Mapping and Evaluating Services for Children with Learning Disabilities and Behaviours that Challenge (MELD):
Stage 2 Statistical Analysis Plan

v1.5 12.03.2025

Sponsor:	University of Warwick
Sponsor ref:	SOC.05/21-22
Funder:	National Institute for Health Research Health Services and Delivery Research
Funder ref:	NIHR 129577
IRAS ref:	310149
ISRCTN	TBC
Protocol No.	v1.5 21.02.2024
Project Duration	Total project length 01 Feb 2021 to 31 December 2023 (35 months) Current Stage 2 protocol 01 March 2022 to 31 December 2023

Table of Contents

1 Research Questions 2
2 Outcome measures 6
2.1 Primary outcome measure 12
2.2 Secondary outcome measures 13
3 Proposed Analysis Framework ..... **Error! Bookmark not defined.**
3.1 Directed Acyclic Graph (DAG) ..... **Error! Bookmark not defined.**
3.2 Two level model specification ..... **Error! Bookmark not defined.**
3.3 Propensity score calculation ..... **Error! Bookmark not defined.**
3.4 Sample size ..... **Error! Bookmark not defined.**
3.5 Missing data ..... **Error! Bookmark not defined.**
3.6 Checking model assumptions ..... **Error! Bookmark not defined.**
- Homogeneity of variance – variance of the residuals across groups is the same. ..... **Error! Bookmark not defined.**
- Residuals are approximately normally distributed – plotting QQ plot .... **Error! Bookmark not defined.**

3.7 Inclusion in analysis/sampling strategy ..... **Error! Bookmark not defined.**
4 References ..... **Error! Bookmark not defined.**

Amendment History

The following amendments and/or administrative changes have been made to this SAP since the implementation of the first approved version.

Amendment No. *(specify substantial/non-substantial)*	**SAP version no.**	**Date issued**	**Summary of changes made since previous version**
1	V1.1	21.11.2022	Addressed edits from SSC statistician, Professor Beth Stuart
2	V1.2	19/01/2023	SMG commented that the BPI in primary outcome needed to be clarified that total score is the primary outcome and its components as secondary measures.
3	V1.3	07/02/2023	Updated to include details on contingency plan if composite measures do not converge to an appropriate solution. Also, comment on potential mediators. Added to address suggestions in SSC meeting (06/02/2023) from SSC statistician, Prof Beth Stuart.
4	V1.4	24/05/2023	Added the VABS sensitivity analysis substituting GO4KIDDS for the VABS.
5	V1.5	12/03/2025	Added analysis relating to external stakeholder satisfaction measure

ROLES AND RESPONSIBILITIES

Trial Statistician: Dr Paul Thompson			
Role: Trial Statistician (Research Fellow in Applied Statistics), CEDAR, University of Warwick			
Date:		**Signature:**	
Senior Statistician: Dr David Gillespie			

Role: Director of the Infection, Inflammation & Immunity Trials Division and a Principal Research Fellow, Centre for Trials Research, Cardiff University			
Date:		**Signature:**	
Chief Investigator: Professor Richard Hastings			
Role: Professor and Cerebra Chair of Family Research; HoD at CEDAR, University of Warwick			
Date:		**Signature:**	
Other non-signatory contributor to the SAP: Dr Samantha Flynn/ Mr Nicholas Manktelow			
Role: MELD Study Manager (Assistant Professor/ Research Assistant)			

1. Introduction

This statistical analysis plan provides guidelines for the final presentation and analysis for MELD stage 2. This plan, along with all other documents relating to the analysis of this observational study, will be stored in the Statistical Analysis Master File electronically and/or in hard signed copy formats.

2. Background

2.1 Rationale and Research Questions

Learning disability (LD), used as the official term in the UK health system, is known as Intellectual Disability internationally. Intellectual Disability/LD is a condition described in ICD-11 as a Disorder of Intellectual Development (Salvador-Carulla et al., 2011). Consistent with contemporary definitions of this condition, LD emerges during the "developmental period" (usually taken to mean before age 18 years), and is characterised by low cognitive ability (using standardised tools an IQ <70) and low levels of adaptive behaviour (such as communication, social skills, independence skills - also assessed using standardised tools). Prevalence studies internationally suggest that approximately 2% of children and adolescents have a LD (Maulik et al., 2011). UK Learning Disability Observatory data also show just over 2% of children in England have been identified by local authorities/schools as having LD (Hatton et al., 2014). Prevalence varies slightly with socio-economic factors but is broadly similar across the UK. In practice, and this is also reflected in the ICD-11 "sub-types" of intellectual disability, it is helpful to distinguish between levels of LD severity: mild (2-3 SDs below the mean on standardised IQ/adaptive behaviour assessments), moderate (3-4 SDs below the mean), and severe/profound (4 or more SDs below the mean). In addition, LD is associated with significantly higher prevalence of other neurodevelopmental conditions; in particular Autism Spectrum Disorder

(ASD). The prevalence of LD among children and adolescents with ASD in UK population-based data has been shown to be as high as 52% (95% CI: 42%, 62%) (Totsika et al., 2011).

Children with LD are also likely to display challenging behaviour (or Behaviours that Challenge; BtC). Approximately 1 in 5 children with LD in the UK in contact with services display BtC (Emerson et al., 2001). Recent analysis of UK population data suggest 10-17% of children with LD show aggression towards others (Emerson et al., 2014). In some settings, prevalence rates are higher (e.g., 53% of children in a special school context; Nicholls et al., 2020). BtC are associated with poor care outcomes for children (e.g., increased exposure to restrictive care), for family carers (e.g., increased stress and mental health problems; Hastings, 2016), and increased costs of care to families (Einfeld et al., 2010) and to health and social care services (Iemmi et al., 2016). Children with more severe LD, and those with LD who also have autism, are more likely to display BtC (Nicholls et al., 2020).

BtC are understood theoretically from a contextual perspective in terms of definition, vulnerability factors, and maintaining processes (Bowring et al., 2019; Hastings et al., 2013). First, BtC are a socially defined health and social care issue, rather than a medical disorder or diagnosis; defined as behaviours that are not typical for the culture the person lives in and that occur at a frequency, severity, or duration that places an individual at risk of harm, places carers or others at risk of harm, or that hinder inclusion in typical community settings (Emerson & Einfeld, 2011). BtC are defined in terms of their effects rather than topography. Nevertheless, individuals with LD often engage in a number of behaviours that are typically considered challenging, regardless of context: injuring themselves (e.g., banging their heads against hard surfaces, eye-poking, skin scratching leading to bleeding), physical aggression towards others (e.g., kicking, biting, pulling hair), physically destructive behaviours (e.g., throwing furniture, pulling down curtains), and other actions (e.g., absconding, high rate unusual repetitive behaviours such as body rocking, inappropriate touching, screaming). The second contextual dimension is that the vulnerability factors for BtC are primarily (though not exclusively) psycho-social, relating to the inequalities and life experiences of people with LD (e.g., impoverished social networks, lack of communication skills, exposure to negative life events including abusive care, barriers to accessing health and care services). The third contextual dimension is that BtC are functional for the person engaging in them – they allow a certain amount of control over the (social) environment: the behaviour/response of others is then the main mechanism through which BtC are maintained and may worsen over time.

Given the prevalence of BtC, and continued high profile care scandals (e.g., BBC Panorama exposés of Winterbourne View in 2011, and Whorlton Hall in 2019), effective community-based services and supports are a national priority (NHS England, 2015). However, NICE guidelines for BtC (2015, 2018) found no high quality evidence relating to the design and organisation of services for children. This study focuses on that evidence gap.

There are currently no data on an England-wide basis about the outcomes and costs associated with different models or characteristics of services for children with LD and BtC. Given the lack of evidence overall, the findings from the current research will be directly relevant to the ongoing planning and delivery of health and social care services across the UK.

Primary research questions

1. For children with intellectual disabilities with behaviour that challenges, is referral into a specialist behaviour that challenges service associated with a larger reduction in child challenging behaviour at 12-months post-referral compared to referral into services organised following other models? (Primary outcome measure: Behaviour Problems Inventory Short Form (BPI-S) – TOTAL SCORE - Rojahn et al., 2012)
2. In addition, we will report the ranked posthoc pairwise comparisons between service models to determine which services show the biggest reductions compared to the specialist behaviour that challenge services.
3. For children with intellectual disabilities with behaviour that challenges, is referral into Learning disability – Childrens and Adolescents Mental Health Services (LD-CAMHS) associated with a larger reduction in child challenging behaviour at 12-months post-referral compared to referral into generic CAMHS?

Secondary research questions

Following the primary research questions, we will also explore the same research questions using the other outcomes listed in section 2.

1. For children with intellectual disabilities with behaviour that challenges, is referral into a specialist behaviour that challenges service associated with a larger reduction in child challenging behaviour at 12-months post-referral compared to referral into services organised following other models? (re-running for each secondary outcome measure)

2. In addition, we will report the ranked posthoc pairwise comparisons between service models to determine which services show the biggest reductions compared to the specialist behaviour that challenge services.
3. For children with intellectual disabilities with behaviour that challenges, is referral into Learning disability – Childrens and Adolescents Mental Health Services (LD-CAMHS) associated with a larger reduction in child challenging behaviour at 12-months post-referral compared to referral into generic CAMHS?

2.2 Objectives

In the proposed research, we will measure outcomes and costs for identified service models (see MELD Stage 1 protocol) by following children and families referred to selected services.

The *research objectives* are to:

1. To assess the outcomes for children with LD and BtC and their carers for different service models.
2. To estimate the economic costs of different service models for children with LD and BtC, and to assess their economic outcomes within a cost-consequences framework.
3. To understand the experiences of children and families, and the processes by which child LD and BtC services impact on the lives of children with LD and BtC, and their families and carers.
4. To describe the ways that services for children with LD and BtC co-produce services with families and carers and children and young people.
5. To inform and support decision-making on commissioning and developing services for children with LD and BtC and their families and carers.

3. Study methods

3.1 Study Design

The study will use systematic sampling of service models identified in Stage 1 of the MELD project and monitoring of outcomes for children and their families referred to the selected services and followed up for 12 months.

3.2 Power/sample size considerations

Given that we have identified 5 different service models, 15 examples of services (three per model) in total will be included at Stage 2 – based on illustrative sample size calculations. To detect a standardised effect size of 0.4 in a one-way ANOVA, with 90% power, a two-sided alpha of 0.05, and assuming five service model types identified in Stage 1, a total sample size of 120 children is required. Inflating this sample size for clustering of children within services, assuming an average number of 13 children recruited from each setting and an intra-cluster correlation coefficient (ICC) of 0.05, the required sample size increases to 195 children from 15 settings. This ICC is likely conservative given the subjective nature of the outcome measures (Adams et al., 2004). In our statistical analysis (but not included in our sample size estimates), we will enhance power and precision by adjusting for individual and cluster level characteristics, including service type (which is by definition a cluster level characteristic) and corresponding baseline measures of the outcome variable. We also allow for 20% drop out indicating a final sample size of N=244.
During selection of services we found that in certain service models, there were unlikely to provide adequate number to sufficiently meet the required N per service, so we reassessed out sample size calculation using varying cluster sizes. The coefficient of variation was calculated at 0.22. Following guidelines from Eldridge, Ashby & Kerry (2006), if the coefficient of variation is below 0.23, then there is a negligible effect due to variable cluster size.

3.3 Framework

The study will aim to determine whether referral into a specialist behaviour that challenges service is associated with a larger reduction in child challenging behaviour at 12-months post-referral compared to referral into other services. Therefore, the study is on the basis of superiority of the specialist behaviour that challenges service model.

3.4 Interim analyses

No interim analyses are planned for this study.

3.5 Stopping rules

The study does not have any stopping rules as there is no basis for stopping the study early given no active intervention or manipulation of services normal functioning is planned.

3.6 Timing of final analysis

All outcomes will be analysed after the database is locked following the last 12 month follow up.

3.7 Timing of outcome assessments

All outcome measures are measured at both baseline and 12 months followup, with the exception of the Experience of service questionnaire (ESQ) as this will only be measured at 12 months follow up.

4. Statistical Principles

4.1 Levels of confidence and p-values

All confidence intervals presented will be 95% and two-sided. In addition, all applicable statistical tests will be 2-sided and will be performed using a 5% significance level.

4.1.1 Adjustment of multiplicity

The overall type I error rate for testing between service models for both the primary endpoint and key secondary endpoints will be controlled at the 2-sided 0.05 significance level.

4.2 Definition of protocol deviation

Any deviations from the original protocol will be submitted as substantial amendments where applicable and recorded in subsequent versions of the protocol and SAP.

4.3 Analysis population

A modified intention to treat population of all participants who provide outcome data.

5. Statistical Principles

5.1 Screening data

The following summaries will be presented for all screened Children and their families (overall and by service model):

Enrolment: the number of days recruiting, the number of families screened, the number of families recruited, the number of screened families not recruited, and the reason for non-recruitment.

5.2 Inclusion in analysis/sampling strategy

All eligible services from stage 1 of the MELD study will be considered (N=161), and sampling of services will be according to a systematic sampling process. Initially, a stratified sampling approach to selecting services that were included in MELD stage 1 analyses was proposed. However, a

potential complication of this procedure is that services may not continue involvement in stage 2. Hence, any achieved apparently random sample would be likely be biased.

We propose the following balancing procedure within service models.

1. Rank services based on criteria indicating how likely they conform to a representative example of the service model
2. Secondly rank according to the quality of existing communication between the services and the research team (e.g., it may not be prudent to approach services whose data were obtained in the end with a FOI request).
3. We will continue to recruit from services until our sample size quota per service model is met. This may mean that an additional service(s) may be included in some service models (Sample size adjustment has been checked and detail included earlier).
4. A screening log describing our approach for each service and the outcome of the approach will be recorded. In addition, we will use service-level data as a proxy (that we could obtain for all services) to explore whether we are at risk of selection bias.

5.3 Eligibility

5.3.1 Inclusion criteria

The inclusion criteria for services are:

1. They were a service included in the analysis for Stage 1, **and**
2. Nothing significant has changed in their service model since the Stage 1 analysis

The inclusion criteria for children referred/referred to services included in Stage 2 are:

1. Child (0-15 years of age) has learning disabilities as defined administratively or otherwise by the service
2. Child has been referred at least in part for support in relation to behaviours that challenge (including those referred to the relevant learning disability/behaviours that challenge pathway in any broader service)
3. Child's parental caregiver consents to take part in the research
4. Child's parental caregiver is able to complete, by questionnaire or interview with a researcher, study outcome measures in English

5.3.2 Exclusion criteria

The exclusion criteria for services are:

1. Services that have been established for less than one year, and so have had little chance to become reasonably stable within the current study timeframe
2. Services that have a typical referral/re-referral rate over 6 months of fewer than 10 children
3. Services that indicate at Stage 1 they would not wish to be contacted about involvement in Stage 2
4. Co-applicants Lovell and Liew's services – to address potential conflicts of interest

5.4 Recruitment

A STROBE flow diagram will be used to summarise the number of families who were:

- assessed for eligibility at screening
- eligible at screening
- ineligible at screening*
- eligible and randomised
- eligible but not randomised*
- lost to follow-up*
- number in each service model in the primary analysis

*reasons will be provided.

5.5 Withdrawal

Participants have the right to withdraw consent for participation in any aspect of the study at any time until data analysis commences. The participants' care will not be affected at any time by declining to participate or withdrawing from the study, or from an individual aspect (e.g., a qualitative interview).

If a participant initially consents but subsequently withdraws from the study, or an individual aspect of it, clear distinction must be made as to what aspect of the study the participant is withdrawing from. These aspects could be:

1. Withdrawal from all future follow-up assessments
2. Withdrawal from qualitative interview
3. Withdrawal from previously collected data
    a. Baseline assessments
    b. Follow-up assessments
    c. Qualitative interview
    d. All of the above
4. Withdrawal of consent to all of the above

Participants who consent and subsequently withdraw should complete the study withdrawal form or the withdrawal form should be completed on the participant's behalf by the Research Assistant/ study team member based on information provided by the participant. This withdrawal form should be sent to the MELD Study email address. Any queries relating to potential withdrawal of a participant should be forwarded to the Study Manager.

5.5.1 Timing of withdrawal

Participants have the right to withdraw consent for participation in any aspect of the trial at any time.

5.6 Lost to follow up

Family carer participants who do not complete the 12-month follow-up data collection will be considered lost to follow-up. The research team will send reminders before they are considered lost to follow-up. The schedule of reminders will be as follows:

One week before sending the questionnaire: A courtesy text/email to let them know to expect it

One week after sending the questionnaire: A reminder email/text

Two weeks after sending the questionnaire: A reminder telephone call, speaking to them if possible or leaving a voicemail

Three weeks after sending the questionnaire: A postal survey (with a minimum dataset version of the questionnaire for family carer participants)

After eight weeks from when the questionnaire was sent, if a response has not been received then a family will be considered lost to follow up.

5.6.1 Presentation of withdrawal/Loss to follow up

The number and % of participants that have withdrawn/loss to follow up from the study will be presented in a table for all stages. Results will be provided by service model, and where possible, we will further breakdown reasons for withdrawal. On the basis of these tabulated results, it may be necessary to include additional factors into our imputation models which will be transparently documented in the final report and any subsequent publication.

5.7 Baseline participant characteristics

5.7.1 List of baseline data

- Baseline demographic questionnaire completed including family living circumstances (including postal code, allowing coding of neighbourhood deprivation);

- gender,
- marital status,
- ethnic group,
- level of education,
- health/disability;
- parent health-related quality of life using the EQ-5D-5L;
- minimal information on recent resource use;
- number of adults and children in household;
- household income and financial hardship;
- where the child with LD lives during a normal week.

5.7.2 Descriptive statistics

Characteristics of each service model will be summarised descriptively. However, no formal statistical comparisons will be undertaken (Senn, 1994). Continuous measures will be reported as a mean, standard deviation (SD), minimum and maximum, while categorical data will be reported as a count and percentage.

6. Analysis

6.1 Outcome definitions

The following outcomes will be measured at baseline and at 12-months post-referral as a part of the observational study, unless otherwise stated:

6.1.1 Primary outcome measure

Child-related outcomes (all completed by the same family carers at baseline and follow up)

- Child BtC measure
  - Behaviour Problems Inventory Short Form (BPI-S) -TOTAL SCORE (Rojahn et al., 2012).

6.1.2 Timing, units, and derivation of primary

Primary outcome is collected at baseline and 12-months post referral. The BPI-S contains 30 items and three subscales including: Self Injurious Behaviour [SIB] (n=8), Aggressive Destructive Behaviour [ADB] (n=10), and Stereotyped Behaviour [SB] (n=12). The BPI-S rates behaviours (previous six months) on a five-point scale (never = 0; monthly = 1; weekly = 2; daily = 3; hourly = 4) and a three-point severity scale (mild = 1; moderate = 2; severe = 3). Each item is scored by multiplying the

frequency and severity scores together and then the sum of these item scores forms a subscale score. The sum of the three subscales gives a BPI-S total score (Bowring et al., 2018). Higher scores indicate more behaviours that challenge (BtC).

6.1.3 Secondary outcome measures

Child-related outcomes (all completed by family carers)

- Child physical health measures
  - Children's sleep habit questionnaire (CSHQ) (Owens et al., 2000)
  - Non-communicating children's pain checklist – revised (NCCPC-R) (Breau et al., 2004)
- Child mental health measure
  - Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997)
- Child quality of life measures
  - EQ-5D-Y Health Questionnaire (Wille et al., 2011) -
  - **EITHER** Paediatric Quality of Life Inventory (PedsQL) infant scales (*only for children aged 0-24 months*) (Varni et al., 2011) **OR**
  - Paediatric Quality of Life Inventory (PedsQL) Generic Core Scales (*only for children aged 2 years+*) (Varni et al., 2001)
- Child skills measure
  - GO4KIDDS (Perry et al., 2015)
- Child services received measure
  - Client service receipt inventory (CSRI) (Beecham & Knapp, 1992)

Family-carer-related outcomes (all completed by family carers)

- Family carer quality of life measure
  - EQ-5D-5L Health Questionnaire (van Hout et al., 2012)
- Family carer wellbeing measure
  - Warwick Edinburgh Mental Wellbeing Scale (WEMWBS) (Tennant et al., 2006)
- Family carer services received measure
  - Client service receipt inventory (CSRI) (Beecham & Knapp, 1992)
- Family carer experience/satisfaction with service (<u>at 12-months only</u>)
  - Experience of service questionnaire (ESQ) satisfaction with care items (Brown et al., 2014)

Service-level outcomes (<u>at 12-months only</u>) (completed by services)

- 'Reach' to population
- Timings of service delivery (e.g., mean time to starting support from referral, and length of time to discharge)
- Service take-up
- Costs at service level

Other-stakeholder-related outcomes (at 12-months only) (completed by external stakeholders)

- Satisfaction of external stakeholders (e.g., referrer, health, social-care, or education professionals) brief measure (*Developed for the purposes of this study*)

6.1.4 Timing, units, and derivation of secondary

Table 1. Descriptions of child or parent/carer reported outcomes

measures	Study timepoints		description	Scoring
	Baseline	12-month follow-up		
Behaviour Problems Inventory Short Form (BPI-S) [total frequency]	X	X	30 items; a five-point scale (never = 0; monthly = 1; weekly = 2; daily = 3; hourly = 4)	Summed score (Continuous) Higher scores indicate more behaviours that challenge (BtC)
Behaviour Problems Inventory Short Form (BPI-S) [Total severity]	X	X	30 items; a three-point severity scale (mild = 1; moderate = 2; severe = 3).	Summed score (Continuous) Higher scores indicate more behaviours that challenge (BtC)
Behaviour Problems Inventory Short Form (BPI-S) [subscale total score - aggression]	X	X	10 items; The BPI-S rates behaviours (previous six months) on a five-point scale (never = 0; monthly = 1; weekly = 2; daily = 3; hourly = 4) and a three-point severity scale (mild = 1; moderate = 2; severe = 3).	(Continuous) Each item is scored by multiplying the frequency and severity scores together and then the sum of these item scores forms a subscale score. Higher scores indicate
Behaviour Problems Inventory Short Form (BPI-S)	X	X	8 items; The BPI-S rates behaviours (previous	

[subscale total score - SIB]			six months) on a five-point scale (never = 0; monthly = 1; weekly = 2; daily = 3; hourly = 4) and a three-point severity scale (mild = 1; moderate = 2; severe = 3).	more behaviours that challenge (BtC)
Behaviour Problems Inventory Short Form (BPI-S) [subscale total score - stereotypy]	X	X	12 items; The BPI-S rates behaviours (previous six months) on a five-point scale (never = 0; monthly = 1; weekly = 2; daily = 3; hourly = 4) and a three-point severity scale (mild = 1; moderate = 2; severe = 3).	
Child Children's sleep habit questionnaire (CSHQ)	X	X	45-items; 3-point scale per item.	Summed score (continuous). higher score indicates more disturbed sleep
Child Non-communicating children's pain checklist (NCCPC-R)	X	X	30 items (7 subscales). Each item is a 4-point scale (0-3)	Summed score (continuous). Higher score indicates greater pain level
Child Strengths and Difficulties Questionnaire (SDQ)	X	X	25 items (5 subscales of 5 items) Each item is a 3 point scale (0-2)	Summed score (continuous, potential skew). scores can be scaled up pro-rata if at least 3 items were completed higher score indicates more problems
Child EQ-5D-Y Health Questionnaire – total utility score	X	X	5 main questions Each item is a 3 point scale (1-3)	a continuous variable between -0.594 and 1.0

Child **Either** Paediatric Quality of Life Inventory (PedsQL) Generic Core Scales **or** Paediatric Quality of Life Inventory (PedsQL) infant scales	X	X	23 items Each item is a 6 point scale (0-5)	Items are reverse scored and linearly transformed to a 0–100 scale (0 5 100, 1 5 75, 2 5 50, 3 5 25, 4 5 0), so that higher scores indicate better HRQOL.
Child GO4KIDDS	X	X	8 items Each item is a 5-point scale	Summed score (continuous, max =40) Higher scores indicate greater skill level and independence
Family carer Warwick Edinburgh Mental Wellbeing Scale	X	X	14-items; Each item is a 5-point scale	Summed score (continuous)
Family carer Experience of service questionnaire (ESQ)		X	12 items; 3-point scale for each item (1-3)	Summed scores (continuous) *Satisfaction with Care* (sum items 1, 2, 3, 4, 5, 6, 7, 11 and 12) *Satisfaction with Environment* (sum items 8, 9 and 10)
Child Client service receipt inventory (CSRI)	X	X	This measure will be used in the health economic analysis, so details can be found in the HEAP.	
Family carer Client service receipt inventory (CSRI)	X	X	This measure will be used in the health economic analysis, so details can be found in the HEAP.	

Table 2. Descriptions of service outcomes (completed by services or stakeholders)

measures	**Study timepoints**		**description**
	Baseline	**12-month follow-up**	

'Reach' to population		X	
Timings of service delivery		X	mean time to starting support from referral, and length of time to discharge
Service take-up		X	
*Costs at service level		X	This measure will be used in the health economic analysis, so details can be found in the HEAP.
Satisfaction of external stakeholders		X	6-point scale for each item, then total score for satisfaction. Higher score, more satisfied.

6.1.5 Order of testing

Secondary outcomes are tested in the order listed in section 6.1.3.

6.2 Analysis Methods

6.2.1 List of methods of presentation

Directed Acyclic Graph (DAG)

It is inevitable that the referral of children to different service types will be a non-random process. For example, children live in different locations in England where different approaches to services are in place. In addition, services may have varying inclusion criteria (e.g., re severity of behaviours that challenge). Therefore, a crude comparison of outcomes of children in different service types will be prone to selection bias. To minimise this, the research team and stakeholders (via PPI consultation – to be completed) have constructed a Directed Acyclic Graph (DAG) prior to starting data collection in Stage 2, that outlines our core assumptions around the causal pathways between service model (exposure) and child outcomes (Williams et al., 2018). The DAG identifies key quantitative data to collect (e.g., confounders, mediators, and moderators) and is as an aid for modelling the relationship between service model and child outcomes. This process is designed to

ensure that additional variables that may have an influence on the direct relationship between the outcome and services models are considered and, if required, adjusted for.

Based on the current DAG (Figure 1), the required adjustment set of variables will include the following: Barriers or ability to access services, children with BtC on referral, child's age, child complexity (a predefined composite score or latent variable), family-level deprivation (also a composite measure), and family wellbeing. No mediator variables were identified within the construction of the DAG (including in discussions with professional and family carer advisory groups) and mediation in this study is not directly of interest to address our primary research questions. The adjustment set will inform the propensity score calculation, but we will not adopt a 'doubly Robust' approach (i.e. using propensity scores and covariates adjustment; Elze et al., 2017) given our limited sample size.

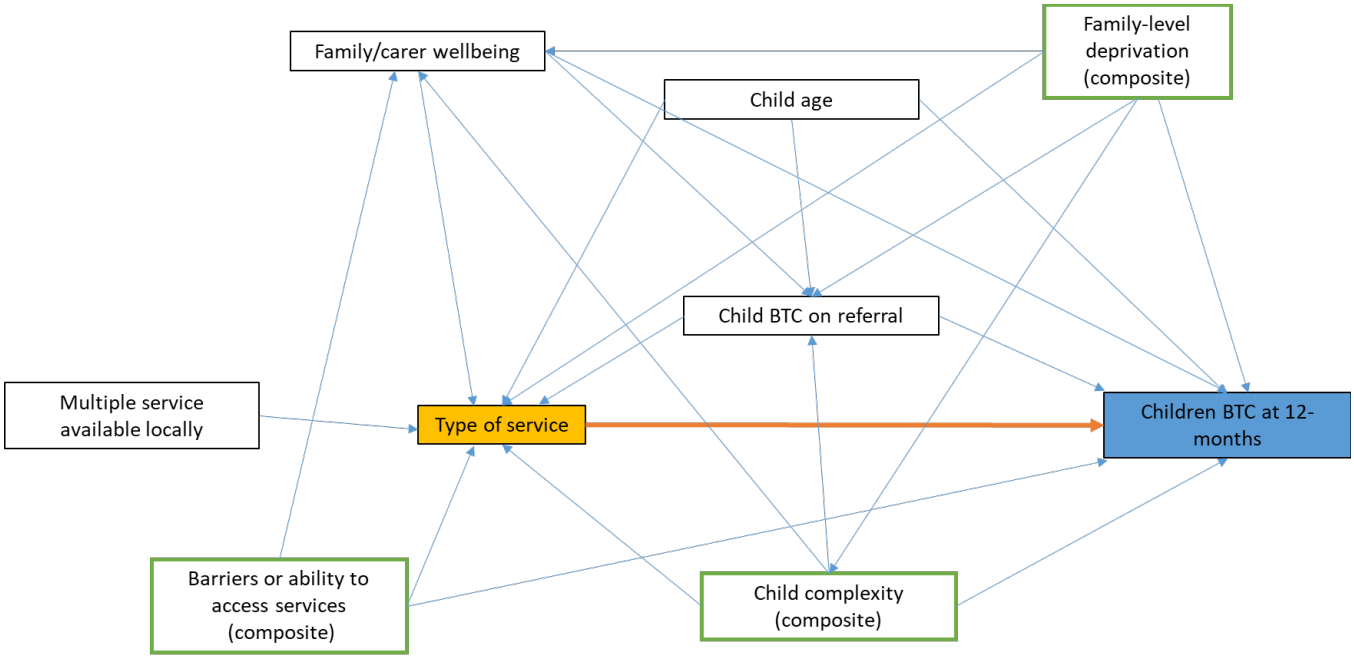


Figure 1: Directed Acyclic Graph for MELD stage 2

The DAG included three composite measures that incorporate information from several measured variables. The composites are used to collate information from observed instruments that measure similar things and create a more informative instrument while also saving on statistical degrees of freedom. The composite scores will be constructed as summed scores of the raw observed scores. Validity of composite scores will be evaluated by calculation of Cronbach's Alpha and a confirmatory factor analysis on items within each composite to ensure that they sit together. Factor analyses will report the following fit indices and thresholds for good model fit: CFI >0.9, TLI >0.9, RMSEA <0.08, and SMR <0.08 (Wang & Wang, 2012). If confirmatory factor analyses fail to converge or provide poor fit to the data, we will identify if specific items are problematic or if model fit can be resolved

by checking the following: i) any items have low loadings on specific factors; ii) whether correlating item residual covariances improves fit; iii) whether certain items need transformation, i.e. collapsing low frequency categories or high skew; iv) check for Heywood cases. It may be that we refine the factor structure by splitting single factors into two if fit indices indicates an improvement. However, we will justify any change in factor structure to ensure sensible and relevant constructs are formed. It may also be the case that some less informative items are dropped from the analysis if necessary and will be justified accordingly. Figure 2 presents the contents of the composite measures.

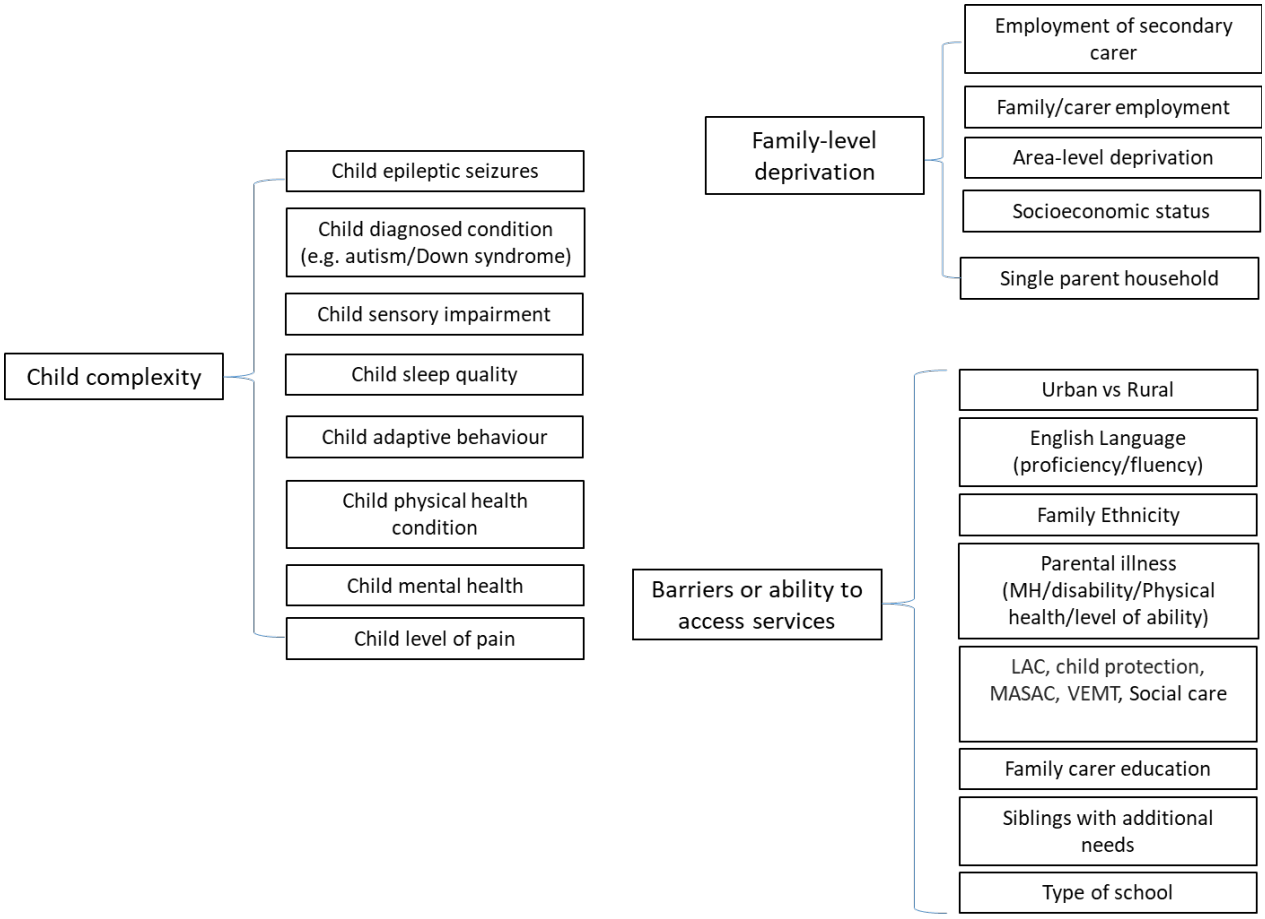


Figure 2: Contents of composite measures: Child complexity, Barriers or ability to access services, and Family-level deprivation

Primary outcome analysis (Two level model specification)

The statistical analysis is specified as a two-level linear regression model (children nested within services), with the child outcome measure at 12-months as the primary outcome, service model included as a fixed effect, and the corresponding child outcome collected at baseline also included as a covariate. Results will be reported as mean differences adjusted for the baseline measure of the outcome, 95% confidence intervals, and p-values. The formal two-level linear model can be specified as follows,

$$\boldsymbol{L1}: Y_{ij} = \beta_{0j} + \beta_1 X1_{ij} + \beta_2 X2_{ij} + r_{ij},\ r_{ij} \sim N\left(0, \sigma^2{}_{|X}\right)$$
$$\boldsymbol{L2}: \beta_{0j} = \gamma_{00} + \gamma_{01} service_model_j + \mu_{0j}, \qquad \mu_{0j} \sim N(0, \tau_2^2)$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$

where, Y are the outcomes; $X1$ are the baseline outcomes values; $service_model$ is the indicator of service model; $X2$ is an additional covariate. μ_{0k} is the random intercept term for service.

Research question 1 (primary and secondary) will be evaluated using the main effect contrasts of 'service models' categories. Two level models will use treatment contrasts with service model, 'Specialist behaviours that challenge services (*BtCExpert*)', as the reference category. Analyses will control the family-wise error rate using the Holm method.

Research question 2 (primary and secondary) will report Tukey's contrasts for pairwise differences between service models and rank them accordingly.

Research question 3 (primary and secondary) will be addressed by reporting multiple comparisons of means (Tukey contrasts) for the pairwise comparison for LD CAMHS vs generic CAMHS, from the primary outcome model addressing Research Question 1.

Propensity score calculation

When fitting models to describe the relationship between our service types and our outcome/s, we will adjust for confounding using propensity score weighting, as this approach tends to preserve the majority of a study sample and therefore maximises precision of estimates. Specifically, by weighting participants' responses based on their propensity score, we aim to control for any biases present that may confound the relationship between service type and outcome (Olmos & Govindasamy, 2015). The multinomial propensity scores for multiple treatments, or service types in this case, are calculated using a multinomial logistic regression (Spreeuwenberg *et al.,* 2010) that include potential confounders as independent variables and the service class as the dependent variable (treatment index). When specifying our propensity score model, we will exclude strong predictors of service model inclusion that are not associated with outcome. The propensity scores will be calculated using the R packages '*PSweight*' and '*WeightIt*'.

Propensity scores will be visually checked for balance by plotting boxplots of scores by service class. We will consider trimming of extreme weights (i.e., probabilities close to zero or one), following the symmetric trimming rules in Crump et al. (2009). We will then estimate the "average treatment

effect in the whole population" (i.e., the effect of service model on outcomes in all referred children) by fitting models with inverse probability of treatment weights and robust sandwich-type estimators for variance estimation. In the event that models including the inverse probability weights do not converge, we will explore other approaches to resolve (i.e. direct adjustment as the number of covariates is relatively small or stratification).

Secondary outcome analysis

All secondary outcome measures will be analysed following the same model setup as the primary outcome. Model assumptions will be checked (section 6.2.3) and any outcomes that do not meet required assumptions will change to an appropriate alternate link function (using a Generalized linear mixed model, section 6.2.4).

Exceptions are the *experience of service questionnaire (ESQ)* measure and external stakeholder satisfaction measure that are only measured at the 12 month follow up, so no baseline adjustment can be included in the model. For the external satisfaction measure, data will be collected anonymously within service, but information on which service the professional belongs will be collected, so that clustering by service can be accounted for using random intercepts within the analysis.

All other service level outcomes, except costs, will be analysed descriptively. These outcomes will be estimated as frequencies and percentages, means and standard deviations, or medians and interquartile ranges as appropriate.

6.2.2 Covariate adjustment

The Directed Acyclic Graph (DAG) process maps all known covariates and determines a minimum adjustment set. All required covariates will be added to the main analysis (see section 6.2.1).

6.2.3 Assumption checking

1. Linearity – plotting residuals vs predictor(s). If a structure is present, then transformation or an alternate model specification is required (i.e. GLM).

2. Homogeniety of variance – variance of the residuals across groups is the same. There is scope to fit models allowing for heterogeneous groups, but the setup is different (Generalized linear mixed model - GLMM).

3. Residuals are approximately normally distributed – plotting QQ plot

6.2.4 Alternative methods if distributional methods not met

If distributional assumptions are not satisfied, as appropriate, a generalized linear model with alternate link function will be used. Alternatively, data transformation could be used but use of the GLM is preferable.

6.2.5 Sensitivity analyses

In addition to the main DAG analysis, we will also present the unadjusted analysis omitting just the minimal adjustment set. The model will be setup following a reduced form to that specified in section 6.2.1.

If we find that the Vineland (VABS-II; Vineland-II adaptive behavior scales) has significant missing data, we will re-run analyses substituting the Go4Kidds measure as a proxy. The correlation between these measures is fairly high at r=0.81 (Pan et al., 2019).

6.3 Missing data

Exploring the impact of missing data on trial outcomes by investigating likely missing data mechanisms and re-fitting the primary outcome within a multiple imputation framework (including exploring MAR and MNAR mechanisms via delta-based controlled multiple imputation).

We will summarise the extent of missing data in all outcomes and their respective control variables. A full multiple imputation strategy will be used if more than 5% of data in the primary model is missing. Alternatively, we will impute if more than 10% of data for a single variable is missing.

Following Granger et al (2019),we will use the multiple imputation by chained equations approach via the mice package in R (Van Buuren and Groothuis-Oudshoorn, 2011) and generate 20 imputed datasets prior to calculation of propensity scores. We will then calculate the propensity scores for each imputed data set and then estimate the intervention effect for each imputed dataset, pooling the results using Rubin's combination rules for standard errors. These estimates will be presented for comparison to the primary analyses results (unimputed data).

6.4 Statistical software

All statistical analyses will use R version 4.1.2 (2021-11-01) with additional packages: tidyverse, VGAM, lme4, lmerTest, performance, mice, psych, ivreg, and ivpack.

7 References

Adams, G., Gulliford, M.C., Ukoumunne, O.C., Eldridge, S., Chinn, S., Campbell, M.J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol.* 57(8), 785-94. doi: 10.1016/j.jclinepi.2003.12.013.

Adams, D., Rose, J., Jackson, N., Karakatsani, E., Oliver, C (2018). Coping Strategies in Mothers of Children with Intellectual Disabilities Showing Multiple Forms of Challenging Behaviour: Associations with Maternal Mental Health. *Behav Cogn Psychother*,46(3):257-275. doi: 10.1017/S1352465817000704. Epub 2017 Nov 29. PMID: 29183406.

Baker, B.L., McIntyre, L.L., Blacher, J., Crnic, K., Edelbrock, C., & Low, C. (2003). Pre-school children with and without developmental delay: Behaviour problems and parenting stress over time. Journal of Intellectual Disability Research, 47, 217– 230.

Balboni, G., Rebecchini, G., Elisei, S., & Tassé, M. J. (2020). Factors affecting the relationship between adaptive behavior and challenging behaviors in individuals with intellectual disabilities and co-occurring disorders. Research in Developmental Disabilities. https://doi.org/10.1016/j.ridd.2020.103718

Barnett, A.G., & Dobson, A.J. (2008). An Introduction to Generalized Linear Models (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315182780

Beck A., Hastings R. P., Daley D. & Stevenson J. (2004) Pro-social behaviour and behaviour problems independently predict maternal stress. *Journal of Intellectual and Developmental Disability* 29, 339–349.

Bowring, D. L., Totsika, V., & Hastings, R. P. (2019). Designing specialist community-based behavioural support teams. *International Journal of Positive Behavioural Support*, 9(2), 4-15.

Bowring, D.L., Totsika, V., Hastings, R.P. and Toogood, S. (2018), Toward data-based clinical decision making for adults with challenging behavior using the Behavior Problems Inventory-Short Form (BPI-S), Tizard Learning Disability Review, Vol. 23 No. 2, pp. 103-110. https://doi.org/10.1108/TLDR-06-2017-0025

Bowring DL, Totsika V, Hastings RP, Toogood S. (2020). Outcomes from a community-based Positive Behavioural Support team for children and adults with developmental disabilities. *J Appl Res Intellect Disabil.*, 33(2):193-203. doi: 10.1111/jar.12660. Epub 2019 Aug 23. PMID: 31441576.

Brown, A., Ford, T., Deighton, J. & Wolpert, W. (2014) Satisfaction in Child and Adolescent Mental Health Services. Administration and Policy in Mental Health and Mental Health Services Research. 41 (4), pp. 436-446.

Candlish, J., Teare, M.D., Dimairo, M. et al. Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: a simulation study. BMC Med Res Methodol 18, 105 (2018). https://doi.org/10.1186/s12874-018-0559-x

Chadwick, O., Cuddy, M., Kusel, Y., & Taylor, E. (2005). Handicaps and the development of skills between childhood and early adolescence in young people with severe intellectual disabilities. Journal of Intellectual Disability Research, 49(12), 877–888. https://doi.org/10.1111/j.1365-2788.2005.00716.x

Crump, R.K, Hotz, V.J., Imbens, G.W. , & Mitnik, O.A. (2009). Dealing With Limited Overlap in Estimation of Average Treatment Effects. Biometrika, 96(1):187–199. doi: 10.1093/biomet/asn055. [p1, 5, 10]

Davies LE, Oliver C. Self-injury, aggression and destruction in children with severe intellectual disability: Incidence, persistence and novel, predictive behavioural risk markers. Res Dev Disabil. 2016 Feb-Mar;49-50:291-301. doi: 10.1016/j.ridd.2015.12.003. Epub 2016 Jan 5. PMID: 26765248.

Eldridge SM, Ashby D, & Kerry S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*, 35(5):1292-300. doi: 10.1093/ije/dyl129. Epub 2006 Aug 30. PMID: 16943232.

Elze, M.C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., ... & Pocock, S.J. (2017). Comparison of propensity score methods and covariate adjustment: Evaluation in 4 cardiovascular studies. Journal of the American College of Cardiology, 69, 345–357.

Estes A., Olson E., Sullivan K., Greenson J., Winter J., Dawson G. *et al*. (2013) Parenting-related stress and psychological distress in mothers of toddlers with autism spectrum disorders. *Brain and Development* 35, 133–138.

Fernández-Ávalos, M. I., Pérez-Marfil, M. N., Ferrer-Cascales, R., Cruz-Quintana, F., Clement-Carbonell, V., & Fernández-Alcántara, M. (2020). Quality of Life and Concerns in Parent Caregivers of Adult Children Diagnosed with Intellectual Disability: A Qualitative Study. International journal of environmental research and public health, 17(22), 8690. https://doi.org/10.3390/ijerph17228690;

Granger, E., Sergeant, J. C., & Lunt, M. (2019). Avoiding pitfalls when combining multiple imputation and propensity scores. Statistics in medicine, 38(26), 5120–5132. https://doi.org/10.1002/sim.8355

Hastings, R. P. (2016). Do children with intellectual and developmental disabilities have a negative impact on other family members? The case for rejecting a negative narrative. International Review of Research in Developmental Disabilities, 50, 165-194.

Hastings R. P. (2002) Parental stress and behaviour problems of children with developmental disability. *Journal of Intellectual and Developmental Disability* 27, 149–160. https://doi.org/10.1348/026151003321164672

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. Journal of Educational and Behavioral Statistics, 32(4), 341–370. https://doi.org/10.3102/1076998606298043

Johnston C., Hessl D., Blasey C., Eliez S., Erba H., Dyer-Friedman J. *et al*. (2003) Factors associated with parenting stress in mothers of children with fragile X syndrome. *Journal of Developmental and Behavioral Pediatrics* 24, 267–275.

Jones, Leah, Hastings, Richard P., Totsika, Vasiliki, Keane, Lisa and Rhule, Neisha. (2014) Child behavior problems and parental well-being in families of children with autism : the mediating role of mindfulness and acceptance. American Journal on Intellectual and Developmental Disabilities, Volume 119 (Number 2). pp. 171-185.

Kiernan, C. & Kiernan, D. (1994). Challenging behaviour in schools for pupils with severe learning difficulties. *Mental Handicap Research, 7* (3), 177-201.

Kurtz, P. F., Leoni, M., & Hagopian, L. P. (2020). Behavioral Approaches to Assessment and Early Intervention for Severe Problem Behavior in Intellectual and Developmental Disabilities. Pediatric clinics of North America, 67(3), 499–511. https://doi.org/10.1016/j.pcl.2020.02.005

Langley, E., Totsika, V., & Hastings, R. P. (2020). Psychological well-being of fathers with and without a child with intellectual disability: a population-based study. Journal of intellectual disability research : JIDR, 64(6), 399–413. https://doi.org/10.1111/jir.12692

Marquis S., Hayes M. V. & McGrail K. (2019) Factors affecting the health of caregivers of children who have an intellectual/developmental disability. Journal of Policy and Practice in Intellectual Disabilities 16.

Mazurek, M. O., Kanne, S. M. &Wodka, E. L. (2013). Physical aggression in children and adolescents with autism spectrum disorders. *Research in Autism Spectrum Disorders, 7,* 455-465. http://dx.doi.org/10.1016/j.rasd.2012.11.004

Murphy, G. H., Beadle-Brown, J., Wing, L., Gould, J., Shah, A. & Holmes, N. (2005). Chronicity of challenging behaviours in people with severe intellectual disabilities and/or autism: a total population sample. *Journal of Autism and Developmental Disorders, 35* (4), 405 – 418, DOI:10.1007/s10803-005-5030-2

Murphy, O., O. Healy, and G. Leader. 2009. “Risk Factors for Challenging Behaviour among 157 Children with Autism Spectrum Disorder in Ireland.” Research in Autism Spectrum Disorders 3: 474–482.

Neece C. & Baker B. (2008) Predicting maternal parenting stress in middle childhood: the roles of child intellectual status, behaviour problems and social skills. *Journal of Intellectual Disability Research* 52, 1114–1128.

McClintock, K., Hall, S. and Oliver, C. (2003), Risk markers associated with challenging behaviours in people with intellectual disabilities: a meta-analytic study. Journal of Intellectual Disability Research, 47: 405-416. https://doi.org/10.1046/j.1365-2788.2003.00517.x

Olmos, A., & Govindasamy, P. (2015). A Practical Guide for Using Propensity Score Weighting in R. Practical Assessment, Research and Evaluation, 20, 13.

Orsmond, G.I., Seltzer, M.M., Krauss, M.W., & Hong, J. (2003). Behavior problems in adults with mental retardation and maternal well-being: Examination of the direction of effects. American Journal on Mental Retardation, 108, 257– 271.

Pan, X., Totsika, V., Nicholls, G., & Paris, A. (2019). Validating GO4KIDDS as a brief measure of adaptive skills in special education settings for children with severe intellectual disability. Journal of applied research in intellectual disabilities : JARID, 32(2), 280–287. https://doi.org/10.1111/jar.12524

Resch JA, Benz MR, Elliott TR. Evaluating a dynamic process model of wellbeing for parents of children with disabilities: a multi-method analysis. Rehabil Psychol. 2012 Feb;57(1):61-72. doi: 10.1037/a0027155. PMID: 22369118.

Ricci L. A. & Hodapp R. M. (2003) Fathers of children with Down's syndrome versus other types of intellectual disability: perceptions, stress and involvement. *Journal of Intellectual Disability Research* 47, 273–284.

Sapiets, S.J., Totsika, V. and Hastings, R.P. (2021), Factors influencing access to early intervention for families of children with developmental disabilities: A narrative review. J Appl Res Intellect Disabil, 34: 695-711. https://doi.org/10.1111/jar.12852

Senn, S. (1994), Testing for baseline balance in clinical trials. Statist. Med., 13: 1715-1726. https://doi.org/10.1002/sim.4780131703

Spreeuwenberg MD, Bartak A, Croon MA, Hagenaars JA, Busschbach JJ, Andrea H, Twisk J, Stijnen T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. Medical Care, 48(2), 166–174.

Taylor, E. et al. (in prep), Mapping and Evaluating Services for Children with Learning Disabilities and Behaviours that Challenge (MELD): Stage 1 - Statistical and descriptive analysis to identify distinct service models.

Tomanik S., Harris G. E. & Hawkins J. (2004) The relationship between behaviours exhibited by children with autism and maternal stress. *Journal of Intellectual and Developmental Disability* 29, 16–26.

Totsika, V., Hastings, R. P., Emerson, E., & Hatton, C. (2019). Early years parenting mediates early adversity effects on problem behaviors in intellectual disability. Child Development, 91(3), 649–664.Totsika V., Hastings R. P., Emerson E., Berridge D. M. & Lancaster G. A. (2015) Prosocial skills in

young children with autism, and their mothers' psychological well-being: longitudinal relationships. *Research in Autism Spectrum Disorders* 13, 25–31.

Totsika, V., Hastings, R. P., Emerson, E., Lancaster, G., & Berridge, D. (2011). A population-based investigation of behavioural and emotional problems and maternal mental health: Associations with autism spectrum dis-orders and intellectual disability. *Journal of Child Psychology and Psychiatry*.,52,91–99. https://doi.org/10.1111/j.1469-7610.2010.02295.x.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI: 10.18637/jss.v045.i03.

Wang, J., & Wang, X. (2012). Structural equation modeling: Applications using Mplus: John Wiley & Sons.

Williams, T.C., Bach, C.C., Matthiesen, N.B., Henriksen, T.B., & Gagliardi, L. (2018). Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res, 84(4). 487-493. doi: 10.1038/s41390-018-0071-3.

Willner P, Goldstein RC. Mediation of depression by perceptions of defeat and entrapment in high-stress mothers. Br J Med Psychol. 2001 Dec;74(Pt 4):473-85. doi: 10.1348/000711201161127. PMID: 11780795.