items. | The participant is asked to rate their opinion from 1 'Strongly Disagree' to 5 'Strongly Agree', with higher scores meaning higher levels of acceptance and perception of acting according to one's valued life directions. |
| Voice Power Differential Scale (Partial) | The number which best describes how you feel in relation to your voice. | 1 | NA | 1. I am much more powerful than my voice<br>2. I am more powerful than my voice<br>3. We have about the same amount of power as each other<br>4. My voice is more powerful than me<br>5. My voice is much more powerful than me | |
| Beck Depression Inventory-II | Past two weeks including today | 21 | NA | Each item is scored on a scale of 0–3<br><br>Total score: sum of item scores | Total score of 0 - 13 is considered minimal range, 14 -19 is mild, 20 - 28 is moderate, |

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| | | | | | and<br>29 – 63 is severe. |
| Depression Anxiety and Stress Scales (DASS-21) | Over the past week | 21 | Subscales – Depression (items 3,5,10,13,16,17,21), Anxiety (Items 2,4,7,9,15,19,20), Stress (items 1,6,8,11,12,14,18). | The rating scale is as follows:<br>0. Did not apply to me at all<br>1. Applied to me to some degree, or some of the time<br>2. Applied to me to a considerable degree, or a good part of time<br>3. Applied to me very much, or most of the time<br><br>Multiply sum by 2 for 21 item version. Do not combine–- interpreted as separate scales. | Depression/ Anxiety/Stress<br>Normal 0-9 /0-7 /0-14<br>Mild 10-13 /8-9 /15-18<br>Moderate 14-20 /10-14 /19-25<br>Severe 21-27 /15-19 /26-33<br>Extremely Severe 28+ /20+ /34+ |
| Warwick-Edinburgh Mental Well-being Scale (WEMWBS) | Last 2 weeks | 14 items<br><br>Responses are made on a five-point scale ranging from 'none of the time' to 'all of the time'. | Total score:<br><br>total scale score is calculated by summing the 14 individual item scores. The minimum score is 14 and the maximum is 70. | All items are worded positively and address aspects of positive mental health. The scale is scored by summing responses to each item answered on a 1 to 5 Likert scale. The minimum scale score is 14 and the maximum is 70. | Higher scores mean a better mental wellbeing<br><br>The English Adults Norm for ages 16 years + for all in 2012:<br><br>All: 52.3 (0.16)<br>Men: 52.5 (0.22)<br>Women: 52.2 (0.20)<br><br>WEMWBS has not been validated as a screening tool to detect individuals with low mental wellbeing, and its psychometric properties mean that it is unlikely to be an efficient screening tool. Scores of 40 or less, however, put individuals in a high-risk category for mental illness. |
| Scale for the Assessment of Positive Symptoms (SAPS) | | 34 items | Hallucination:<br>Items 1 to 7<br>Delusion:<br>Items 8 to 20<br>Bizarre Behaviour:<br>Items 21 to 25<br>Positive Formal Thought Disorder:<br>Items 26 to 34 | SAPS measures positive symptoms on a 34 item, 6-point scale. (Scores 0 to 5)<br><br>SAPS Total (Composite) score = sum (of SAPS items 1-6, 8-19, 21-24, and 26-33) | The total SAPS score ranges from 0-150 and the global SAPS score of 0-20.<br><br>Higher scores indicate greater symptom severity<br><br>The |

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| | | | | SAPS (Global) Summary score = sum (of SAPS items 7, 20, 25, and 34, which include hallucinations, delusions, bizarre behaviour, and thought disorder global rating scores, respectively) | positive symptoms include hallucinations, delusions, bizarre behaviour, and positive formal thought disorder. |
| PSYRATS – Delusions | Last week | 6 items | Total score: all items<br><br>Distress: Item 4 and 5<br><br>Frequency: Items 1, 2, 3, 6 | 5-point response choice (scores 0 to 4)<br><br>Total score: Sum of scores | Higher scores mean higher severity of the psychotic symptoms of Delusion |
| Clinical Assessment Interview for Negative Symptoms (CAINS) | Past week | 13 items | 1. Motivation and Pleasure (MAP) Scale - 9 items<br><br>2. Expression (EXP) scale - 4 items | Items are rated on a 5-point (0-4) scale, with anchor points ranging from the symptom being absent or no impairment (0) to severe deficit (4)<br><br>MAP total score (sum)<br>EXP total score (sum) | It should be noted that lower "impaired" scores on several of the items may be within the range of normal variation in the general population. |
| Adapted CHOICE Short Form | Over the last week | 11 items | NA | Numeric rating scale 0 (worst) to 10 (best)<br><br>Sum of all items<br><br>Range 0 to 100 | |
| Adapted Mini-Trauma and Life Events (TALE) Checklist | Liftetime (<16yrs old and >16yrs old) | 4 items | 1. Verbal Abuse<br>2. Neglect<br>3. Physical Abuse<br>4. Sexual Abuse | Binary score for each item (Yes/No), once or multiple times, <16 or >16yrs old. | Endorsing an item indicates this form of abuse has been experienced, additional responses indicate whether this was on multiple occasions and in childhood, adulthood or both. |
| International Trauma Questionnaire | | 18 | PTSD (items P1 to P9)<br><br>CPTSD (items C1 to C9) | Dimensional scoring for PTSD and CPTSD.<br>Scores can be calculated for each PTSD and DSO symptom cluster and summed to produce PTSD and DSO scores.<br>PTSD<br>Sum of Likert scores for P1 and P2 = Re-experiencing in the here and now score (Re) | An individual can receive either a diagnosis of PTSD or CPTSD, not both. If a person meets the criteria for CPTSD, that person does not also receive a PTSD diagnosis.<br><br>For Diagnostic scoring for PTSD and CPTSD, please |

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| | | | | Sum of Likert scores for P3 and P4 = Avoidance score (Av)<br>Sum of Likert scores for P5 and P6 = Sense of current threat (Th)<br>PTSD score = Sum of Re, Av, and Th<br>DSO<br>Sum of Likert scores for C1 and C2 = Affective dysregulation (AD)<br>Sum of Likert scores for C3 and C4 = Negative self-concept (NSC)<br>Sum of Likert scores for C5 and C6 = Disturbances in relationships (DR)<br>DSO score = Sum of AD, NSC, and DR | check the manual. |
| TVAQ - Trauma Voice Associations Questionnaire (Partial) | | 3 items | NA | Binary responses:<br><br>0: No, not at all<br>1: Yes | |
| The Relationships Questionnaire (RQ) (Partial) | | 1 item<br><br>I am uncomfortable getting close to others. I want emotionally close relationships, but I find it difficult to trust others completely, or to depend on them. I worry that I will be hurt if I allow myself to become too close to others. | NA | 1: Disagree Strongly<br>2<br>3<br>4: Neutral/Mixed<br>5<br>6<br>7: Agree Strongly | |
| Client Service Receipt Inventory (CSRI) | | | | Will be reported and analysed by Health Economists | |
| EQ-5D-5L | | | | Based on algorithm<br><br>Will be reported and analysed by Health Economists | |
| COVID-19 Context Questionnaire | | 6 items | | Descriptor / No scoring | |
| Psychological and Psychosocial Interventions Log | | | | Descriptor / No scoring | |
| Antipsychotic Medications Log | | | | Descriptor / No scoring | |
| Withdrawal Form | | | | CONSORT / No scoring | |

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| Working Alliance Inventory - Short Form Revised * | | | | Descriptor / No scoring | |
| Therapy Attendance Log | | | | Descriptor / No scoring | |
| Adverse Events Log | | | | Safety / No scoring | |
| Indices of multiple deprivation (IMD)<br><br>This index is a measure of the extent to which a particular area suffers from a lack of resources, opportunities, and services relative to other areas. The index is based on a range of indicators such as income, employment, education, and health. | Baseline | 2<br>Rank and quintiles | NA | To assess levels of deprivation in different regions of England and Scotland, we will obtain a deprivation index for each region.<br><br>To obtain the deprivation index, we will gather data from the relevant government websites.<br><br>These indices are updated every few years and the most recent update was in 2019, and we will ensure we use the latest version available.<br><br>For Scotland, we will consult the Scottish Government's SIMD (Scottish Index of Multiple Deprivation). The most recent update was in 2020, and we will ensure we use the latest version available. | A higher quintile score indicates a lower level of deprivation, whereas a lower quintile score indicates a higher level of deprivation. For example, if a region is in the 5th quintile, it means that it is among the least deprived regions in the country, whereas a region in the 1st quintile is among the most deprived regions. |
| Ethnicity | Baseline | 1 item | NA | The Macro database records Ethnicity based on the following categories:<br><br>1.White<br>2.Black Caribbean<br>3.Black African<br>4.Black-Other<br>5.Indian<br>6.Pakistani<br>7.Chinese<br>8.Other | Ethnicity Recategorization:<br><br>Considering the substantial representation of the "Other" category and the relatively low number of individuals belonging to certain ethnicities (too low for meaningful statistical analysis), such as "Chinese". To address this, we recategorize participants' ethnicities as follows for |

| Measure in MACRO | Timescale | Number of items | Subscales | Scoring | Interpretation |
|---|---|---|---|---|---|
| | | | | | statistical analysis purposes:<br><br>1.White<br>2.Black or mixed Black<br>3.South Asian or Mixed South Asian<br>4.Other |

# 8. Appendix 2: Dummy tables for primary publication

**Table 1: Baseline Characteristics – Demographics**

| Measure | | Brief AVATAR | Extended AVATAR | TAU | Total |
|---|---|---|---|---|---|
| Age – years (mean (SD)) | | | | | |
| Gender identity – self reported (n (%)) | Female | | | | |
| | Male | | | | |
| | Not reported/disclosed | | | | |
| Ethnicity (n (%)) | White | | | | |
| | Black or mixed Black | | | | |
| | South Asian or Mixed South Asian | | | | |
| | Other | | | | |
| Marital status (n (%)) | Single | | | | |
| | In a Relationship | | | | |
| | Cohabiting | | | | |
| | Married or Civil Partnership | | | | |
| | Divorced | | | | |
| | Widowed | | | | |
| Primary living situation (n (%)) | Living alone (+/- children) | | | | |
| | Living with husband/wife (+/- children) | | | | |
| | Living together as a couple (+/ - children) | | | | |
| | Living with parents | | | | |
| | Living with other relatives | | | | |
| | Living with others | | | | |
| Living with, as best described (n (%)) | Alone | | | | |
| | Partner/spouse without children | | | | |
| | Partner/spouse with children | | | | |
| | Children only | | | | |
| | Parents/other family members | | | | |
| | One or several friends | | | | |
| | Housemates/lodgers/tenants (not friends) | | | | |
| | Other | | | | |
| Age when first started hearing voices – Years (mean (SD)) | | | | | |
| Duration of contact with MH services – Years (mean (SD)) | | | | | |
| Highest level of schooling (n (%)) | Primary school | | | | |
| | Secondary no exams qualifications | | | | |
| | Secondary (O/ CSE equivalent) | | | | |
| | Secondary (A level equivalent) | | | | |
| | Vocational Education/ college | | | | |
| | University degree/ professional qualification | | | | |
| Main current working status (n (%)) | Unemployed | | | | |
| | Employed full-time | | | | |
| | Employed part-time | | | | |
| | Self-employed | | | | |
| | Retired | | | | |
| | Student | | | | |
| | Housewife/husband | | | | |

**Table 2: Primary and secondary outcomes**

| | Unadjusted, Mean (SD) | | | Brief vs TAU | Extended vs TAU |
|---|---|---|---|---|---|
| Outcome | Brief AVATAR n= XX | Extended AVATAR n= XX | TAU n= XX | Adjusted Difference (SE); p-value (95% CI): Cohen's d | Adjusted Difference (SE); p-value (95% CI): Cohen's d |
| Baseline | | | | - | - |
| 16 Weeks | | | | | |
| 28 Weeks | | | | | |

# 9. Appendix 3: Example analysis code

Data will be in long format with two rows for each participant, one for 16- and 28-week time points.

Example variable names:

- pid: participant identifier
- treat: randomised arm of the participant
- timepoint: follow-up timepoint (categorical)
- baseline: baseline measure of the outcome
- centre: centre (used as a stratification factor)
- vc: voice characterisation (used as a stratification factor)
- outcome: outcome measure
- therapist: therapist identifier

Example analysis code:

*Model for continuous outcomes analysed using mixed effect model:

```
mixed outcome i.treat##i.timepoint baseline vc i.centre ||
therapist: treat, nocons || pid:

margins treat, at(timepoint==16) pwcompare(effects)

margins treat, at(timepoint==28) pwcompare(effects)
```

*Model for binary outcomes analysed using mixed effect model:

```
melogit outcome i.treat##i.timepoint baseline vc i.centre ||
therapist: treat, nocons || pid:

melogit, or

melogit outcome i.treat##ib28.timepoint baseline vc i.centre ||
therapist: treat, nocons || pid:
```