Emotion Regulation in Children (ERiC):
Statistical analysis plan

Version history log

Version	Date	Details of Change
0.1	10 Feb 2026	---
1	5 Mar 2026	Changes in response to TSC comments: - Added target sample size (p 2) - Updated reference on fidelity scale (p 6) - Added explanation of why age & baseline SDQ are included as covariates in outcome models (p 10) - Added explanation of how the primary outcome analysis is valid under MAR and provided reference (p11) - Clarified that conditional mean imputation is a planned procedure for item missingness only, not other types of missing values (p 12) - Added multiple imputation under not missing at random assumptions as a potential sensitivity analysis, in line with the trial protocol (p 12) - Added brief explanation of the role of mediation analysis within the trial (p 13)
2	12 Aug 2026	Changed calculation of summary scores for the EAQ subscales to mean of items, instead of sum of items. This ensures that results are comparable to most other published studies using the EAQ. The change was made after the (blinded) data had been shared with the statisticians. However, the change will make no difference to the substantive interpretation, as the mean of the items is a linear transformation of the sum.

1. Study Summary

For full details see the trial protocol, published as: Midgley, N., Mortimer, R., Carter, M., Casey, P., Coffman, L., Edbrooke-Childs, J., ... & McFarquhar, T. (2023). Emotion regulation in children (ERiC): A protocol for a randomised clinical trial to evaluate the clinical and cost effectiveness of Mentalization Based Treatment (MBT) vs Treatment as Usual for school-age children with mixed emotional and behavioural difficulties. *Plos one*, *18*(8), e0289503.

Title	Emotion Regulation in Children
Short title	ERiC
Chief Investigator	Prof. Nick Midgley
Statistician	Dr Peter Martin
Design	An individually randomised controlled superiority trial with children aged 6–12. Participants will be randomised to receive either the intervention (Mentalization Based Therapy, MBT) or usual care.
Primary objective	To evaluate the clinical and cost-effectiveness of MBT compared to treatment as usual for school age children with emotional and behavioural difficulties.
Population	Children aged 6–12 referred to CAMHS with mixed mental health problems (emotional and behavioural) as primary problem
Sample size	320 children (recruitment target)
Study Type	Interventional randomised controlled trial
IRAS Number	316392

2. Introduction

2.1. Purpose and scope of the statistical analysis plan

This statistical analysis plan was written by Peter Martin and describes the main statistical analyses to be applied to the data from the ERiC trial. This document does not cover health economic analyses.

2.2. Timing of Analysis

The analyses described within this analysis plan will begin to be performed after all the data from the primary endpoint (end of treatment) have been entered, checked and locked and this analysis plan has been finalised. Data relating to treatment session attendance and treatment dropout will only become available after the 40-week follow-up. Therefore analyses of these data, as well as the complier average causal effects analysis that depends on attendance data, will be performed once these data have been entered, checked and locked. Further analyses of 40-week follow-up data will be performed after all the data from the 40-week-follow up have been entered, checked and locked.

2.3. Data checking

Before analysis and database lock, basic checks will be performed on the quality of the data, focusing on identifying:

- Missing data
- Data outside expected range
- Other inconsistencies between variables, e.g. in the dates the questionnaires were completed

If any inconsistencies are found, the corresponding values will be double checked with the researchers and corrected if necessary in the source data. This checking process and subsequent changes will be documented.

3. Description of the trial

3.1 Intervention

The adaptation of MBT used in this study is a manualized, transdiagnostic model designed for children aged 6–12 and their carers. It aims to promote adaptive ER by attending to the capacity to understand oneself and others in terms of underlying mental states. MBT consists of 6–8 sessions, delivered fortnightly, which can flexibly involve different members of the family, with a primary focus on promoting mentalizing and emotion regulation in the parent-child relationship.

3.2 Control condition

The control group are offered treatment as usual (TAU). As there is no single evidence-based treatment for this group, and because practice is not standardised across CAMH services in England, TAU is likely to include CBT, parenting groups, and/or children's social skills groups. Scoping work with CAMH services has confirmed that 6–8 fortnightly sessions would be the usual length of intervention for this client group.

3.3 Randomization

Participants will be randomised to the intervention or control arm with a 1:1 ratio. We will use randomized permuted blocks stratified by CAMH service and by age group (6-9 years vs 10-12 years old). The randomisation will be undertaken by a member of the research team via an electronic database using a pre-programmed, concealed randomized list. For details see the trial protocol.

3.4 Duration of the treatment period and frequency of follow up

In both arms, data will be collected at baseline, mid-treatment (8 weeks from baseline), end of treatment (16 weeks from baseline) and 40 weeks from baseline (follow-up). Data on participants' experiences of therapy will be collected at the end of treatment only.

4. Data Collection

4.1 Participant characteristics

The following demographic characteristics will be collected at baseline from the parent or carer. About the child: gender, age, ethnicity, type of school, presence of special educational needs, social care involvement, eligibility for free school meals. About the carer: age, gender, household size, partnership status, tenancy, long-term illness, highest qualification, employment status, income, relationship to child.

4.2 Outcome data

Outcome data will be collected at baseline, mid-treatment (8 weeks from baseline), end of treatment (16 weeks from baseline) and 40 weeks from baseline (follow-up).

4.2.1 Primary outcome

The primary outcome is the parent-rated Strengths and Difficulties Questionnaire (SDQ) Total Difficulties Score (Goodman, 1997, 2001), measured at the end of treatment (16 weeks after baseline assessment).

4.2.2 Secondary outcomes

The following secondary outcomes will be measured:

- Me and My Feelings Questionnaire (Deighton et al., 2013; Patalay et al., 2014), two subscales:
    - Emotional Difficulties
    - Behavioural Difficulties
- Progress towards personalized treatment goals: Goals-based outcome measure
- Parenting/carer stress: Parental Stress Index – Short Form (Haskett et al., 2006):
    - Total Score, and three subscales:
    - Parental Distress
    - Dysfunctional Interactions
    - Difficult Child
- Parental/carer mentalizing capacity: Parental Reflective Functioning Questionnaire (Luyten et al., 2017), with two subscales (a third subscale, Certainty about Mental States, is not a secondary outcome in this study)
    - Interest and Curiosity in Mental States
    - Pre-mentalizing modes
- Child Emotion Regulation: Emotional Awareness Questionnaire (self-report) (Rieffe et al., 2008), six subscales:
    - Differentiating Emotions
    - Verbal Sharing of Emotions
    - Not Hiding Emotions
    - Bodily Awareness
    - Attending to Others' Emotions
    - Analyses of Emotions
- Child Emotion Regulation: Emotion Regulation Checklist (carer report) (Reis et al., 2016; Shields & Cicchetti, 1997), two subscales:
        - Emotion regulation subscale
        - Emotion Lability/Negativity subscale
- Carer Emotion Regulation: Difficulties in Emotion Regulation Scale (carer self-report) (Gratz & Roemer, 2004): Total Score

Four further variables are listed as secondary outcomes in the study protocol:

- The Child and Adolescent Service Use Schedule (CA-SUS) and the Service Use Record, a questionnaire completed by CAMHS team to record the services offered and attended between baseline and end of treatment, and end of treatment and follow up. These will be used to estimate treatment costs and other service use costs. Their scoring and analysis is described in a separate Health Economics Analysis Plan.

- The Test of Emotional Comprehension and the Parent-Child Interaction Task. These were tested in the pilot study and administered to a subsample of participants only. They will thus not be part of the formal outcome analysis of the trial, although exploratory and secondary analyses will be performed on the data (not described in this document).

Details of all outcome measures and scoring methods are given in Table 1.

4.2.3 Mediating mechanisms

Carer emotion regulation (DERS) and child emotion regulation (EAQ), measured at mid-treatment, are hypothesized to be joint mediators. The mediation analysis strategy is described in Section 6.

4.3 Fidelity

Treatment fidelity will be assessed using the supervisor-report version of the MBT (Child) Fidelity Scale (Morris et al., n.d.).

4.4 Acceptability

We will describe the distribution of the number therapy sessions attended separately for the intervention and control arms. We will also document the proportion of treatment dropout separately for the two arms. Treatment dropout will be defined as the therapist recording the treatment as having ended prematurely.

Table 1: Outcome measures and scoring

Outcome	Details and scoring
SDQ Total Difficulties	The SDQ Total Difficulties Score is calculated as the sum of the 20 items comprising the subscales of Emotional Problems, Conduct Problems, Peer Problems, and Hyperactivity/Inattention. Some items are reversed such that a higher score indicates more difficulty. Each item is rated on a 3-point scale (Not True, Somewhat True, Certainly True) and scored 0, 1, 2. The SDQ Total Difficulties Score ranges from 0 to 40. A higher score indicates more, or more severe, difficulties. Individual missing items can be imputed using individual mean imputation (prorating) by sub-scale, as long as a maximum of two items are missing in a particular subscale (https://www.sdqinfo.org/py/sdqinfo/c0.py). For example, if two items from the Emotional Problems subscale are missing for a participant, the mean of the remaining three items for that participant will be calculated and used to replace the two missing values. If any subscale has more than two items missing for an individual, the SDQ Total Difficulties Score will be considered missing for that individual at that time point.
Me and My Feelings Questionnaire: Emotional Difficulties	This is a 10-item scale. Each item (e.g. "I am unhappy") is scored on a three-point response scale (0 = never, 1 = sometimes, and 2 = always). The scale score is calculated as the sum of the 10 items.
Me and My Feelings Questionnaire: Behavioural Difficulties	This is a 6-item scale. Each item (e.g. "I hit out when I am angry") is scored on a three-point response scale (0 = never, 1 = sometimes, and 2 = always). One item ("I am calm") is reverse coded. The scale score is calculated as the sum of the 6 items.
Parenting Stress Index – Short Form (PSI): Total Score	The PSI total score is calculated as the sum of the three subscale scores (see below). The total score can be calculated if each of the three subscales has at most one item missing (Abidin, 2012). If one or more subscales have more than one item missing, the PSI total score will be considered a missing value.
PSI: Parental Distress	This is a 12-item scale. Each item (e.g. "I feel trapped by my responsibilities as a parent") is scored on a 5-point response scale (from 1 = Strongly disagree to 5 = Strongly agree). The scale score is calculated as the sum of the 12 items. A subscale score can be prorated if one item is missing. If more than one item is missing, the subscale will be considered a missing value.
PSI: Dysfunctional Interactions	This is a 12-item scale. Each item (e.g. "My child rarely does things that make me feel good") is scored on a 5-point response scale (from 1 = Strongly disagree to 5 = Strongly agree for most items, other 5-point scales for some items). The scale score is calculated as the sum of the 12 items. A subscale score can be prorated if one item is missing. If more than one item is missing, the subscale will be considered a missing value.
PSI: Difficult Child	This is a 12-item scale. Each item (e.g. "My child gets upset easily over the smallest thing") is scored on a 5-point response scale (from 1 = Strongly disagree to 5 = Strongly agree for most items, other 5-point scales for some items). The scale score is calculated as the sum of the 12 items. A subscale score can be prorated if one item is missing. If more than one item is missing, the subscale will be considered a missing value.
Parental RF: Pre-mentalizing modes	This is a 6-item scale. Each item (eg "My child cries around strangers to embarrass me.") is scored on a 7-point scale (1 to 7). The scale score is calculated as the mean of

Outcome	Details and scoring
	the six items. Higher values indicate stronger presence of pre-mentalizing modes, which is an aspect of poorer RF.
Parental RF: Interest and Curiosity in Mental States	This is a 6-item scale. Each item (eg "I try to see situations through the eyes of my child.") is scored on a 7-point scale (1 to 7). One item ("I believe there is no point in trying to guess what my child feels.") is reverse coded. The scale score is calculated as the mean of the six items. Higher values indicate more interest/curiosity, which is an aspect of higher RF.
Emotion Awareness Questionnaire (EAQ): Differentiating Emotions	This is a 7-item subscale of the EAQ. Each item (eg "I am often confused or puzzled about what I am feeling") is rated on a three-point response scale: 1 = not true, 2 = sometimes true, 3 = often true. Some items are reverse coded. Subscale scores are calculated as the means of the scores of subscale items. Higher values indicate higher emotion awareness.
EAQ: Verbal Sharing of Emotions	This is a 3-item subscale of the EAQ. Items (eg "I can easily explain to a friend how I feel inside") are rated and averaged as described above. Higher values indicate Verbal Sharing.
EAQ: Not Hiding Emotions	This is a 5-item subscale of the EAQ. Items (eg "When I am upset, I try not to show it") are rated and averaged as described above. Higher values indicate Not Hiding.
EAQ: Bodily Awareness	This is a 5-item subscale of the EAQ. Items (eg "When I am sad, my body feels weak") are rated and averaged as described above. Higher values indicate Bodily Awareness.
EAQ: Attending to Others' Emotions	This is a 5-item subscale of the EAQ. Items (eg "If a friend is upset, I try to understand why") are rated and averaged as described above. Higher values indicate Attending to Others' Emotions.
EAQ: Analyses of Emotions	This is a 5-item subscale of the EAQ. Items (eg "It is important to understand how I am feeling") are rated and averaged as described above. Higher values indicate Analyses of Emotions.
Emotion regulation checklist for children (ERC, Shields & Cicchetti 1997): Emotion regulation Scale	This is an 8-item subscale of the ERC. Each item (e.g. "Can recover quickly from episodes of upset or distress") is rated on a 4-point response scale from 1= Rarely or Never to 4 = Almost Always. Some items are reverse coded. A total score is calculated as the sum of the item scores.
ERC: Emotion Liability/Negativity Scale	This is a 16-item subscale of the ERC. Each item (e.g. "Exhibits wide mood swings") is rated on a 4-point response scale from 1= Rarely or Never to 4 = Almost Always. Some items are reverse coded. A total score is calculated as the sum of the item scores.
Difficulties in Emotion Regulation Scale (DERS)	This is a 36-item questionnaire that has six subscales. Each item (eg "When I'm upset, I lose control over my behaviours") is measured on a 5-point response scale (1 = Almost Never to 5 = Almost Always). Several items are reverse coded. The six subscales are: Nonacceptance of emotional responses, Difficulties engaging in goal-directed behaviour; impulse control difficulties; lack of emotional awareness; limited access to emotion regulation strategies; lack of emotional clarity. Each subscale score is calculated as the sum of its items, and the total score is calculated as the sum of the subscales.
Goals-based outcome measure (GBO)	Participants can specify up to three goals at baseline, where each goal is rated on a scale from 0 (no progress) to 10 (goal achieved) as a starting point. At subsequent time points, progress towards the goal is rated on the same response scale.

5. Data analysis

Analyses will be carried out based on the intention to treat principle, comparing the groups as randomised regardless of compliance with the intervention. The primary analysis will be performed on observed outcome values (without imputation, except missing item imputation discussed below to enable us to calculate total scores for scales). All statistical hypothesis tests will be two-sided. Confidence intervals will be symmetric around the point estimate and using the 95 % level of confidence.

5.1 Recruitment and representativeness of recruited patients

A consort diagram will be constructed to describe the flow of subjects through the trial (http://www.consort-statement.org/). The diagram will detail the number of subjects: invited to participate; agreeing to enter the study (with reasons for refusal); receiving the intervention (with reason for not receiving this); followed up and withdrawn (with reasons).

5.2 Baseline characteristics

Baseline characteristics of the participants will be summarised by treatment arm to gauge the balance in characteristics between the randomised groups. The results will be presented as means, standard deviations, medians and inter-quartile ranges for numeric variables; and frequencies and percentages for categorical variables. No statistical hypothesis testing will be used.

5.3 Adherence to treatment

Adherence to the MBT intervention is defined as attending at least four MBT sessions. We will document the distribution of attended sessions in both arms as outlined above in Section 4.4.

5.4 Attrition

Some loss to follow-up is expected over the 16-weeks period leading up to the primary endpoint, and the 40-week follow-up period. Reasons for missing outcome data will be described, and the frequency and percentage of subjects with missing data, by reason, will be provided for each outcome and for each randomised group.

5.5 Adverse event reporting

Adverse events (AE) and serious adverse events (SAE) will be summarised. These events are defined in the protocol.

5.6 Primary and secondary outcome analysis: general considerations

Although this is an individually randomized trial, participants are clustered within CAMH services. The intervention is delivered by therapists that have been specifically trained. Treatment as Usual is delivered by different therapists within the same clinical teams that have not been trained in the intervention. Hence study participants are also clustered within therapists. Data from two timepoints (mid-intervention and post-intervention) will be used in the analysis, and measurement occasions are clustered within participants. Therefore, a statistical model that fully respects the nested structure of the data is a four-level model with random effects for the participant, the therapist, and the CAMH service. However, in practice we expect the average number of participants treated by the same therapist to be small, and the proportion of therapists who treat only a single participant to be large. In this situation, problems with model convergence are common. In the event that the four-level model does not converge and cannot be fitted, we will attempt to fit a 3-level model with random intercept terms for participant and CAMH service only.

5.7 Analysis of primary outcome

The primary outcome is the SDQ Total Difficulties Score at the end of treatment. A mixed-effects linear regression model will estimate the effect of the intervention on the SDQ Total Difficulties Score. The model will adjust for child's age group and the SDQ score at baseline, as these are assumed to be predictive of outcomes. Random intercepts will be included to account for clustering of participants within CAMH services and therapists, and to account for clustering of observations (mid-treatment and end-of-treatment) within participants. The evidence for a difference in SDQ Total Difficulties scores at the end of treatment will be assessed via a two-sided t-test of the coefficient of the treatment indicator variable (β_1 in the equation below). A 5% significance level will be used as a cut-off to support a finding of superiority of MBT over TAU, or vice versa. The primary analysis will be an intention-to-treat analysis and include all participant with a valid measurement of the SDQ at baseline and at either mid-treatment, end-of-treatment, or both.

$$SDQ_{ijkl} = \left(\beta_0 + u_{jkl} + u_{kl} + u_k\right) + (\beta_1)MBT_{jkl} + (\beta_2)TIME_{ijkl} + (\beta_3)MBT_{jkl} \times TIME_{ijkl} + (\beta_4)SDQ.0_{jkl} + (\beta_5)AGE_{jkl} + \varepsilon_{ijkl}$$

(Model 1)

where

- SDQ_{ijkl} is the SDQ Total Difficulties score post-baseline (at time *i*); a higher score indicates the presence of more and/or more severe difficulties;
- *l* = 1, …, *L* indicates the CAMH services;
- *k = 1, …, n_L* indicates the therapists*; n_L* denotes the number of therapists situated in the *l*th service;

- $j = 1, \ldots, n_k$ indicates the participants; n_k denotes the number of participants treated by the k^{th} therapist;
- $i = 2, 3$ indicates the time points; 2 denotes T2 (mid-treatment), 3 denotes T3 (end of treatment);
- β_0 is the fixed intercept;
- $u_{jkl} \sim N(0, \sigma^2_{u.jkl})$ is a random intercept accounting for differences in outcomes between participants;
- $u_{kl} \sim N(0, \sigma^2_{u.kl})$ is a random intercept accounting for differences in outcomes between therapists;
- $u_l \sim N(0, \sigma^2_{u.l})$ is a random intercept accounting for differences in outcomes between services;
- $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ is a random error that accounts for between-participant variability in outcomes;
- MBT_{jkl} is the treatment indicator variable, coded 1 for MBT and 0 for TAU; the associated coefficient β_1 represents the difference in primary outcome between the two groups at the primary endpoint (relative treatment efficacy);
- $TIME_{ijkl}$ is the time indicator, coded 0 for the primary endpoint, and 1 for mid-treatment;
- $MBT_{jkl} \times TIME_{ijkl}$ specifies the interaction of treatment and time; the associated coefficient represents the difference in relative treatment efficacy between the two time points;
- $SDQ.0_{jkl}$ is the participant's baseline SDQ score;
- AGE_{jkl} is the child participant's age group, coded 0 for 6-9 years old and 1 for 10-12 years old.

Participants will be included in this analysis if they provide either a mid-treatment or an end-of-treatment outcome score, or both. In the presence of loss to follow-up, estimates from this model would be valid under the assumption that data are missing at random conditional on the variables included in the model (Sullivan et al., 2018).

5.8 Analysis of secondary outcomes

The results for the secondary outcomes will be presented as estimates with 95% confidence intervals. *P*-values will not be reported. Analyses will compare groups defined by intention to treat and include all those with available data.

5.8.1 Model for secondary outcomes

All secondary outcomes are numeric scales. With the exception of Goals-based outcomes, the analogous four-level model as for the primary outcome will be fit for each secondary outcome, adjusting for each outcome's respective baseline score.

5.8.2 Goals-based outcomes

Each participant can specify up to three goals, and progress ratings are made at all follow-up time points. Progress ratings of goals are thus nested within measurement occasions (time points). We will attempt to fit a five-level model, analogous to the primary outcome model, but with an additional random intercept for time point, so that the individual goal is the unit of analysis. If this model proves too complex too fit, we will conduct exploratory analyses to ascertain whether dropping one or several levels of clustering from the analysis model will enable us to fit the model.

5.9 Missing items in scales and subscales of secondary outcomes

Participants who provide data at baseline or follow-up may sometimes respond to some but not all items that comprise a psychometric scale. This paragraph specifies procedures for dealing with item-missingness. For our primary outcome, SDQ Total Difficulties, a procedure for individual mean imputation (prorating) is specified to deal with missing item values. The same is true for the Parental Stress Index. See Table 1 for relevant procedures. For secondary outcome scales that do not have a specified method for dealing with item missingness, we will proceed as follows: If individual items in scales or subscales are missing, we will use individual mean imputation if 20% or fewer items are missing for an individual in a questionnaire. For example, in a scale with 10 items, individual mean imputation will be applied to individuals with up to 2 items missing. The average value for the complete items will be calculated for that individual and used to replace the missing values. The scale score will be calculated based on the complete values and these replacements. If more than 20 % of scale items are missing for an individual, this scale will be treated as a missing value for that individual.

5.10 Supportive analyses

The following supportive analyses will be carried out:

- *Adjustment for baseline characteristics*. For the primary and secondary outcomes, using the same modelling approaches as described previously, the treatment effect will be estimated adjusting for any concerning imbalances in baseline characteristics.
- *Loss to follow-up*: If either considerable loss to follow-up or considerable differences in follow-up rates between treatment arms is encountered, we will consider conducting information-anchored sensitivity analyses under a range of Not Missing At Random (NMAR) assumptions, as indicated in the trial protocol (Cro et al., 2016,

2019). These analyses would be complex and require input from non-statistician members of the project team to specify a range of plausible MCAR processes, and are not prespecified in this statistical analysis plan.

- *Complier Average Causal Effect (CACE) analysis*. A CACE analysis will be carried out for the primary outcome, including only those participants in the treatment arm who adhered to the intervention. Adherence in the intervention arm is defined as attending four or more MBT sessions (See Section 5.3.).

5.11 Model checking and dealing with convergence issues

If a multilevel model for a given outcomes yields a boundary estimate (zero) for a random intercept, we will re-estimate the model without the random intercept. If a model fails to converge, we will in the first instance attempt to re-estimate the model without the random intercept for therapist (as explained in Section 5.6). If this simplified model does not converge either, we will conduct exploratory analyses to investigate whether dropping another random intercept term (eg for CAMH service) will enable us to fit the model.

The linear mixed effects models for our primary and secondary outcomes assume that the errors (ε_{ijkl}) are normally distributed and homoscedastic. This will be checked using residual plots. The same analyses will also check the compatibility of the data with the assumption of linearity in the relationship between each outcome's baseline score and its post-baseline scores. Since our outcome measures are scales with fixed minima and maxima, outliers with high leverage are unlikely to occur, but we will check for influential observations also. If substantial departures from normality, homoscedasticity or linearity occur, additional sensitivity analyses using suitable transformations of the relevant outcome variables will be considered. If there are influential observations, sensitivity analyses will be conducted to investigate the influence of these observations on the treatment effect estimate.

5.12 Blinding

The statistician will conduct the analysis of primary and secondary outcomes blind to treatment allocation. The trial manager will send a data set for analysis where treatment arms are concealed within the treatment variable. Once all analyses have been completed and discussed with the Principal Investigator, the statistician will be unblinded for the purpose of preparing the statistical report of the trial analyses.

6. Mediation Analysis

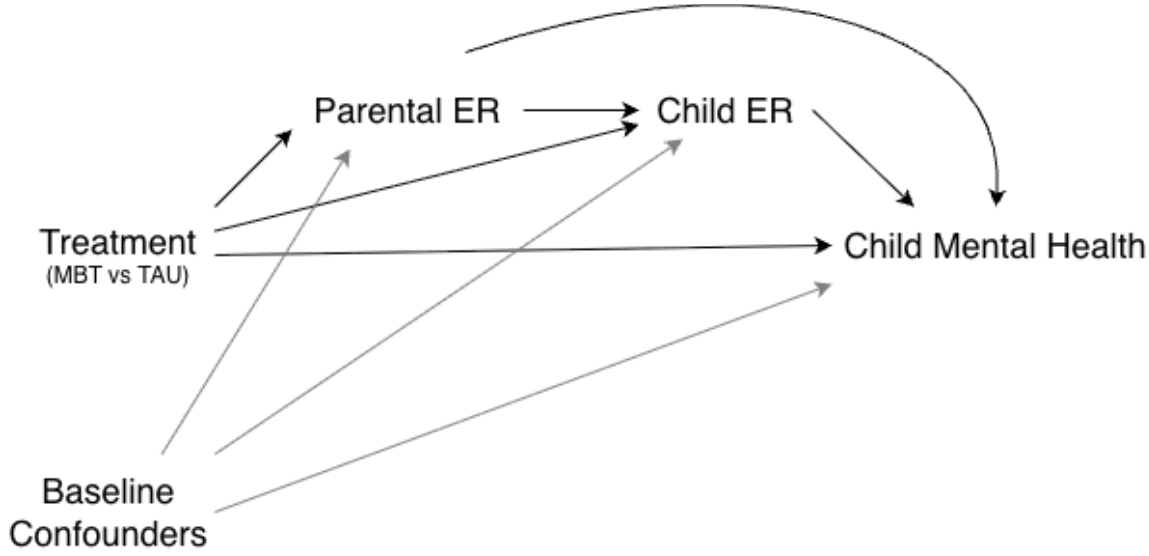
The mediation analysis will be conducted after the statistician has been unblinded. In what follows, the mediation analysis is described in some detail. This is partly due to the

relatively new methods to be used, and partly due to its conceptual importance for the trial, which made detail pre-registration of the methods desirable. Nonetheless, the mediation analysis is a secondary analysis, and no specific power calculation was performed for this part of the analysis.

6.1 Assumptions underlying the mediation analysis

Two mediators are posited: parental emotion regulation (parental ER, measured by the DERS) and child emotion regulation (child ER, measured by the six subscales of the EAQ). Thus there are seven mediating variables to consider. The mid-treatment measures of these variables will be considered to be mediators on the causal path from treatment (MBT vs TaU) to mental health outcome (SDQ, measured at the end of treatment). The directed acyclic graph (DAG) specifying the assumed causal relationships is shown in Figure 1. Note that we assume a causal ordering whereby, within the time frame from baseline up to mid-treatment, parental ER may affect child ER, but not the other way around. The six EAQ subscales are considered as joint mediators without an internal causal ordering.

Figure 1: Directed acyclic graph of assumed causal relationships between treatment, outcome, mediators, and covariates.



Note: ER: Emotion Regulation

6.2 Analytic strategy for mediation analysis

The assumption of causal ordering among the mediators informs a sequential strategy for the mediation analysis (T. J. VanderWeele, 2015; T. VanderWeele & Vansteelandt, 2013): in Stage 1, we will estimate the mediation effect of parental ER only. In Stage 2, we will will estimate the joint mediation effect of parental ER and child ER. Note that it is not easily possible to estimate the mediation effect of child ER only, since according to our assumed DAG parental ER would act as an intermediate confounder. Also note that, if the assumption of a causal ordering between parental ER and child ER is not deemed plausible, the estimation in Stage 2 would still be valid as an estimation of the joint mediation effect of both parental and child ER.

We will conduct a causal mediation analysis using the following variables:

- Treatment: randomized to MBT vs randomized to TaU
- Mediator 1: parental emotion regulation, measured by the DERS at mid-treatment
- Mediators 2-7: child emotion regulation, measured by the six EAQ subscales at mid-treatment
- Outcome: child mental health, measured by the SDQ Total Difficulties score at the end of treatment
- Covariates (measured at baseline):
    - SDQ Total Difficulties; DERS; EAQ subscales; CAMH service
    - All demographic participant characteristics:

- About the child: gender, age (6-9 vs 10-12 yrs old), ethnicity, type of school, presence of special educational needs, social care involvement, eligibility for free school meals.
- About the carer: age, gender, household size, partnership status, tenancy, long-term illness, highest qualification, employment status, income, relationship to child.

Baseline demographic variables may be dichotomized or otherwise discretized to aid with estimation in the presence of sparseness. Some demographic variables may be dropped if they are found to be highly collinear with one another, to be unrelated to treatment, mediators and outcome, and/or if the models described below become difficult or unstable due to the large number of covariates.

6.3 Research questions for mediation analysis

In the event that the primary outcome analysis shows evidence in favour of a treatment effect on child mental health, the mediation analysis will assess how much, if any, of that effect is due to mediation via (1) parental ER, and (2) parental ER and child ER jointly.

On the other hand, if the primary outcome analysis fails to show evidence in favour of a treatment effect, the mediation analysis will assess whether this result can be illuminated by investigating whether the treatment failed to affect the mediator, whether the mediator failed to affect the outcome, or whether there was paradoxical mediation (whereby the treatment effects mediated and not mediated via parent and child ER may act in opposite directions).

6.4 Analysis model

We will follow the inverse probability weighting approach described in VanderWeele (T. J. VanderWeele, 2015, p. 122ff) and VanderWeele & Vansteelandt (2013). This involves estimating three counterfactual outcomes:

- $Q_1 = E\left[Y_{tau,M_{tau}}\right]$: the expected value of the outcome when the treatment is TAU, and the mediator(s) have the value they would have under TAU
- $Q_2 = E\left[Y_{mbt,M_{mbt}}\right]$: the expected value of the outcome when the treatment is MBT, and the mediator(s) have the value they would have under MBT
- $Q_3 = E\left[Y_{mbt,M_{tau}}\right]$: the expected value of the outcome when the treatment is MBT, but the mediator(s) have the value they would have under TAU

These quantities will be estimated as follows:

Q_1: We take the weighted average of the outcome among the TaU participants, where the weights are equal to the probability of receiving TaU, divided by the probability of receiving TaU given the covariates. Formally:

$$Q_1 = \sum_{i=1}^{N_{TAU}} w_i Y_i$$

Where:

- $w_i = \frac{P(T=TAU)}{P(T=TAU \mid \boldsymbol{C}=\boldsymbol{c_i})}$,
- N_{TAU} is the number of TAU participants,
- Y_i is the outcome score for the ith TAU participant
- C is the vector of covariates, and c_i are the observed covariates values for the ith TAU participant.

The probability $P(T = TAU \mid \boldsymbol{C} = \boldsymbol{c_i})$ will be estimated using logistic regression.

Q_2: We take the weighted average of the outcome among the MBT participants, where the weights are equal to the probability of receiving MBT, divided by the probability of receiving MBT given the covariates. Formally:

$$Q_2 = \sum_{j=1}^{N_{MBT}} v_j Y_j$$

Where:

- $v_j = \frac{P(T=MBT)}{P(T=MBT \mid \boldsymbol{C}=\boldsymbol{c_j})}$
- N_{MBT} is the number of MBT participants,
- Y_j is the outcome score for the jth MBT participant
- C is the vector of covariates, and c_j are the observed covariates values for the jth MBT participant.

Estimates of the probability $P(T = MBT \mid \boldsymbol{C} = \boldsymbol{c_j})$ will be derived from the same logistic regression used to estimate the w_i (see above).

Q_3: We model the outcome using a linear regression model that includes as predictors the treatment, the mediator(s), and all covariates, as well as treatment-mediator interaction(s) and, in the case of multiple mediators, mediator-mediator interactions. Using the estimates from this model, we calculate the predicted outcome among the TaU participants under the counterfactual event that they had received MBT, but using their own observed values for the mediator(s) and covariates. That is, we predict what the outcome would have been for

TaU participants had they received MBT instead, but if their mediators had not been affected by MBT. To calculate Q_3, we again weight these predicted outcome values by w_i defined above.

Formally:

$$Q_3 = \sum_{i=1}^{N_{TAU}} w_i \, E[Y|T = MBT, \boldsymbol{M} = \boldsymbol{m_i}, \boldsymbol{C} = \boldsymbol{c_i}]$$

Where:

- M is the vector of mediators, and m_i are the observed values of *M* for the i[th] TAU participant.

The quantities Q1, Q2, and Q3 can then be used to estimate the (marginal) natural direct effect (NDE) and natural indirect effect (NIE) as follows:

$$\widehat{NDE} = Q_3 - Q_1$$
$$\widehat{NIE} = Q_2 - Q_3$$

Intuitively, the natural direct effect considers the counterfactual effect of MBT on the outcome if MBT had *not* affected the mediators. The natural indirect effect considers the counterfactual effect on the outcome if the treatment had affected the outcome *only* through the mediators.

Our procedure will result in the following estimates:

Stage 1:

- Marginal natural indirect effect of the treatment on child mental health through pathways via parental ER
- Marginal natural direct effect of the treatment on child mental health through pathways not via parental ER

Stage 2:

- Marginal natural indirect effect of the treatment on child mental health through pathways via parental ER and child ER jointly
- Marginal natural direct effect of the treatment on child mental health through pathways not via either parental ER or child ER

6.5 Estimation of confidence intervals and dealing with missing values

Missing values may occur in covariates, mediators, and in the outcome. We will exclude participants from the analysis whose outcome values are missing at both the mid-treatment

and end-of-treatment time points. For all other participants, we will use multiple imputation with chained equations to impute missing values. The imputation model will include all variables used in the mediation analysis (listen in section 6.2), and additionally will include the following auxiliary variables:

- SDQ measured mid-treatment
- EAQ and DERS measured at the end of treatment
- If other secondary outcomes, such as PSI, PRFQ and/or ERC, have considerably higher proportions of completeness compared to the mediators at mid-treatment, we will consider adding them as auxiliary variables

The imputation model will include interactions of parental ER with all six measures of child ER. Imputation will be performed separately for the two treatment arms, thereby implicitly allowing for interactions of the treatment with all other variables in the imputation model. We will impute 20 data sets, unless the percentage of participants with incomplete data is considerably higher than 20 %, in which case we will impute a number of data sets equal to or higher than this percentage, rounding up to the nearest multiple of 10 %. Point estimates of natural direct and indirect effects will be calculated as the arithmetic means of the estimates derived from the imputed data sets.

To estimate 95 % confidence intervals for the NDE and NIE, we will use the non-parametric bias-corrected and accelerated bootstrap with 2000 bootstrap samples. Assuming 20 imputed data sets, we shall draw 100 bootstrap samples from each imputed data set to obtain 2000 bootstrap samples overall.

6.6 Further analyses of mediation

If the results of the mediation analyses described above would benefit from further illumination through estimation of specific parameters, such as treatment-mediator interactions or mediator-mediator interactions, we will consider exploring an alternative estimation strategy, in addition to the approach outlined above, using the regression-based approach to mediation analysis with multiple mediators described in VanderWeele (2015, pp. 114–122)

7. Further analyses

Once data from the 40-week follow-up will have become available, we will conduct further analyses of the primary and secondary outcomes, in order to establish the relative efficacy of MBT vs TaU in the medium term. We will extend the longitudinal mixed effects model defined in Section 5.7 to include the 40-week follow-up. The time variable will be defined

thus that the coefficient of the treatment arm indicator measures the difference in mean outcome between the MBT and TaU arms at 40-week follow-up.

References

Abidin, R. (2012). *Parenting Stress Index*. Psychological Assessment Resources.

Cro, S., Carpenter, J. R., & Kenward, M. G. (2019). Information-anchored sensitivity analysis: theory and application. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *182*(2), 623–645. https://doi.org/10.1111/rssa.12423

Cro, S., Morris, T. P., Kenward, M. G., & Carpenter, J. R. (2016). Reference-based Sensitivity Analysis via Multiple Imputation for Longitudinal Trials with Protocol Deviation. *The Stata Journal: Promoting Communications on Statistics and Stata*, *16*(2), 443–463. https://doi.org/10.1177/1536867x1601600211

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., Martin, A., Patalay, P., & Wolpert, M. (2013). The Development of a School-Based Measure of Child Mental Health. *Journal of Psychoeducational Assessment*, *31*(3), 247–257. https://doi.org/10.1177/0734282912465570

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *38*(5), 581–586. http://www.ncbi.nlm.nih.gov/pubmed/9255702

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*(11), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Gratz, K. L., & Roemer, L. (2004). Multidimensional Assessment of Emotion Regulation and Dysregulation: Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale 1. In *Journal of Psychopathology and Behavioral Assessment* (Vol. 26, Number 1).

Haskett, M. E., Ahern, L. S., Ward, C. S., & Allaire, J. C. (2006). Factor structure and validity of the parenting stress index-short form. *Journal of Clinical Child and Adolescent Psychology*, *35*(2), 302–312. https://doi.org/10.1207/s15374424jccp3502_14

Luyten, P., Mayes, L. C., Nijssens, L., & Fonagy, P. (2017). The parental reflective functioning questionnaire: Development and preliminary validation. *PLoS ONE*, *12*(5). https://doi.org/10.1371/journal.pone.0176218

Morris, E., Dwyer Hall, H., & Midgley, N. (n.d.). MBT (child/family) Fidelity Scale. In *Unpublished Manuscript*. Anna Freud National Centre for Children and Families / University College London.

Patalay, P., Deighton, J., Fonagy, P., Vostanis, P., & Wolpert, M. (2014). Clinical validity of the Me and My School questionnaire: A self-report mental health measure for children and adolescents. *Child and Adolescent Psychiatry and Mental Health*, *8*(1). https://doi.org/10.1186/1753-2000-8-17

Reis, A. H., De Oliveira, S. E. S., Bandeira, D. R., Andrade, N. C., Abreu, N., & Sperb, T. M. (2016). Emotion Regulation Checklist (ERC): Preliminary studies of cross-cultural adaptation and validation for use in Brazil. *Temas Em Psicologia*, *24*(1), 97–116. https://doi.org/10.9788/TP2016.1-07

Rieffe, C., Oosterveld, P., Miers, A. C., Meerum Terwogt, M., & Ly, V. (2008). Emotion awareness and internalising symptoms in children and adolescents: The Emotion Awareness Questionnaire revised. *Personality and Individual Differences*, *45*(8), 756–761. https://doi.org/10.1016/j.paid.2008.08.001

Shields, A., & Cicchetti, D. (1997). Emotion Regulation Among School-Age Children: The Development and Validation of a New Criterion Q-Sort Scale. In *Developmental Psychology* (Vol. 33, Number 6).

Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., & Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical Methods in Medical Research*, *27*(9), 2610–2626. https://doi.org/10.1177/0962280216683570

VanderWeele, T. J. (2015). *Explanation in causal inference: methods for mediation and interaction*. OxfordUniversityPress.

VanderWeele, T., & Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic Methods*, *2*(1), 95–115. https://doi.org/10.1515/em-2012-0010