Statistical Analysis Plan (SAP)

Study Name:	Mind VR Trial
Study Title:	Hybrid Type 2 Implementation-Effectiveness Trial of Mind VR ('Mind VR Trial')
Trial Registration:	ClinicalTrials.gov; NCT06894836
Ethics Approval Number:	HREC 2024.31083
Sponsor	Orygen 35 Poplar Rd Parkville, VIC, 3058, AUSTRALIA
SAP Version:	1.0, References Protocol Version 4.0
Coordinating principals	Dr Imogen Bell and A/Prof Jennifer Nicholas
Date Finalised:	9/9/26
Prepared by:	Orygen Data Science & Analytical Methods team, Dr Imogen Bell and A/Prof Jennifer Nicholas
Approved by:	Dr Imogen Bell and A/Prof Jennifer Nicholas

Change History

Version	Date	Description of Change	Author
1.0	9/9/26	Initial SAP version	Orygen Data Science & Analytical Methods team, Dr Imogen Bell and A/Prof Jennifer Nicholas

Statement of Compliance

This SAP is consistent with the approved trial protocol and adheres to ICH E9 (R1), SPIRIT 2025, CONSORT 2025 and its relevant extensions, including for pragmatic, hybrid, and implementation trials. It also complies with ethical standards such as the Declaration of Helsinki.

Table of Contents

List of Abbreviations and definitions of terms

Abbreviation	Definition
AE	Adverse Event
AFQ-Y	Avoidance and Fusion Questionnaire for Youth
CFQ	Cognitive Fusion Questionnaire
CI	Confidence Interval
CTCAE	Common Terminology Criteria for Adverse Events
FIM	Feasibility of Implementation Measure
GAD-7	Generalised Anxiety Disorder 7-item scale
GCP	Good Clinical Practice
HREC	Human Research Ethics Committee
IQR	Interquartile Range
ITT	Intention-to-Treat
K10	Kessler Psychological Distress Scale
LOCF	Last Observation Carried Forward
MI	Multiple Imputation
Mind VR	Trial name (VR intervention)
MPoD-s	Metacognitive Processes of Decentering – State
MPoD-t	Metacognitive Processes of Decentering – Trait
PHQ-8	Patient Health Questionnaire – 8 items
PP	Per-Protocol
PTQ	Perseverative Thinking Questionnaire
ReQoL	Recovering Quality of Life Scale
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Standard Deviation
TFAQ	Theoretical Framework of Acceptability Questionnaire
VIMS	Visually Induced Motion Sickness questionnaire
VR	Virtual Reality
WAI-SR	Working Alliance Inventory – Short Revised
WEMWBS	Warwick-Edinburgh Mental Well-Being Scale
YP	Young People

Introduction

This Statistical Analysis Plan (SAP) describes the pre-specified statistical methods to be used in the analysis of data from the Mind VR Trial: a single-arm, hybrid Type 2 implementation-effectiveness trial evaluating a virtual reality (VR)-based psychological intervention for depression and anxiety among young people aged 12–25 years attending participating headspace services in Australia.

This trial is primarily designed to assess the feasibility and acceptability of delivering the MIND VR intervention within routine youth mental health settings, with preliminary effectiveness as a secondary aim. Given the single-arm design, effectiveness analyses are within-group pre-post comparisons only and should be interpreted as hypothesis-generating estimates to inform a future randomised controlled trial, rather than as evidence of efficacy. There is no randomisation, control group, or blinding in this trial. The trial is registered at ClinicalTrials.gov: https://clinicaltrials.gov/study/NCT06894836.

The following change was made after trial registration:

- Increased young person sample size target from 75 to 150. Sample size was increased due to targets being met within the first half of the recruitment window, to honor the intervention access window commitments (12month) made with partner services. Increased sample provides increased power for secondary analysis.

This SAP was developed prior to database lock. Any amendments after this point will be documented, dated, version-controlled, and the rationale recorded. This document should be read alongside the trial protocol (Version 4.0, 15 December 2025, HREC 2024.31083). Any analysis performed, not prospectively defined in this document, will be labelled as post hoc and exploratory.

1. Overview and objectives of study design

This is a single-arm effectiveness-implementation trial. All enrolled young people receive the Mind VR intervention. There is no concurrent control group within the trial design. Pre-post comparisons within participants serve as the primary effectiveness analytic approach.

The overall aims of this study were to evaluate the feasibility, acceptability, implementation, and preliminary effectiveness of the MIND intervention within routine youth mental health care. Specifically:

Primary objectives were to:

a) Assess the feasibility of the study protocol to inform the design and conduct of a future full-scale trial of the Mind VR intervention.
b) Evaluate the feasibility and acceptability of implementing the Mind VR intervention within routine youth mental health care settings.
c) Evaluate the safety of the Mind VR intervention when delivered in routine youth mental health care settings.

Secondary objectives were to:

d) Explore the preliminary effectiveness of the Mind VR intervention on improving the primary clinical outcomes of depression and anxiety
e) Explore the preliminary effectiveness of the Mind VR intervention on improving the secondary clinical outcomes of wellbeing, quality of life, and psychological distress

f) Explore the preliminary effectiveness of the Mind VR intervention on improving the mechanistic outcomes of decentering, repetitive negative thinking, psychological flexibility, and cognitive defusion

Additional exploratory aims were to:

g) Examine whether the potential therapeutic impact of the Mind VR intervention on depression and anxiety is mediated by changes in underlying mechanistic targets (i.e. decentering, repetitive negative thinking, psychological flexibility, and cognitive defusion).
h) Examine predictors of the implementation and effectiveness of the Mind VR intervention
i) Identify barriers and facilitators to implementing the MIND intervention in routine care settings, including implementation
j) Examine young people's experiences of the Mind VR intervention in routine care

1.1 Study Timeline and Assessment Schedule

Young people complete assessments at three timepoints across a 16-week period. Each assessment session is expected to take approximately 45 minutes. An optional qualitative interview will also be offered to 20-30 young people at or following the post-intervention assessment to explore their experiences of receiving VR-based intervention (see section 4.3).

Timepoint	Window	Key Assessments
Baseline (T0)	Week 0–1	Demographics, PHQ-8, GAD-7, K10, PTQ, MPoD-t, CFQ, AFQ-Y, WEMWBS, ReQoL, CEQ
Mind VR intervention	Week 1–10 (±2 weeks)	4–5 Mind VR sessions delivered by clinician; MPoD-s and VR experience items collected within each session; adverse events monitored throughout
Post-intervention (T1)	Week 10–12 (±1 week)	PHQ-8, GAD-7, K10, PTQ, MPoD-t, CFQ, AFQ-Y, WEMWBS, ReQoL, TFAQ, WAI-SR, VIMS Subset qualitative interviews
Follow-up (T2)	Week 14–16 (±1 week)	PHQ-8, GAD-7, K10, PTQ, MPoD-t, CFQ, AFQ-Y, WEMWBS, ReQoL

Clinicians and service staff are assessed over a substantially longer period, 16 months across the full trial, to capture implementation processes, workforce outcomes, and service-level feasibility indicators that unfold over time. This is distinct from the 16-week young person assessment window described above.

Timepoint	Window	Key Activities
Pre-intervention (Month 0)	Month 0	Informed consent, demographics, prior digital tool use, implementation leadership and climate scales, confidence and competence questionnaire, half-day Mind VR training (clinicians only)
Intervention delivery period	Month 1–15	Delivery of Mind VR sessions (clinicians); clinician post-session surveys; monthly group supervision (clinicians); research meetings and touchpoints; adverse event monitoring

Post-intervention (Month 16)	Month 16 (±4 weeks)	TFAQ, FIM, confidence and competence questionnaire (clinicians); qualitative interview or focus group (clinicians and service staff)

2. Endpoints and Estimands

This section specifies the precise technical definition of each endpoint. Endpoints are organised into primary (feasibility, acceptability, and safety) and secondary (preliminary effectiveness objectives) domains. For primary objectives the corresponding estimand, that is, exactly what quantity will be estimated in each analysis, is outlined. For secondary and exploratory outcomes, these are outlined in the general statistical analysis section (see 6.1 General Statistical Considerations).

2.1 Primary objectives — Feasibility, acceptability and safety

Domain	Endpoint	Estimand
Protocol feasibility	Recruitment rate, retention rate (T1 and T2), session attendance, assessment completion rate, trial-related adverse event rate	Proportion reported with 95% CI; interpreted descriptively in the context of published feasibility benchmarks.
Implementation feasibility and acceptability	Feasibility: FIM score; clinician adoption rate; number of sessions delivered; proportion of sessions using Mind VR Acceptability: Recommendation item (1-6 scale) assessed separately for young people and clinicians. TFAQ mean score (7-item, 1–5 scale) assessed separately for young people, clinicians, and service staff	Feasibility: Mean FIM score with 95% CI; proportion of trained clinicians delivering Mind VR to at least one young person Acceptability: Number and proportion endorsing 'agree' and 'strongly agree' Mean TFAQ score with 95% CI per group; interpreted descriptively with reference to the scale range and midpoint
Safety	The number of serious adverse events specifically related to the intervention or trial as determined during assessments (e.g. symptom exacerbation),	Number reported over the trial period (total, trial related, and intervention related)

Commented [CS1]: Maybe clarify participant group (young people or clinicians)

2.2 Secondary objectives — Preliminary Effectiveness

Outcome	Measure	Timepoints
Primary clinical outcomes		
Depression	PHQ-8 (0–24)	T0, T1, T2
Anxiety	GAD-7 (0–21)	T0, T1, T2
Secondary clinical outcomes		
Psychological distress	K10 (10–50)	T0, T1, T2
General wellbeing	WEMWBS	T0, T1, T2
Quality of life	ReQoL	T0, T1, T2
Mechanistic outcomes		
Repetitive negative thinking	PTQ	T0, T1, T2
Trait decentering	MPoD-t	T0, T1, T2
Cognitive fusion	CFQ	T0, T1, T2
Psychological inflexibility	AFQ-Y	T0, T1, T2

2.3 Exploratory objectives

Domain	Endpoint
Predictors of outcome	Baseline youth and clinician sociodemographic and clinical variables as predictors of PHQ-8/ GAD-7 or K10 (T1 & T2)
Mediation	MPoD-t and PTQ change (T0 to T1) and session change in MPoD-s as mediator of PHQ-8/ GAD-7 or K10 change (T1/T2).
Implementation determinants	Barriers and facilitators to Mind VR implementation identified from clinician and service staff interviews (16 weeks) and process data (across intervention period).
Young people's experiences	Experiences of receiving VR-based intervention in routine care (T1)

3. Sample Size and Power Calculations

The target sample size of 75 young people and 20 clinicians and service staff (combined) was determined by the primary feasibility, acceptability, and safety aims of the trial, not by a formal power calculation. This sample size enables stable estimates of proportions (e.g. recruitment rate, acceptability) with adequate precision and allows thorough assessment of implementation factors across a broad range of services.

Revised sample size of 150 young people allows for within-group pre-post effectiveness comparisons, at 80% power (alpha = 0.05, two-tailed paired t-test) to detect within-group effect sizes from small (d = 0.23) upwards.

Commented [JN2]: to revise based on answer to above

Commented [IB3]: I would prefer a framing that this allows a determination (i.e. t test with sig value) of the within-group changes as a result of the intervention exposure with follow up trial needed to confirm effectiveness against a control group. This trial is large enough for us to have certaintly over whether the intervention worekd for this group.

4. Analysis Populations

Analysis populations are therefore defined functionally in the table below and referenced throughout Section 5.

Population	Definition	Used For
Enrolled Population (5.1)	All consented young people completing at least the baseline assessment (T0). Participants who withdraw after baseline are included with available data up to withdrawal.	Primary population for all effectiveness analyses
Completer Population (5.2)	Enrolled population participants completing at least three of the four to five Mind VR sessions and the post-intervention assessment (T1).	Sensitivity analyses for effectiveness outcomes to assess consistency among those receiving an adequate intervention dose
Implementation Evaluation Population (5.3)	All consented clinicians who complete the pre-intervention training session, and all consented service staff.	All clinician and service staff analyses

Commented [IB4]: Nice

Protocol deviations affecting eligibility or intervention delivery will be reviewed prior to analysis and documented on the trial deviation log. The impact of major deviations on the enrolled and completer populations will be reported in a deviation summary table.

5. Statistical Analysis

5.1 General Statistical Considerations

Statistical analyses will be reported using summary tables, figures, and data listings. Continuous variables will be summarised with counts, means, standard deviations, medians, confidence intervals, minimums, and maximums. Categorical variables will be summarised by counts and by percentage of participants. Two-tailed tests will be used throughout with a significance level of alpha = 0.05, except where otherwise stated.

Prior to analysis, data will be screened for implausible values and range violations against the known scoring ranges for each measure. Any out-of-range values will be flagged and queried against source data in REDCap before being treated as missing if unresolvable. Outliers will be investigated on a case-by-case basis but will not be excluded unless there is clear evidence of data entry error; the reason for any exclusion will be documented. Duplicate records will be checked and resolved prior to analysis. A data cleaning log will be maintained documenting all decisions made during screening.

No interim analyses are planned. Feasibility indicators may be monitored periodically by the investigator team for operational purposes but this does not constitute a formal interim analysis and will not influence trial continuation decisions.

Analysis populations are defined in Section 5 and referenced for each analysis below.

5.2 Descriptive Statistics

Baseline characteristics may be summarised for the Enrolled Population (Section 5.1) and the Implementation Evaluation Population (Section 5.3) as follows:

- Continuous variables: mean, standard deviation (SD), median, and number of non-missing observations. The median, interquartile range (IQR), and range (minimum, maximum) will also be used in the case of highly skewed data.
- Categorical variables: frequency counts and percentages

Summaries will be presented separately for young people, clinicians, and service staff. No formal statistical testing will be performed on baseline characteristics given the single-arm design.

5.3 Primary Analysis

5.3.1 Feasibility Outcomes

Conducted in the Enrolled Population (Section 5.1) and Implementation Evaluation Population (Section 5.3). Each feasibility indicator will be reported as a proportion or mean with standard deviation as appropriate. Findings will be interpreted descriptively by the investigator team in the context of published feasibility benchmarks and used to inform recommendations for a future trial. No formal pass/fail criteria are pre-specified; the investigator team will make a collective judgement about trial feasibility based on the pattern of findings across indicators.

The following feasibility indicators are specified in the trial protocol:

Feasibility Indicator	Definition	Reported As
Recruitment and consent rates	Proportion of eligible young people, clinicians, and service staff who consent to participate, reported separately for each group	Proportion with 95% CI per group
Dropout rates	Proportion of enrolled participants who withdraw or are discontinued prior to completing the trial, reported separately for young people, clinicians, and service staff	Proportion with 95% CI per group
Assessment completion	Proportion of enrolled young people completing assessments at each timepoint (T0, T1, T2)	Proportion with 95% CI per timepoint
Session attendance	Mean number of Mind VR sessions completed per participant; proportion completing at least three sessions	Mean and SD; proportion
Clinician adoption	Proportion of trained clinicians who deliver Mind VR to at least one young person during the trial period	Proportion with 95% CI

5.3.2 Acceptability Outcomes

Conducted in the Implementation Evaluation Population (Section 5.3) and Enrolled Population (Section 5.1). Acceptability will be assessed using the TFAQ, scored as a mean of seven items on a five-point Likert scale. Mean scores with standard deviations and 95% confidence intervals will be reported separately for young people, clinicians, and service staff. An a priori acceptability benchmark of a mean TFAQ score of ≥3.5 (out of 5) will be used to indicate adequate acceptability. Acceptability is also assessed using the proportion of young people and clinicians who agree they would recommend Mind VR to a peer.

5.3.3 Safety Outcomes

Trial- and intervention-related adverse events	Number and proportion of participants experiencing at least one trial- or intervention-related adverse event	Count and proportion; SAEs listed individually

Conducted in the Enrolled Population (Section 5.1). All adverse events (AEs) and serious adverse events (SAEs) will be listed individually and summarised by:

- Frequency and proportion of participants experiencing at least one AE or SAE
- Causality classification (related, possibly related, not related)
- Severity grade (CTCAE v5.0)
- Resolution status at end of trial

No formal statistical testing will be performed on safety outcomes. All SAEs will be listed individually regardless of causality.

5.4 Secondary Analyses

5.4.1 Within-Group Pre-Post Effectiveness Analyses

Conducted in the Enrolled Population (Section 5.1) as the primary analysis. All enrolled participants with data at a given timepoint contribute to mean scores at that timepoint. Mean change scores and paired t-tests are based on participants with complete data at both timepoints; the N contributing to each estimate will be reported separately. Sensitivity analyses will be repeated in the Completer Population (Section 5.2) to assess consistency of findings among those receiving an adequate intervention dose. For each secondary continuous effectiveness outcome (PHQ-8, GAD-7, K10, PTQ, MPoD-t, CFQ, AFQ-Y, WEMWBS, ReQoL), the following will be reported:

- Mean score at T0, T1, and T2 with standard deviations and 95% confidence intervals
- Mean change from T0 to T1 and from T0 to T2
- Paired samples t-test comparing T0 to T1 scores (two-tailed, alpha = 0.05)
- Cohen's d effect size with 95% confidence interval for T0 to T1 change, calculated as mean change divided by SD of baseline scores
- Paired samples t-test and Cohen's d for T0 to T2 change

If data are non-normally distributed (assessed using Shapiro-Wilk test), Wilcoxon signed-rank tests will be used in place of paired t-tests and reported accordingly. No adjustment for multiple comparisons will be made given the exploratory nature of secondary and exploratory analyses.

5.5 If a historical comparator group is secured, additional effectiveness analyses will compare PHQ-8, GAD-7 and/or K10 change scores from baseline to equivalent follow-up timepoints (T1/T2) between the two groups. Handling of Missing Data

Available case analysis will be used as the primary approach, utilising all participants with complete data for each specific analysis. This approach is appropriate given the feasibility nature of the trial and the expectation of informative missingness.

The extent and pattern of missing data will be described for all primary and secondary outcomes at each timepoint. Baseline variables will be compared between completers and non-completers at T1 to assess whether missingness is associated with observed characteristics.

Sensitivity analyses using multiple imputation (MI) will be conducted for the primary effectiveness outcomes (PHQ-8, GAD-7) if missing data at T1 exceeds 10% of the enrolled population. MI will be conducted using the mice package in R or equivalent alternative, with 50 imputed datasets and predictive mean matching for continuous outcomes. Imputation models will include all baseline covariates, demographic variables, and auxiliary variables associated with missingness. Results will be combined using Rubin's rules. MI results will be compared to available case analysis results; material differences will be discussed in the trial report.

Commented [GU5]: This doesn't line up with the objectives because the WEMWBS etc are presented as exploratory - perhaps we should include anything related to effectivenss (i.e. the cohen's d analyses) as secondary outcomes, starting with PHQ and GAD7? Then exploratory can be all the mediation and qually stuff?

Commented [IB6]: Ah I see - it's all laid out here. My prior comments call for this and it looks approrpiate so can mostly dismiss any call outs above re this unless edits seem approriate.

Commented [CS7]: Doesn't the primary analysis need to be in the Completer Population?

Commented [ST8R7]: I'm following the ITT principle, so using all enrolled participants with some data? Or is per-protocol preferred for feasibility trials? what do you think?

Commented [CS9R7]: It's tricky when there is no comparator group and only have within group comparisons - in practice it will end up being per protocol anyway won't it for simple mean change? but I guess could do this as 2 time point simple mixed effects so can use the baseline, even though ultimately the change will still be based on those with 2 time points? I think we need to check other trial examples

Commented [SMT10R7]: Good point. I think I will add the mixed effects model

Commented [IB11]: Looks good!

Last observation carried forward (LOCF) will not be used as it is not recommended for clinical trial analysis.