A Pragmatic Randomized Multiple Baseline Trial Evaluating Knowledge Insight Tools (KIT), a Cognitive Behavioural Therapy-informed School-based Counselling Intervention for Secondary School-aged Children and Young People with Anxiety and Mood Difficulties in the UK

Statistical Analysis Plans

NCT ID: Not yet assignment

Document Date: 01/08/2023

Statistical methods

Statistical methods for primary and secondary outcomes {20a}

Primary Outcome

The primary outcome is YP-CORE scores, assessed weekly throughout the baseline and intervention phases. We will analyze within-person and between-person change in YP-CORE scores using descriptive and inferential statistics suitable for multiple baseline designs.

Descriptive methods include summary statistics (e.g., measures of central tendency and dispersion) and visual analysis, where we will inspect time-series graphs of session-by-session scores for stability during the baseline phase, negative trends in the intervention phase, variability within and between phases, and the general degree of overlap between baseline and intervention scores using range lines (Lane & Gast, 2014).

Inferential methods include piecewise multilevel growth models to estimate within-person and between-person differences in average scores (i.e. levels) and the direction/rate of change in scores (i.e. slopes) between the baseline and intervention phases (e.g., Reut et al., 2023). Models will include at least two levels, with repeated observations of YP-CORE scores at level 1 nested within each young person at level 2. We will include additional levels if there is significant variation associated with specific therapists, schools, or regions. We will estimate fixed effects, which include intercepts and slopes averaged across young people for both the baseline and intervention phases (see below for piecewise coding of phases). We will also estimate random effects, which include between-person variation in intercepts and slopes for the baseline and intervention phases.

Two dummy-coded variables each representing an intercept for the baseline phase and intervention phase will be included in the model at level one. Each intercept variable will be coded to reflect the individually-varying break-points between baseline and intervention phases. Take, for instance, a young person with a three-week baseline phase followed by a four-week intervention phase. The baseline intercept variable for this young person would be coded as 1, 1, 1, 0, 0, 0, 0, and the intervention intercept variable as 0, 0, 0, 1, 1, 1, 1. This coding will produce estimates of the mean baseline and intervention YP-CORE scores averaged across young people with individually varying phase lengths. Baseline and intervention intercepts can then be compared using a two-tailed Wald test to estimate a treatment effect; mean YP-CORE scores during the intervention phase are predicted to be significantly lower than mean YP-CORE scores during the baseline phase.

Two more dummy-coded variables will be included at level one to estimate the slopes for the baseline and intervention phases. The baseline slope variable will be coded for the weeks until the first intervention session, whilst the intervention slope variable will be coded for the weeks since the first intervention session. Continuing the example above of a young person with a three-week baseline phase and four-week intervention phase, the baseline slope variable would be coded as -3, -2, -1, 0, 0, 0, 0, and the intervention slope variable would be coded as 0, 0, 0, 0, 1, 2, 3. Note that the first intervention session is coded as a ‘0’ across baseline and intervention time variables because it serves as the pre-treatment baseline (since the YP-CORE will be administered at the beginning of the first intervention session, before the intervention starts). This coding will produce estimates of the mean slopes during the baseline and intervention phases; a baseline slope not significantly different from zero, plus a significant negative intervention slope, would indicate a treatment effect. Furthermore, random intercepts and slopes will allow us to quantify stability and intraindividual variability in YP-CORE trajectories that would be observed in the visual analysis.

Between-person covariates, including the randomly allocated start week, absolute time in weeks since the first participant was recruited, age, sex, ethnicity, free school-meal eligibility, and special education needs status, will be added to level two of the model to control for the confounding effects of method variables and demographics. We will also include interactions within time variables to investigate non-linear trajectories in YP-CORE scores, and interactions between time and demographic variables to investigate whether demographics moderate the treatment effect.

Multilevel models tend to be robust to deviations from their parametric assumptions (Maas & Hox, 2004). However, we will explore non-parametric alternatives, like Simulation Modelling Analysis (e.g., Dunn et al., 2019), if our data heavily violate these assumptions.

We will also estimate treatment effects and their effect size in line with standard outcomes reported in randomized controlled trials. For the treatment effect, we will use randomization tests, which, unlike more common tests of repeated measures like paired *t*-tests, do not make distributional assumptions or assume homogeneous variances (Bulté & Onghena, 2009). This is because the sampling distribution is based on random permutations of the observed data (i.e. a randomization distribution). We will also calculate a measure of effect size: Shadish et al.'s (2014) adapted *d*-statistic for single-case designs or Tau-U if there are trends in the baseline phase (Manolov, Losada, Chacón-Moscoso, & Sanduvete-Chaves, 2016).

Secondary Outcomes

Secondary outcomes include the RCADS, KIT Fidelity Checklist, and Implementation Survey. The RCADS will be completed at the start, middle and end of the intervention. We will use measures of central tendency and dispersion (e.g., means and standard deviations) to report group-level differences in scores between the baseline and intervention phases. Furthermore, we will evaluate the statistical significance of the difference in group means using parametric tests (e.g., repeated *t*-test), non-parametric (e.g., Wilcoxon Signed-rank test) tests, and regression models controlling for covariates, and quantify the size of the difference using the standardized mean difference (i.e. Cohen's *d*).

We will also calculate Jacobson and Truax's (1991) clinically significant and reliable change indices for the RCADS. Clinically significant change tells us the proportion of young people who start the intervention in the clinical range and finish the intervention in the non-clinical or recovery range. There are different methods for calculating the thresholds for clinical and non-clinical/recovery ranges. We will use the RCADS' established clinical norms to determine the clinical range (e.g., *T* scores

> 69) and non-clinical range (e.g., *T* scores < 65; Chorpita, Moffitt, & Gray, 2005).

Reliable change refers to whether the changes observed in scale scores (both improvements and deteriorations) over the course of an intervention are greater than the change expected due to measurement error alone (Jacobson & Truax, 1991). We will calculate the proportion of young people who demonstrate reliable improvement, reliable deterioration, and no reliable change in the RCADS after receiving KIT. We will calculate reliable change indices from the sample data using Jacobson and Truax's (1991) formula for reliable change. We will also compare our findings to previous reports of reliable change indices for the RCADS (Edbrooke-Childs, Wolpert, Zamperoni, Napoleone, & Bear, 2018). Finally, we will determine the proportion of young people who showed both clinically significant and reliable change, since one can show reliable improvement without it being clinically significant, and vice versa.

The Fidelity Checklist will be scored in various ways. Traditionally, clinical researchers calculate an index of the proportion of practitioners demonstrating a prespecified level of treatment fidelity. For each KIT intervention with a young person, we will calculate proportions of completeness for each KIT phase across sessions and a total completeness score. That is, we will score the presence of each item on the checklist as a '1' and calculate the proportion of items scored within each phase. Only the first instance of the item throughout the intervention will be counted, not the frequency of the item across sessions. We will also create a total score by summing the subscale scores together. We will then determine a threshold for a 'complete' KIT intervention (e.g., scoring at least 75% of items within each subscale, across all subscales), and determine the proportion of KIT interventions meeting this threshold. This will allow us to conduct sensitivity analyses with complete KIT interventions only vs. incomplete KIT interventions. Since fidelity data will be assessed over time, we can also control for session-by-session fidelity scores in the multilevel growth models of YP-CORE trajectories or analyze patterns/profiles of scores on the fidelity checklists that predict better outcomes.

As for the Implementation Survey, we will use measures of central tendency and dispersion to get an overall sense of how practitioners experienced implementing

KIT. We will also examine the distribution of responses for each item to determine what practitioners favoured most/least about KIT. Some questions, e.g., confidence around delivering KIT, can be used as moderators in the multilevel growth curve models. Depending on the number and richness of responses, we will analyze free-text responses with thematic analysis to draw out practitioners' views of the advantages and barriers to delivering KIT.

Interim analyses {21b}

We will estimate the sample's conditional power after recruiting 50% of the target (*N* = 30) or if we have not met our recruitment target by the first deadline (February 2024), whichever comes first. If the conditional power estimate indicates that we need more participants than the initial target of 60, we will extend recruitment to a second wave with an adjusted alpha level for subsequent analyses.

Stopping guidelines will be based on an O'Brien-Fleming-like alpha-spending function (Ciolino, Kaizer, & Bonner, 2023). We will use the weights outlined by O'Brien-Fleming for our sample size re-estimation analysis pre-specified above (e.g., after we have recruited 50% of our sample), but will pair this with a spending function that is based on the amount of data collected at the time and does not require a pre-specified time-point (so we can re-estimate the required sample size if we have not reached at least 50% of our sample by February 2024).

We will also conduct safety checks by examining trends in the YP-CORE data, particularly if we encounter multiple adverse events. Safety checks will involve visual analysis of YP-CORE scores for each participant during the baseline and intervention phases to assess for any consistent, statistically reliable and clinically significant negative trends. Reliable and clinically significant deterioration in YP-CORE scores that can be linked to the trial protocol will be raised with Adverse Events Oversight Group, who will decide on whether to discontinue the trial.

Methods for additional analyses (e.g. subgroup analyses) {20b}

We will examine (and control for) group differences in demographics and random baseline period by including these variables as covariates in the multilevel growth models. We will also conduct sensitivity analyses that exclude young people whose KIT interventions did not meet fidelity requirements as well as young people who did not start the intervention on their allocated start week, either because they/their practitioner did not attend the intervention session or the first intervention session fell on a school closure and was moved to a another date.

Methods in analysis to handle protocol non-adherence and any statistical methods to handle missing data {20c}

We will analyze all available outcomes on an intention-to-treat basis, including young people who did not start the intervention on the randomly allocated start week (e.g., due to school closures or pupil/practitioner absences) or complete the intervention. Missing data patterns will partly be caused by different baseline and intervention lengths; however, multilevel models can handle missing data due to varying treatment lengths (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009). We will assess whether clinical variables (e.g., baseline symptom scores), the randomly allocated start week, and demographic variables, predict dropout. Depending on proportion of missingness (e.g., >30% of observations) and its implied mechanisms, we will run a sensitivity analysis whilst handling missing data with methods such as multiple imputation.