

Case Study 1: An Evidence Based Practice Review Report

Theme: School based interventions for children and young people with special educational needs (SEN)

How effective is mentalization based therapy/treatment (MBT) at reducing psychological distress in children and young people?

Table of Contents

| | |
|--|----|
| Summary..... | 2 |
| Introduction | 3 |
| What is MBT? | 3 |
| Relevance of MBT to Educational Psychological Practice..... | 5 |
| Critical Review of the Evidence Base..... | 6 |
| Literature Search and Screening Strategy..... | 6 |
| Weight of Evidence (WoE)..... | 12 |
| Participants..... | 13 |
| Research Design | 15 |
| Setting Relevance – Educational vs. Clinical Significance..... | 16 |
| MBT Interventions..... | 18 |
| Outcome Measures | 20 |
| Conclusions and Recommendations..... | 29 |
| References..... | 32 |
| Appendices | 36 |
| Appendix A - Studies Excluded from Systematic Review with Reasons..... | 36 |
| Appendix B - Summary of Included Studies and Effect Sizes..... | 47 |
| Appendix C - Weight of Evidence Information | 54 |
| Appendix D - Coding Protocols..... | 67 |

Summary

Mentalization-based therapy/treatment (MBT) is an integrative approach that draws from a wide range of theories, but predominantly from psychodynamic and attachment theory (Bateman & Fonagy, 2004). It aims to develop mentalization, which is the ability to interpret behaviours based on mental states of one's self and others, and understand intentions via psychotherapy (Bateman & Fonagy, 2004). This systematic literature review aimed to evaluate the evidence for MBT based interventions in alleviating psychological distress in children and young people with Special Educational Needs and Disabilities (SEND). The review identified seven studies that used MBT as an intervention to support children and young people experiencing psychological distress. There was heterogeneity in terms of research designs, how MBT were implemented, and the outcome measures used. However, the research suggests that MBT shows promise in helping children and young people experiencing psychological distress. Recommendations are made regarding how future research can further develop the evidence base of MBT for children and young people.

Introduction

What is MBT?

MBT is a psychotherapy that uses psychodynamic approaches to support individuals who have difficulties with mentalizing (Bateman & Fonagy, 2004). Mentalizing involves the processes of introspection; perceiving oneself from the outside and others from the inside; and interpreting human behaviours as intentional human states e.g. thoughts, feelings, beliefs, desires, and reasons (Bateman & Fonagy, 2006). Some individuals have difficulties mentalizing which can include a lack of self-awareness, poor theory of mind, and lack the capacity to makes sense of misunderstandings. Mentalizing has been described as the ability to develop stable internal self-representations and secure relationships (Bateman & Fonagy, 2004). For this reason, MBT can also involve family therapy and group formats. Treatment can support individuals to learn from conscious or unconscious attempts to understand mental states in themselves and others (Bateman & Fonagy, 2006). Treatment can also support individuals to 'differentiate and separate out their own thoughts and feelings from those around them' (Grohol, 2016, p1).

In accordance to mentalization theory, psychological distress can be related to attachment (Bateman & Fonagy, 2004). The quality of soothing responses by care givers can influence the development of a child's sense of safety in the world, their own internal working model of relationships, and ability to mentalize (Bateman & Fonagy, 2004). If a care giver repeatedly gives effective soothing responses, a child can internalise caregivers self-soothing responses to self-regulate their own emotions, feel safe to mentalize about their own and other people's state of minds, and develop a secure attachment style (Bateman & Fonagy, 2004). In contrast, if a caregiver does not give effective soothing responses, a child can struggle to self-

regulate their own emotions, find it unsafe to mentalize about their own and other people’s state of minds, and develop an insecure attachment style (Bateman & Fonagy, 2004).

MBT therefore aims to develop a therapeutic attachment based relationship, to help participants to develop their capacity to mentalize (Bateman & Fonagy, 2006). This involves the exploration of their own and other people’s mental states, in order to develop an internal working model of themselves and others that is more rational and coherent (Bateman & Fonagy, 2006). By increasing participants’ ability to mentalize, this can help participants to understand their own and other people’s minds more effectively, and therefore regulate their emotions and distress (Bateman & Fonagy, 2006).

The features of MBT can be seen in Table 1, which revolves around the exploration of mental states from different perspectives (Bateman & Fonagy, 2013). The aim of these steps is to facilitate mentalization so that individuals can be more curious about other people’s mental states and develop greater understanding of their own mental states (Bateman & Fonagy, 2013). The focus is on exploration of mentalization in the present as opposed to past experiences (Bateman & Fonagy, 2013). With this exploration, the aim is to maintain a moderate level of emotional states (Bateman & Fonagy, 2013).

Table 1

The Features of a MBT Intervention (Bateman & Fonagy, 2013)

| Therapeutic Steps | Features |
|-------------------|---|
| 1. | Demonstrating empathy with the patient’s current subjective state |
| 2. | Exploration, clarification and, if appropriate, challenge |
| 3. | Identifying affect and establishing an affect-focus |
| 4. | Mentalizing the relationship |

Bateman and Fonagy (2013) also discussed the therapist stance. This includes taking a non-expert position, being patient in order to explore differences in perspectives, accepting different perspectives and exploration of experiences. This therapist stance is important in order to explore the process of mentalizing within a safe attachment relationship, and this relationship is considered to be important for therapeutic change to occur (Bateman & Fonagy, 2013).

Relevance of MBT to Educational Psychological Practice

With the improvement of children's and young people's mental health a priority of national agenda, educational settings are focused on improving the social emotional mental health (SEMH) of children and young people (Public Health England, 2016).

Improving mentalization processes may be important, considering this is related to understanding relationships, understanding of the self, motivation and flexibility. All of which informs how one acts and responds in the world. Indeed, research has shown that SEMH may be related to deficits in mentalization (Fonagy & Bateman, 2011). There is a growing evidence base for MBT, where much of the research has been conducted on adults (Bateman and Fonagy, 2013). MBT is being shown to be effective for a wide range of difficulties including borderline personality disorder (BPD), antisocial personality disorder, substance abuse, eating disorders, at-risk mothers with infants and children, and other comorbidities (Bateman & Fonagy, 2013). Mentalization deficit is considered to be central to the pathology of people with a diagnosis of borderline personality disorder (Bateman & Fonagy, 2013). Furthermore, neuro-imaging studies demonstrated the effects of MBT on areas of the brain that is related to higher order thinking (Fonagy et al., 2005). It is

also being used with families and adolescents, in schools, and in managing social groups (Asen & Fonagy, 2011).

Educational psychologists can be a key profession in improving outcomes for children with SEMH given their specialist training and role. Their training gives them expertise to critically consume research and apply psychological theories to practice (British Psychological Society, 2015). Their role involves individual casework and systemic work to improve SEMH outcomes in children, which is in line with the Special Educational Needs Code of Practice (Department for Education and Department of Health, 2015). Given the evidence base for MBT, MBT may potentially be helpful for children with SEMH difficulties. Educational psychologists have the skill set to determine whether this is the case and can apply MBT in practice when working with schools.

Considering the need to improve children's mental health in educational settings, and the growing evidence base for MBT with adult populations, it is important to review how MBT can be effective for children and young people. Therefore, the literature review question is, how effective is MBT at reducing psychological distress in children and young people?

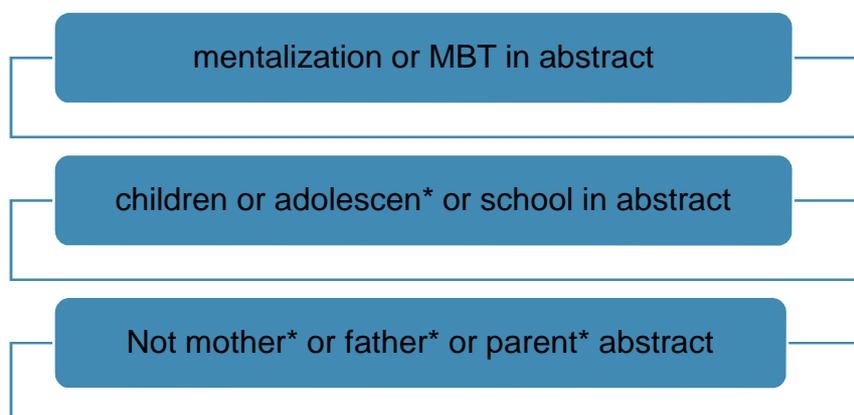
Critical Review of the Evidence Base

Literature Search and Screening Strategy

Preliminary searches led to the development of search terms used in this review. The preliminary search identified that a large proportion of research focused on the effectiveness of MBT for parents. As the focus of this review is on children and young people and not parents, search terms were used to exclude research that focused on parents.

A literature search was carried out in January and February 2018 using electronic databases. Databases were selected in Medline, British Education index, ERIC and PsychINFO. The following search terms were selected:

Figure 1. Online Database Search Terms



In addition to the search terms the following parameters were applied; English; and empirical studies. From these 50 studies were found with no duplicates. Abstracts were screened according to inclusion and exclusion criteria (Table 2) and 41 articles did not meet criteria and were excluded (see Appendix A for rationale of exclusion). Nine remaining articles were read in full and two were excluded. A PRISMA flow chart representing the data screening can be seen in Figure 2.

Seven studies were included for review and were subsequently evaluated using Gough's (2007) Weight of Evidence Framework. Group Designs were evaluated using Gersten et al. (2005) coding protocol and Small N Designs evaluated using Horner et al. (2005) coding protocol.

Table 2

Inclusion and Exclusion Criteria

| Criterion | Criteria for Inclusion | Criteria for Exclusion | Rationale |
|---------------------|---|--|--|
| 1. Language | Publication in English language. | Publication not in English language. | Reviewer needs to be able to comprehend the text. Also, lack of translating resources. |
| 2. Publication Type | Published in a journal. | Not published in a journal e.g. dissertations, books. | Involves editorial approval which maintains the quality of studies accepted. |
| 3. Type of Study | Studies that collected primary empirical data with pre and post measures. | Study that did not collect primary empirical data e.g. book reviews, descriptive articles, literature reviews, studies investigating the mechanisms and process. | The review is interested in the effectiveness of MBT. |
| 4. Intervention | Intervention based on MBT. | Intervention that was not informed by MBT. MBT intervention that was targeted towards adults. | The review aims to investigate the effectiveness of MBT for children and young people. |

| Criterion | Criteria for Inclusion | Criteria for Exclusion | Rationale |
|----------------------|---|---|--|
| 5. Participants | <p>Participants from studies are 18 years or under.</p> <p>Participants are identifiable with SEND.</p> | <p>Participants described as over 19.</p> <p>Participants are not identifiable with SEND.</p> | <p>The effectiveness of MBT for children and young people. The UK legal age of adulthood is 18 (Office for National Statistics, 2017) this is the cut-off point to capture school aged participants and adolescents.</p> <p>High levels of SEND within children and young people, highlighting the area of need (Public Health England, 2016).</p> |
| 6. Outcome variables | <p>Outcome measures relevant to psychological distress e.g. depression, anxiety, aggression, BPD and self-harm.</p> | <p>Outcome variables do not measure psychological distress e.g. physical health outcomes.</p> | <p>Current topical concerns regarding reducing SEMH difficulties in the children and young people population.</p> |

Figure 2. Study Selection Process – PRISMA Flow Chart

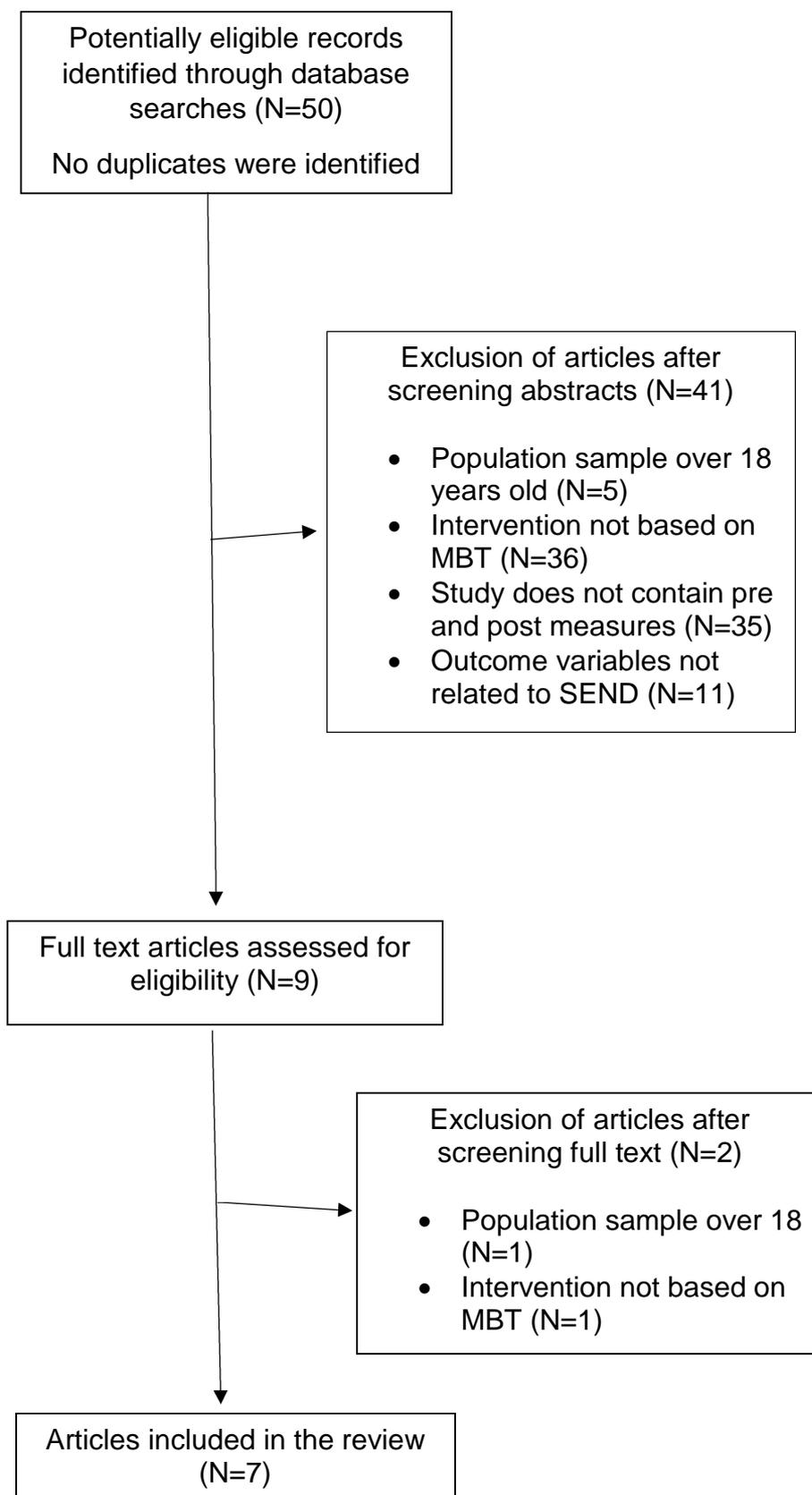


Table 3

Final Studies Included in the Systematic Review

| Final Studies |
|--|
| Fonagy, P, Twemlow, S, W., Vernberg, E, M., Nelson, J, M., Dill, E, J., Little, T, D., & Sargent, J, A. (2009). A Cluster Randomized Controlled Trial of Child-Focused Psychiatric Consultation and a School Systems Focused Intervention to Reduce Aggression. <i>Journal of Child Psychology and Psychiatry</i> , 50 (5): 607–16. https://doi.org/10.1111/j.1469-7610.2008.02025 |
| Halfon, S., Bekar, O., & Gürleyen, B. (2017). An Empirical Analysis of Mental State Talk and Affect Regulation in Two Single-Cases of Psychodynamic Child Therapy. <i>Psychotherapy</i> , 54 (2): 207–19. https://doi.org/10.1037/pst0000113 |
| Hauber, K., Boon, A, E., & Vermeiren, R. (2017). Examining Changes in Personality Disorder and Symptomology in an Adolescent Sample Receiving Intensive Mentalization Based Treatment: A Pilot Study. <i>Child and Adolescent Psychiatry and Mental Health</i> , 11 (58): 1-7. https://doi.org/10.1186/s13034-017-0197-9 |
| Laurensen, E, M, P., Hutsebaut, J., Feenstra, D, J., Bales, D, L., Noom, M, C., Busschbach, J, J, V., Verheul, R., & Luyten, P. (2014). Feasibility of Mentalization-Based Treatment for Adolescents with Borderline Symptoms: A Pilot Study. <i>Psychotherapy, Children and Adolescents</i> , 51 (1):159–66. https://doi.org/10.1037/a0033513 |
| Ramires, V, R, R., Schwan, S., & Midgley, N. (2012). Mentalization-Based Therapy with Maltreated Children Living in Shelters in Southern Brazil: A Single Case Study. <i>Psychoanalytic Psychotherapy</i> , 26 (4): 308–26. https://doi.org/10.1080/02668734.2012.730546 |
| Rossouw, T, I., & Fonagy, P. (2012). Mentalization-Based Treatment for Self-Harm in Adolescents: A Randomized Controlled Trial. <i>Journal of the American Academy of Child & Adolescent Psychiatry</i> , 51 (12):1304–13. https://doi.org/10.1016/j.jaac.2012.09.018 |
| Twemlow, S, W., Fonagy, P., Sacco, F, C., Vernberg, E., & Malcom, J, M. (2011). Reducing Violence and Prejudice in a Jamaican All Age School Using Attachment and Mentalization Theory. <i>Psychoanalytic Psychology</i> , 28 (4): 497–511. https://doi.org/10.1037/a0023610 |

Weight of Evidence (WoE)

The seven studies included in the review were evaluated according to Gough's (2007) Weight of Evidence (WoE) framework. Each study was reviewed on methodological quality (WoE A), methodological relevance (WoE B) and topic relevance (WoE C).

The criteria for WoE A were selected and designed with randomised control trials (RCT), Cohort and small N studies in mind. A numerical rating from 0-3 was given in relation to the methodological quality of each study. Gersten et al. (2005) and Horner et al. (2005) coding protocols were used (see Appendix D).

The criteria for WoE B and WoE C were specified between 'low' and 'high' with tailored descriptors to suit the relevance of each study's research design, relevance of MBT intervention, and the relevance of each study weighted against the review question.

The overall effectiveness of each study was calculated by averaging the scores across WoE A, B and C measures to form WoE D. Appendix C contains further information about the WoE criteria and the descriptors used to evaluate the studies. Table 4 is a table showing the WoE A – D, across each study.

Table 4

Weight of Evidence Ratings

| Study Authors | Quality of Methodology (WoE A) | Relevance of Methodology (WoE B) | Relevance of evidence to the review question (WoE C) | Overall Weighting (WoE D) |
|---------------------------------|--------------------------------|----------------------------------|--|---------------------------|
| Fonagy et al., 2009 | High (2.8) | High (3) | High (3) | High (2.9) |
| Halfon, Bekar & Büşra, 2017 | High (2.4) | Low (1) | Low (1) | Medium (1.5) |
| Hauber, Boon & Vermeiren, 2017 | High (2.8) | Medium (2) | Medium (2) | Medium (2.3) |
| Laurensen et al., 2014 | High (2.8) | Medium (2) | Low (1) | Medium (1.9) |
| Ramires, Schwan & Midgley, 2012 | Low (1.1) | Low (1) | Low (1) | Low (1) |
| Rossouw & Fonagy, 2012 | High (2.8) | High (3) | Medium (2) | High (2.6) |
| Twemlow et al., 2011 | Low (1.4) | Medium (2) | Low (1) | Medium (1.5) |

Participants

1,151 participants took part in six of the included studies. It is unknown how many participants took part in the Twemlow et al. (2011) study because the numbers were not explicitly specified. The known age range of the studies were between 6 – 22. This shows that the included studies captured a wide range of ages of children and young adults experiencing psychological distress.

Few studies characterised the sample, therefore it is difficult to know how generalisable the interventions were to different demographics e.g. ethnicity. However, the studies were based on a wide range of countries suggesting that the results are generalisable to a wide range of cultural backgrounds. Females were significantly overrepresented when compared to males. Two studies included a sample comprising of 100% females (Halfon, Bekar & Büşra, 2017; Laurensen et al., 2014) and two studies included a sample comprising of over 80% female participants (Hauber, Boon & Vermeiren, 2017; Rossouw & Fonagy, 2012). This impacted a criterion of WoC, which was for the sample to include both females and males. Three studies did not meet this criteria (Halfon, Bekar & Büşra, 2017; Laurensen et al., 2014; Ramires, Schwan & Midgley, 2012). The overrepresentation of females in this literature review, could be due to the majority of studies that focused on young people with a diagnosis of BPD. Research has shown that females are more likely to get a diagnosis of BPD compared to males (Skodol & Bender, 2003). Inevitably, this affected the generalisability of these studies due to gender imbalances of the sample populations. This was evaluated in WoE A, and detailed in Appendix C.

Most studies had sufficient reporting of participant characteristics and received the maximum score for this. Two studies did not receive maximum scores for sufficient reporting of participants' characteristics (Ramires, Schwan & Midgley, 2012; Twemlow et al., 2011), which was one of the quality indicators for WoE A. In Ramires, Schwan and Midgley's (2012) study, it was not reported how participants were selected. Whilst in Twemlow et al.'s (2011) study, minimal participants' characteristics were reported.

Halfon, Bekar and Büşra (2017), screened for psychotic symptoms, significant developmental delays, significant risk of suicide attempts and drug abuse. Similarly, Laurensen et al. (2014) excluded on the basis of those who presented with psychopathy and severe learning difficulties. Therefore, it is unknown whether MBT is significantly effective for children and young adults who present with these difficulties. Hauber, Boon and Vermeiren (2017) worked with 115 adolescents and young adults who had a large range of comorbidities. For example, mood disorder; eating disorder; ADHD; substance dependence; dissociative disorder and obsessive compulsive disorder. This supports the generalisability of MBT being effective for a wide range of difficulties.

Research Design

RCT gained the highest WoE B overall rating. RCTs are the most effective design to support the review question. This is because the design uses randomisation to spread out participant characteristics, and therefore supports the validity of the results. On the contrary, small n designs have limited sample populations and therefore received the lowest WoE B overall rating. Indeed, Fonagy et al.'s (2009) study is the gold standard for research design in this literature review. This is because of different features in their design such as; a cluster-level RCT with stratified restricted allocation; and the study included an evaluation of efficacy after two years of intervention and an evaluation of effectiveness after one year. Support was withdrawn in order to reduce the intensity of a research protocol and to replicate real world conditions. Practitioners and teachers were provided with monitored support and supervision to ensure the intervention was delivered well. The efficacy in the effectiveness stages of the design supported internal and external validity.

Rossouw and Fonagy's (2012) research design was also an RCT. A relatively small sample size of participants were allocated via minimization, which controlled for various characteristics, such as past hospital admissions and demographics. The study was able to identify that MBT intervention led to less self-harm compared to treatment as usual (TAU) group. The two RCT designs used outcome measures to assess for long-term follow-up.

The cohort studies included in the literature review (Hauber, Boon & Vermeiren, 2017; Laurensen et al., 2014; Twemlow et al., 2011) tested for the effect of MBT over time. These studies received an overall medium WoE B rating. Without a comparison group this reduced the internal validity and therefore held less weight which is reflected in WoE B. However, Hauber, Boon and Vermeiren (2017), identified the difficulties of establishing RCT in clinical practice. This can pose ethical dilemmas for samples that poses suicidal risk.

Case studies (Halfon, Bekar & Büşra, 2017; Ramires, Schwan & Midgley, 2012) investigated the effects of MBT over a limited number of participants. These studies received an overall low WoE B rating. Within the two case studies in this literature review, they used either one or two participants. Whilst case studies can be advantageous in terms of allowing a more in depth examination, due to the limited numbers of participants, it reduces generalisability. Within the context of this literature review question, this design type was limited in its capacity to answer the question.

Setting Relevance – Educational vs. Clinical Significance

Fonagy et al.'s (2009) study was relevant to an educational setting because MBT was used to inform a systems approach within a school. This was a high quality study as defined by all WoEs, that used a large sample size and looked at efficacy

as well as effectiveness. This study was also highly relevant in comparison to other studies and is the only study to have received an overall high WoE C rating. This is because the intervention used was based on MBT, it was in an educational setting, and the sample included both females and males. Fonagy and colleagues (2009) randomly assigned 10 schools to intervention conditions, 'contrasted two school-wide interventions with a TAU control group in a cluster-randomized control trial, and included one-year post-intervention follow-up. They also stratified for low income students. Likewise, Twemlow et al. (2011) also provided MBT systemically to a school setting. However, the relevance of Twemlow et al.'s (2011) study was undermined because not all of the intervention implemented were based on MBT and it did not report the proportion of gender in its sample. As a result, it received an overall low WoE C rating.

All other studies used clinical populations where MBT was often delivered in a hospital setting. This reduces the generalisability of these studies in relation to supporting children and young people experiencing psychological distress in educational settings, which was reflected in WoE C ratings. However, in the Laurensen et al. (2014) study, it was reported that the practitioners strongly encouraged adolescents receiving treatment to attend school. This increased the relevance of this study, by encouraging the young people to generalise mentalization to an educational setting. The study that included a sample of both females and males within a clinical setting received an overall medium WoE C rating (Hauber, Boon & Vermeiren, 2017), whilst the studies that were in a clinical setting and the sample only included one sex received an overall low WoE C rating (Halfon, Bekar & Büşra, 2017; Laurensen et al., 2014; Ramires, Schwan & Midgley, 2012).

MBT Interventions

There were variations in terms of how MBT was implemented across the seven studies. Two studies utilised a wider community systemic approach. Fonagy et al. (2009) implemented a school-wide intervention called 'Creating a Peaceful School Learning Environment' (CAPSLE). The premise was to intervene in the current school culture and to use MBT to reduce aggression rates, bullying and victimisation amongst the school population (Fonagy et al., 2009). This was because it was assumed that each member of the school could contribute to preventing bullying behaviours (Fonagy et al., 2009). The theoretical framework behind this version of MBT was that by increasing mentalization for each school member, this would reduce bullying behaviours (Fonagy et al., 2009). Strategies to increase mentalization included, a positive climate campaign on bullying, consultation and training of school staff, self-defence training and peer mentoring (Fonagy et al., 2009).

Twemlow et al. (2011) also used a whole school approach and targeted a range of children which had trickle down effects to a larger section of the school population. The study used mentalization and power dynamic focused activities to encourage positive interrelationships between school, children and staff. Whilst this improved these relationships, it is difficult to know if the effects of change were due to power dynamic interventions or MBT. Furthermore, the community built a workshop to create a safe space and to positively impact school culture (Twemlow et al., 2011). The study did not specify how the adults involved in the intervention were trained. The study referred to minimal psychoanalytic supervision, however, this is not specified in detail. No manuals were provided or supplements, which effected the study's WoE A rating. It could be that the intervention did not need intensive training

or supervision and therefore the use of 'caregivers' to support the implementation of the interventions was all that was required. The study referred to some of the activities used in the intervention as 'easy to supervise' (Twemlow et al., 2011, p503), but does not incorporate any further details.

Other studies implemented MBT in a more traditional format, in either individual or group formats. Hauber, Boon and Vermeiren's (2017) study implemented group MBT where the participant received 'intensive' five days a week MBT with partial hospitalisation. The study did not mention the supervision of practitioners, therefore quality assurance regarding the implementation of MBT is uncertain. Rossouw and Fonagy's (2012) study used structured MBT as weekly individual sessions and monthly family therapy. Each session was 50 minutes. The practitioner who implemented MBT also received six days training. Laursenssen and colleagues (2014), included a range of interventions including weekly group psychotherapy, individual psychotherapy, art therapy, writing therapy, and mentalizing cognitive therapy. Ramires and colleagues (2012) and Halfon and colleagues (2017) only delivered individual psychotherapy, with Halfon and colleagues (2017) including parent sessions. Both of which were supervised by a more experienced clinician.

The majority of the studies ensured fidelity to MBT protocols, however two studies did not (Ramires, Schwan & Midgley, 2012; Twemlow et al., 2011). This made it difficult to ascertain whether MBT was implemented appropriately. As a result, both of these studies received the lowest score for the quality indicator relating to implementation of intervention as part of WoE A. The differences between the rest of the studies were due to the clarity of the implementation of the intervention. Two studies received the highest score for the quality indicator relating

to implementation of intervention as part of WoE A (Laurensen et al., 2014; Rossouw & Fonagy, 2012). The other studies received a medium score for the quality indicator relating to implementation of intervention as part of WoE A, due to having less clear information regarding how their intervention was implemented (Halfon, Bekar & Büşra, 2017; Hauber, Boon & Vermeiren, 2017; Rossouw & Fonagy, 2012). Interestingly, of the two studies that were based in community settings, the onus was on teachers or community members to help deliver the interventions, with support from experts. Whereas the studies that delivered MBT in clinical settings were delivered by trained clinicians. This may be due to differences in interventions (whole school approach vs. traditional MBT) and target populations (school community vs. clinical populations).

Outcome Measures

Five studies focused on personality disorders and/or psychological distress (Halfon, Bekar & Büşra, 2017; Hauber, Boon & Vermeiren, 2017; Laurensen et al., 2014; Ramires, Schwan & Midgley, 2012; Rossouw & Fonagy, 2012). These were measured via self-reports and interviews, which included a wide range of different questionnaires and interview schedules. Those that measured a form of personality disorder, demonstrated high content validity as they were based on the Diagnostic and Statistical Manual (DSM; American Psychiatric Association, 2013).

Two studies were based in educational settings and focused on other outcome measures relevant to these settings (Fonagy et al., 2009; Twemlow et al., 2011). These studies included measures such as victimization, aggression, disruptive classroom behaviours, confiscation of weapons and exam results.

Two studies stood out for the way they collected outcome measures. Halfon and colleagues (2017) used questionnaires to gather multiple perspectives in order

to triangulate their assessment of whether MBT was effective. These included clinician, teacher and parent perspectives. Though it is unclear why a child's perspective was missing, which undermined the validity of the results, as the perspective of whom the MBT intervention was delivered to was missing. Fonagy and colleagues (2009) were the only study to use behavioural assessment methods, such as classroom observations. The benefit of this is that it provided more objective data compared to questionnaires and interview.

The reporting of psychometric properties varied between the studies. Four studies did not cite the validity of the outcome measures they used (Fonagy et al., 2009; Ramires, Schwan & Midgley, 2012; Rossouw & Fonagy, 2012; Twemlow et al., 2011). This made it difficult to ascertain how valid the measures were. In addition, in Laurensen and colleague's (2014) study, the BSI and the Severity Indices of Personality Problems (SIPP-188) had an internal reliability of less than .7, which suggested that these measures had a relatively low level of reliability. In summary, the studies identified in this review suggested that MBT is effective for people with a diagnosis of personality disorder, depression, general psychological distress, victimization, aggression, self-harm, disruptive classroom behaviours, and reports of confiscating weapon.

Findings

The studies included in this review evaluated the effectiveness of MBT interventions on a wide range of outcome measures related to psychological distress. Across these studies MBT has been delivered individually, in groups, and using a whole school approach. For studies that used a RCT design, outcome measures were collected at multiple time points (Fonagy et al., 2009; Rossouw & Fonagy, 2012). As a result, these studies calculated effect sizes based on different

regression models, and compared the results with a treatment as usual control group (Roberts & Monaco, 2006). For studies that used a cohort design, there were no comparison groups (Hauber, Boon & Vermeiren, 2017; Laurensen et al., 2014; Twemlow et al., 2011). In both Hauber and colleague's (2017) and Laurensen and colleague's (2014) studies, effect sizes were calculated based on pre and post measures for within participants (Becker, 1988). Effect sizes were unable to be calculated for three studies (Halfon, Bekar & Būşra, 2017; Ramires, Schwan & Midgley, 2012; Twemlow et al., 2011). In Twemlow and colleague's (2011) study, the effect sizes and the descriptive statistics of outcomes related to psychological distress were not reported. Thus, effect sizes could not be calculated. In Ramires and colleague's (2012) single case study, only two data points were reported. As a result, no effect size could be calculated. Likewise, In Halfron and colleague's (2012) case study involving two participants, only two data points were reported, and so effect sizes could not be calculated. However, they did provide reliable change index scores (RCI; Jacobsen et al., 1999). RCI determines whether the magnitude of an intervention's effects are of clinical significance. This contrasts to effect sizes, as effect sizes do not determine whether an intervention effect is of clinical significance. In addition, effect sizes can be used to make comparison across different studies, which RCI cannot. Nonetheless, as RCI is a measure of the magnitude of an intervention's effect, this data was extracted.

The findings of the included studies have been separated as psychological distress relating to mental health difficulties, personality disorders, and aggression and victimization in schools (Tables 5 – 7 respectively). Regarding psychological distress relating to mental health difficulties, a broad range of outcomes were measured. This includes measures relating to general measures of distress,

depression, self-harm, and anxiety disorders. Large effect sizes were found in measures of general distress (Hauber, Boon & Vermeiren, 2017; Laurensen et al., 2014). Whilst small effect sizes were found for measures of depression and self-harm (Rossouw & Fonagy, 2012). The smaller effect sizes found in Roussouw and Fonagy's study (2012) may be due to different outcomes measures and use of RCT design instead of cohort design. Halfon and colleague's (2017) study documented RCI instead of effect sizes. This found that MBT was clinically effective for a range of measures that measured anxiety disorders and more general measures of psychological distress.

For findings relating to personality disorders, large effect sizes were found in relation to a range of personality disorders presentations (Hauber, Boon & Vermeiren, 2017, Laurensen et al., 2014). However, in Rossouw and Fonagy's (2012) study, only small effect sizes were found when the outcome measures measured BPD. The smaller effect size may be due to focusing on one type of personality disorder and it may be due to the use of RCT design compared to cohort designs. For findings relating to aggression and victimization in schools, small effect sizes were found for a range of measures relating to aggression and victimization.

To summarise, MBT was found to have varying effect sizes, ranging from small to large, on a range of outcome measures relating to mental health, personality disorders, and aggression and victimization in school. However, small effect sizes were found for study designs that used a RCT, with larger effect sizes found for study designs that used a cohort design.

Table 5: Effect sizes for psychological distress relating to mental health difficulties

| Authors | Measures | MBT | | | Control | | | Effect size | Effect size descriptor | p | Overall WoE D rating |
|-----------------------------|--|-----|-------------|-------------|---------|-------------|-------------|-------------|------------------------|------|----------------------|
| | | N | Pre M(SD) | Post M (SD) | N | Post M (SD) | Post M (SD) | | | | |
| Halfon, Bekar & Büşra, 2017 | Internalizing problems subscale of the CBCL | 2 | 65 (1.18) | 57 (1.82) | N/A | N/A | N/A | RCI =.28 | N/A | <.05 | Medium (1.5) |
| | Anxiety problems subscale of the CBCL | 2 | 65.5 (1.24) | 59 (1.73) | N/A | N/A | N/A | RCI =.5 | N/A | <.05 | |
| | Anxiety disorder subscale of the SCARED | 2 | 63 (1.73) | 41.5 (2.07) | N/A | N/A | N/A | RCI =.3.92 | N/A | <.05 | |
| | Separation anxiety disorder subscale of the SCARED | 2 | 8.5 (1.2) | 5 (1.91) | N/A | N/A | N/A | RCI =.2.29 | N/A | <.05 | |
| | Internalizing problems subscale of the TRF | 2 | 61 (1) | 50.5 (1.75) | N/A | N/A | N/A | RCI =.5.59 | N/A | <.05 | |
| | Anxiety problems subscale of the TRF | 2 | 63 (1) | 31.5 (2.14) | N/A | N/A | N/A | RCI =.5.53 | N/A | <.05 | |
| | Emotional lability subscale of the ERC | 2 | 3.28 (1.21) | 2.75 (0.96) | N/A | N/A | N/A | RCI =.1.34 | N/A | NS | |

| Authors | Measures | MBT | | | Control | | | Effect size | Effect size descriptor | p | Overall WoE D rating |
|---------------------------------|---|-----|--------------|--------------|---------|--------------|--------------|-------------|------------------------|--------|----------------------|
| | | N | Pre M(SD) | Post M (SD) | N | Post M (SD) | Post M (SD) | | | | |
| Halfon, Bekar & Büşra, 2017 | Emotional regulation subscale of the ERC | 3 | 2.065 (.66) | 2.75 (0.84) | N/A | N/A | N/A | RCI =.1.4 | N/A | NS | Medium (1.5) |
| Hauber, Boon & Vermeiren, 2017 | Self report of General Psychological Distress(SCL-90) | 62 | 241 (51.8) | 189.8 (64.8) | N/A | N/A | N/A | d=.87 | Large | <,.000 | Medium (2.3) |
| Laurensen et al., 2014 | Self report of Psychological Distress(BSI) | 11 | 2.2 (.19) | 1.33(.74) | N/A | N/A | N/A | d= 1.46 | Large | <.001 | Medium (1.9) |
| Ramires, Schwan & Midgley, 2012 | Self report of depression (CDI) | 1 | 40 (N/A) | 5 (N/A) | N/A | N/A | N/A | N/A | N/A | N/A | Low (1) |
| Rossouw & Fonagy, 2012 | Self-harm subscale of the RTSHI | 33 | 3.12 (.52) | 1.33 (1.26) | 35 | 3.08 (0.59) | 2.01 (1.24) | d=.28 | Small | <.05 | High (2.6) |
| | Self-report of depression(MFQ) | 33 | 17.46 (4.84) | 9.26 (7.3) | 35 | 16.32 (4.38) | 11.54 (6.74) | d=0.3 | Small | <.01 | |

Notes. BSI= Brief Symptom Inventory; CBCL= Child Behaviour Checklist; CDI= Child’s Depression inventory; ERC = Emotion Regulation Checklist; MFQ = Mood and Feelings Questionnaire; NS = Not Significant; RTSHI= Risk Taking and Self-Harm Inventory; SCARED= Screen for Child Anxiety Related Emotional Disorders; SCL-90= Symptom Check List 90; TRF= Teacher Rating Form.

Table 6: Effect sizes for psychological distress relating to personality disorder

| Authors | Measures | MBT | | | Control | | | Effect size | Effect size descriptor | p | Overall WoE D rating |
|--------------------------------|---|-----|-------------|-------------|---------|-------------|-------------|-------------|------------------------|-------|----------------------|
| | | N | Pre M(SD) | Post M (SD) | N | Post M (SD) | Post M (SD) | | | | |
| Hauber, Boon & Vermeiren, 2017 | Structured Clinical Interview for DSM personality disorders (SCID-II) | 62 | 1.42 (1.21) | 0.48 (0.78) | N/A | N/A | N/A | d= .92 | Large | <.001 | Medium (2.3) |
| Laurensen et al., 2014 | Self control subscale of Self-report of Personality Problems (SIPP-118) | 11 | 3.26(9.1) | 4.56(1.17) | N/A | N/A | N/A | d= 1.29 | Large | <.01 | Medium (1.9) |
| | Social concordance subscale of Self-report of Personality Problems (SIPP-118) | 11 | 5.18(.72) | 5.66(.68) | N/A | N/A | N/A | d= 1.29 | Large | <.01 | |

| Authors | Measures | MBT | | | Control | | | Effect Size | Effect size descriptor | p | Overall WoE D rating |
|------------------------|--|-----|------------|--------------|---------|-------------|--------------|-------------|------------------------|-------|----------------------|
| | | N | Pre M(SD) | Post M (SD) | N | Post M (SD) | Post M (SD) | | | | |
| Laurensen et al., 2014 | Identity Integration subscale of Self-report of Personality Problems (SIPP-118) | 11 | 2.72(.55) | 3.87(1.07) | N/A | N/A | N/A | d=.7 | Large | <.05 | Medium (1.9) |
| | Relational capacities subscale of Self-report of Personality Problems (SIPP-118) | 11 | 3.49(.61) | 4.06(1) | N/A | N/A | N/A | d=.72 | Large | <.05 | |
| | Responsibility subscale of Self-report of Personality Problems (SIPP-118) | 11 | 3.9(.67) | 4.33(.86) | N/A | N/A | N/A | d=.58 | Large | <.05 | |
| Rossouw & Fonagy, 2012 | Clinical Interview for BPD (CI-BPD) | 33 | 0.75(0.7*) | 0.33 (0.09*) | 35 | 0.7(0.7*) | 0.58 (0.09*) | d=0.34 | Small | <.041 | High (2.6) |

Notes. * = Proportional Odds Ratio; CI-BPD= Clinical Interview of Borderline Personality Disorder; SCID-II= Structured Clinical Interview for DSM Disorders; SIPP-118= Severity Indices of Personality Disorders.

Table 7: Effect sizes for psychological distress relation to aggression and victimization in schools

| Authors | Measures | MBT | | | Control | | | Effect size | Effect size descriptor | p | Overall WoE D rating |
|---------------------|--|-----|----------------|---------------|---------|---------------|---------------|-------------|------------------------|------|----------------------|
| | | N | Pre | Post | N | Pre | Post | | | | |
| Fonagy et al., 2009 | Peer nominations of aggression | 563 | 98.2 (10.2) | 101.7 (11.15) | 360 | 97.8 (9.11) | 102.7 (13.66) | d= .25 | Small | <.05 | High (2.9) |
| | Peer nominations of victimization | 563 | 98.7 (9.74) | 100.7 (9.25) | 360 | 97.6 (10.63) | 102.8 (14.04) | d=.3 | Small | <.01 | |
| | Peer nominations of aggressive by standing | 563 | 98.1 (10.41) | 101.2 (10.65) | 360 | 97.6 (9.49) | 102.7 (13.47) | d=.2 | Small | <.05 | |
| | Victimization (Peer experience questionnaire) | 563 | 100.64 (10.89) | 99.2 (10.41) | 360 | 99.70 (11.57) | 99.9 (11.57) | d= .06 | Small | NS | |
| | Aggression (Peer experience questionnaire) | 563 | 100.4 (11.63) | 100.2 (11.62) | 360 | 98.2 (10.63) | 99.7 (11.19) | d= .3 | Small | NS | |
| | Aggression as legitimate (peer experience questionnaire) | 563 | 99.1 (11.83) | 100.9 (11.63) | 360 | 96.5 (8.92) | 99.0 (9.49) | d= .09 | Small | NS | |

Notes. NS = Not Significant

Conclusions and Recommendations

The literature review highlights that there has been a limited amount of research investigating the effectiveness of MBT for children and young people, and only two studies were based in an educational setting (Fonagy et al., 2009; Twemlow et al., 2011). Of the studies identified in this review, there is a heterogeneity in terms of research designs, implementation of MBT, and outcome measures. This made it difficult to draw conclusions about the studies.

Nonetheless the review tentatively suggests that MBT has potential to improve psychological distress for children and young people in educational settings. The potential usefulness of MBT to improve psychological distress for children and young people in educational settings, is supported by variations in different implementations of MBT (individual, group, whole school approach) and wide range of outcome measures. In addition, large to small effect sizes were found for a range of outcomes relating to mental health, personality disorders, and aggression and victimization. These effect sizes tended to reduce to small effect sizes, with studies that used RCT (Fonagy et al., 2009; Rossouw & Fonagy, 2012). Furthermore, the age range of the participants of all studies in this review were between 6 to 22, and the effectiveness of MBT was documented across a range of different countries (Jamaica, UK, Brazil, Netherlands, Turkey, USA). This supports the generalisability of the results. A key study that stands out regarding the use of MBT for children and young people in an educational setting is the Fonagy et al. (2009) study. This was an RCT with a large sample (563) that found a small effect size for an MBT based whole school intervention to address aggression and victimization in the context of bullying.

Further research is needed to evaluate how effective MBT is to reduce psychological distress for children and young people in educational settings. This is due to a range of critiques of the current literature. This being the limited amount of research in this area, two studies were conducted within an educational setting (Fonagy et al., 2009; Twemlow et al., 2011), the majority of studies being conducted on a clinical sample, an overrepresentation of females in this current review, and limited amount of studies using a RCT design. Therefore, the current evidence base for whether MBT is effective for children and young people in educational settings is very limited. However, the current research warrants further research to explore how MBT can be effective for different types of outcomes relating to psychological distress for children and young people in educational settings.

The literature review has identified several recommendations. Practitioners may want to consider using Fonagy et al.'s (2009) MBT based whole school approach to address bullying. However, further research would benefit from comparing this MBT based intervention with other interventions addressing bullying, to compare which interventions are most effective for addressing bullying. Further research is recommended to build the evidence base of MBT for SEMH within educational settings. The current evidence suggests that MBT can be effective for a range of clinical samples. Future research may want to replicate these findings within educational settings, with the outcome measures being related to anxiety disorders, depression, self-harm, general psychological distress, and personality disorders. It will be helpful for future research to also ensure a significant proportion of males are included in their sample, given the overrepresentation of females identified in this review. Likewise, future research can benefit from collecting a range of

sociodemographic data, such as ethnicity, sexuality, and socioeconomic status. This will help to determine how generalisable findings are.

References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
- Asen E., & Fonagy, P. (2011). Mentalization-based therapeutic interventions for families. *Journal of Family Therapy*, 34 (4): 347–370.
- Bateman, A., & P. Fonagy. (2000). Effectiveness of psychotherapeutic treatment of personality disorder. *British Journal of Psychiatry*, 177 (2):138–143.
- Bateman, A. W. and Fonagy, P. (2004). *Psychotherapy of Borderline Personality Disorder: mentalisation based treatment*. Oxford University Press, Oxford.
- Bateman, A. W. and Fonagy, P. (2006). *Mentalization-Based Treatment for Borderline Personality Disorder: A Practical Guide*. Oxford University Press, Oxford.
- Bateman., A & Fonagy, P. (2013). Mentalization-Based Treatment. *Psychoanalytic Inquiry*, 33 (6) :595-613.
- Beck, E., Sune, B., Matthias, G., Stig, P., Liselotte, P., Jesper, P., & Erik, S. (2016). Mentalization-Based Treatment in Groups for Adolescents with Borderline Personality Disorder (BPD) or Subthreshold BPD versus Treatment as Usual (M-GAB): Study Protocol for a Randomized Controlled Trial. *Trials*, 17 (314): 1-13. <https://doi.org/10.1186/s13063-016-1431-0>.
- Becker, B. J. (1988). Synthesizing standardized mean change measures. *British Journal of Mathematical and Statistical Psychology*, 41, 257–278. <http://doi.org/10.1111/j.2044-8317.1988.tb00901.x>
- Bo, S., Sharp, C., Beck, E., Pedersen, J., Matthias, G., and Simonsen, E. (2017). First Empirical Evaluation of Outcomes for Mentalization-Based Group Therapy for Adolescents with BPD. *Personality Disorders: Theory, Research, and Treatment*, 8 (4): 396–401. <https://doi.org/10.1037/per0000210>.

- British Psychological Society (2015). Standards for the accreditation of educational psychology training in England, Northern Ireland & Wales. Available at: [https://www.bps.org.uk/sites/bps.org.uk/files/Accreditation/Educational%20Accreditation%20\(England,%20NI,%20Wales\)%202017_WEB.pdf](https://www.bps.org.uk/sites/bps.org.uk/files/Accreditation/Educational%20Accreditation%20(England,%20NI,%20Wales)%202017_WEB.pdf) (Accessed: 04 June 2018).
- Department for Education and Department of Health (2015) *Special educational needs and disability code of practice: 0 to 25 years*. Available at: <https://www.gov.uk/government/publications/send-code-of-practice-0-to-25> (Accessed: 04 June 2018).
- Fonagy, P., & Bateman, A. (2011). The widening scope of mentalizing: a discussion. *Psychology and Psychotherapy*, 84 (1): 98–110.
- Fonagy, P., Twemlow, W, S., Vernberg, M, E., Nelson, M, J., Dill, J, E., Little, D, T., & Sargent, A, J. (2009). A Cluster Randomized Controlled Trial of Child-Focused Psychiatric Consultation and a School Systems-Focused Intervention to Reduce Aggression. *Journal of Child Psychology and Psychiatry*, 50 (5): 607–16. <https://doi.org/10.1111/j.1469-7610.2008.02025.x>.
- Fonagy, P., Target, M., Cottrell, D., Phillips, J., Kurtz, Z., & Kurtz, A. (2005). *What Works for Whom? A Critical Review of Treatments for Children and Adolescents*, 2ndEdn. New York, NY: Guilford Press.
- Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C. & Innocenti, M. S. (2005). Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education. *Exceptional Children*, 71(2): 149-164.
- Gough, D. (2007). Weight of Evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, 22(2): 213–228.
doi:10.1080/02671520701296189

- Grohol, J. (2016). Mentalization Based Therapy (MBT). Psych Central. Retrieved on March 17, 2018, from <https://psychcentral.com/lib/mentalization-based-therapy-mbt/>
- Halfon, S., Ozlem B., & Büşra G. (2017). An Empirical Analysis of Mental State Talk and Affect Regulation in Two Single-Cases of Psychodynamic Child Therapy. *Psychotherapy, 54* (2): 207–19. <https://doi.org/10.1037/pst0000113>.
- Hauber, K., Boon, E, A., & Vermeiren, R. (2017). Examining Changes in Personality Disorder and Symptomology in an Adolescent Sample Receiving Intensive Mentalization Based Treatment: A Pilot Study. *Child and Adolescent Psychiatry and Mental Health, 11* (58): 1-7. <https://doi.org/10.1186/s13034-017-0197-9>.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. & Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children, 71*(2): 165-179.
- Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300–307. <http://dx.doi.org/10.1037/0022-006X.67.3.300>
- Laurensen, M, P, E., Hutsebaut, J., Feenstra, J, D., Bales, L, D., Noom, J, M., Busschbach, J, V, J., Verheul, R., & Luyten, P. (2014). Feasibility of Mentalization-Based Treatment for Adolescents with Borderline Symptoms: A Pilot Study. *Psychotherapy, Children and Adolescents, 51*(1): 159–66. <https://doi.org/10.1037/a0033513>.
- Midgley, N., Besser, J, S., Dye, H., Fearon, P., Gale, T., Jefferies-Sewell, K., & Irvine, K. (2017). The Herts and Minds Study: Evaluating the Effectiveness of Mentalization-Based Treatment (MBT) as an Intervention for Children in Foster Care with Emotional and/or Behavioural Problems: A Phase II, Feasibility, Randomised

Controlled Trial. *Pilot And Feasibility Studies*, 3 (12): 1–12.

<https://doi.org/10.1186/s40814-017-0127-x>.

Ougrin, D., Tranah, T., Stahl, D., Moran, P., & Asarnow, R, J. (2015). Therapeutic Interventions for Suicide Attempts and Self-Harm in Adolescents: Systematic Review and Meta-Analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54 (2): 97–107. <https://doi.org/10.1016/j.jaac.2014.10.009>.

Public Health England (2016). *The mental health of children and young people in England*. PHE Publications: London.

Ramires, R, R, V., Schwan, S., & Midgley. (2012). Mentalization-Based Therapy with Maltreated Children Living in Shelters in Southern Brazil: A Single Case Study. *Psychoanalytic Psychotherapy*, 26 (4): 308–26. <https://doi.org/10.1080/02668734.2012.730546>.

Roberts, J. K., & Monaco, J. P. (2006). *Effect size measures for the two-level linear multilevel model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rossouw, I, T., & Fonagy, F. (2012). Mentalization-Based Treatment for Self-Harm in Adolescents: A Randomized Controlled Trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(12): 1304–13. <https://doi.org/10.1016/j.jaac.2012.09.018>.

Skodol, A, E., & Bender, D, S. (2003). Why are women diagnosed borderline more than men? *Psychiatric Quarterly*, 74(4):349-60.

Twemlow, Stuart W., Fonagy, P., Sacco, C, F., Vernberg, E., & Malcom, M, J. (2011). Reducing Violence and Prejudice in a Jamaican All Age School Using Attachment and Mentalization Theory. *Psychoanalytic Psychology*, 28 (4): 497–511. <https://doi.org/10.1037/a0023610>.

Appendices

Appendix A - Studies Excluded from Systematic Review with Reasons

Studies Excluded at Abstract Screening:

| Study Reference | Reason |
|---|---|
| Aguert, M., Le Vallois, C., Martel, K., & Laval, V. (2018). "That's really clever!" Ironic hyperbole understanding in children. <i>Journal of Child Language</i> , 45(1), 260–272. https://doi.org/10.1017/S0305000917000113 | Not MBT, Not SEND |
| Aung, A. T., Pickworth, W. B., & Moolchan, E. T. (2004). History of marijuana use and tobacco smoking topography in tobacco-dependent adolescents. <i>Addictive Behaviors</i> , 29(4), 699–706. https://doi.org/10.1016/j.addbeh.2004.02.012 | Not MBT, Not SEND |
| Bilewicz, M., Mikołajczak, G., & Babińska, M. (2017). Speaking about the preborn. How specific terms used in the abortion debate reflect attitudes and (de)mentalization. <i>Personality and Individual Differences</i> , 111, 256–262. https://doi.org/10.1016/j.paid.2017.02.018 | Not MBT, not SEND, no pre and post measures |
| Bychkova, T., Hillman, S., Midgley, N., & Schneider, C. (2011). The psychotherapy process with adolescents: A first pilot study and preliminary comparisons between different therapeutic modalities using the Adolescent Psychotherapy Q-Set. <i>Journal of Child Psychotherapy</i> , 37(3), 327–348. https://doi.org/10.1080/0075417X.2011.614754 | No pre and post measures |

| Study Reference | Reason |
|--|---|
| <p>Cho, Y. H., & Huang, Y. (2014). Exploring the links between pre-service teachers' beliefs and video-based reflection in wikis. <i>Computers in Human Behavior</i>, 35, 39–53. https://doi.org/10.1016/j.chb.2014.02.022</p> | <p>Not MBT, not on children and young people, not on SEND</p> |
| <p>Conway, F., Oster, M., & Szymanski, K. (2011). ADHD and complex trauma: A descriptive study of hospitalized children in an urban psychiatric hospital. <i>Journal of Infant, Child & Adolescent Psychotherapy</i>, 10(1), 60–72. https://doi.org/10.1080/15289168.2011.575707</p> | <p>No MBT or pre and post measures</p> |
| <p>Di Lorenzo, M., Maggiolini, A., & Suigo, V. A. (2015). A developmental perspective on adolescent psychoanalytic psychotherapy. An Italian study with the Adolescent Psychotherapy Q-Set. <i>Research in Psychotherapy: Psychopathology, Process and Outcome</i>, 18(2), 102–113.</p> | <p>Not focused on children or young people, no pre and post measures,</p> |
| <p>D'Onofrio, E., Pace, C. S., & Cavanna, D. (2015). Qualitative research in adolescent psychotherapy: Attachment and Reflective Functioning as psychotherapy's outcomes of an adolescent with anorexia nervosa. <i>Research in Psychotherapy: Psychopathology, Process and Outcome</i>, 18(2), 93–101.</p> | <p>No pre and post measures</p> |

| Study Reference | Reason |
|---|---|
| <p>Farkas, C., Vallotton, C. D., Strasser, K., Santelices, M. P., & Himmel, E. (2017). Socioemotional skills between 12 and 30 months of age on Chilean children: When do the competences of adults matter? <i>Infant Behavior & Development</i>, 49, 192–203. https://doi.org/10.1016/j.infbeh.2017.09.010</p> | No MBT, no SEND |
| <p>Fonagy, P. (1999). The transgenerational transmission of holocaust trauma: Lessons learned from the analysis of an adolescent with obsessive-compulsive disorder. <i>Attachment & Human Development</i>, 1(1), 92–114. https://doi.org/10.1080/14616739900134041</p> | No MBT, no pre and post measures |
| <p>Fossati, A., Feeney, J., Maffei, C., & Borroni, S. (2011). Does mindfulness mediate the association between attachment dimensions and Borderline Personality Disorder features? A study of Italian non-clinical adolescents. <i>Attachment & Human Development</i>, 13(6), 563–578. https://doi.org/10.1080/14616734.2011.608993</p> | No MBT, no pre and post measures |
| <p>Fossati, A., Feeney, J., Maffei, C., & Borroni, S. (2014). Thinking about feelings: Affective state mentalization, attachment styles, and borderline personality disorder features among Italian nonclinical adolescents. <i>Psychoanalytic Psychology</i>, 31(1), 41–67. https://doi.org/10.1037/a0033960</p> | No MBT intervention, no pre and post measures |

| Study Reference | Reason |
|---|--|
| <p>Gavazzi, I. G., Ornaghi, V., & Antoniotti, C. (2011). Children's and adolescents' narratives of guilt: Antecedents and mentalization. <i>European Journal of Developmental Psychology</i>, 8(3), 311–330.</p> <p>https://doi.org/10.1080/17405629.2010.491303</p> | <p>NO MBT, no SEND, no pre and post measures</p> |
| <p>Goodman, G., Midgley, N., & Schneider, C. (2016). Expert clinicians' prototypes of an ideal child treatment in psychodynamic and cognitive-behavioral therapy: Is mentalization seen as a common process factor? <i>Psychotherapy Research</i>, 26(5), 590–601.</p> <p>https://doi.org/10.1080/10503307.2015.1049672</p> | <p>No children or young adult, No pre and post measures,</p> |
| <p>Goodman, G., Reed, P., & Athey-Lloyd, L. (2015). Mentalization and play therapy processes between two therapists and a child with Asperger's disorder. <i>International Journal of Play Therapy</i>, 24(1), 13–29.</p> <p>https://doi.org/10.1037/a0038660</p> | <p>No pre and post measure</p> |
| <p>Ha, C., Sharp, C., Ensink, K., Fonagy, P., & Cirino, P. (2013). The measurement of reflective function in adolescents with and without borderline traits. <i>Journal of Adolescence</i>, 36(6), 1215–1223.</p> <p>https://doi.org/10.1016/j.adolescence.2013.09.008</p> | <p>No MBT intervention, no pre and post measure</p> |

| Study Reference | Reason |
|---|---|
| <p>Heede, T., Runge, H., Storebø, O. J., Rowley, E., & Hansen, K. G. (2009). Psychodynamic milieu-therapy and changes in personality—What is the connection? <i>Journal of Child Psychotherapy</i>, 35(3), 276–289.</p> <p>https://doi.org/10.1080/00754170903237286</p> | No MBT |
| <p>Ingleby-Cook, G., & Dobel-Ober, D. (2013). Innovations in practice: Group work with children who are in care or who are adopted: Lessons learnt. <i>Child and Adolescent Mental Health</i>, 18(4), 251–254.</p> | No pre and post measures |
| <p>Lorentzen, E., Nilsen, H., & Traeen, B. (2008). Will it never end? The narratives of incest victims on the termination of sexual abuse. <i>Journal of Sex Research</i>, 45(2), 1–1.</p> <p>https://doi.org/10.1080/00224490801987473</p> | No children and young people, No MBT, no pre and post measures |
| <p>McCullough, E., Stedmon, J., & Dallos, R. (2014). Narrative responses as an aid to understanding the presentation of maltreated children who meet criteria for autistic spectrum disorder and reactive attachment disorder: A case series study. <i>Clinical Child Psychology and Psychiatry</i>, 19(3), 392–411.</p> <p>https://doi.org/10.1177/1359104513503353</p> | No MBT, no pre and post measures |

| Study Reference | Reason |
|---|---|
| <p>Muñoz Specht, P., Ensink, K., Normandin, L., & Midgley, N. (2016). Mentalizing techniques used by psychodynamic therapists working with children and early adolescents. <i>Bulletin of the Menninger Clinic</i>, 80(4), 281–315. https://doi.org/10.1521/bumc.2016.80.4.281</p> | No pre and post measures |
| <p>O'Malley, F. (2003). Mentalizing in the psychotherapy of an disturbed adolescent girl. <i>Bulletin of the Menninger Clinic</i>, 67(2), 150–157. https://doi.org/10.1521/bumc.67.2.150.23442</p> | No pre post measures |
| <p>Pelosi, A., Zorzi, G., & Corsano, P. (2014). The “Body Image Control in Photos questionnaire” (BICP): A new tool for the analysis of self-presentation of body image on Facebook during adolescence. <i>Applied Psychology Bulletin</i>, 269(62), 42–52.</p> | NO MBT no pre post measures |
| <p>Psycharis, S. (2013). Examining the effect of the computational models on learning performance, scientific reasoning, epistemic beliefs and argumentation: An implication for the STEM agenda. <i>Computers & Education</i>, 68, 253–265. https://doi.org/10.1016/j.compedu.2013.05.015</p> | No children and young people, no MBT, no SEND |

| Study Reference | Reason |
|--|---|
| <p>Quek, J., Newman, L. K., Bennett, C., Gordon, M. S., Saeedi, N., & Melvin, G. A. (2017). Reflective function mediates the relationship between emotional maltreatment and borderline pathology in adolescents: A preliminary investigation. <i>Child Abuse & Neglect</i>, 72, 215–226. https://doi.org/10.1016/j.chiabu.2017.08.008</p> | <p>No MBT, no pre and post measures</p> |
| <p>Rouzer, S. K., Cole, J. M., Johnson, J. M., Varlinskaya, E. I., & Diaz, M. R. (2017). Moderate maternal alcohol exposure on gestational day 12 impacts anxiety-like behavior in offspring. <i>Frontiers in Behavioral Neuroscience</i>, 11. https://doi.org/10.3389/fnbeh.2017.00183</p> | <p>No children or young people, no MBT, no SEND</p> |
| <p>Rueda, P., Cabello, R., & Fernández-Berrocal, P. (2013). Preliminary validation of Spanish “Eyes Test-Child Version.” <i>Ansiedad y Estrés</i>, 19(2–3), 173–184.</p> | <p>No MBT, no pre and post measures, no SEND</p> |
| <p>Rutherford, H. J. V., Wareham, J. D., Vrouva, I., Mayes, L. C., Fonagy, P., & Potenza, M. N. (2012). Sex differences moderate the relationship between adolescent language and mentalization. <i>Personality Disorders: Theory, Research, and Treatment</i>, 3(4), 393–405. https://doi.org/10.1037/a0028938</p> | <p>No MBT, no pre and post measure</p> |

| Study Reference | Reason |
|---|-------------------------------------|
| <p>Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). Theory of mind and emotion regulation difficulties in adolescents with borderline traits. <i>Journal of the American Academy of Child & Adolescent Psychiatry</i>, 50(6), 563–573. https://doi.org/10.1016/j.jaac.2011.01.017</p> | No MBT and no pre and post measures |
| <p>Sharp, C., Williams, L. L., Ha, C., Baumgardner, J., Michonski, J., Seals, R., & Fonagy, P. (2009). The development of a mentalization-based outcomes and research protocol for an adolescent inpatient unit. <i>Bulletin of the Menninger Clinic</i>, 73(4), 311–338. https://doi.org/10.1521/bumc.2009.73.4.311</p> | No pre and post measures |
| <p>Smetana, M. (2017). Recurring similarity: The meaning of musical objects in music therapy for adolescents with structural disorders. <i>Nordic Journal of Music Therapy</i>, 26(2), 105–123. https://doi.org/10.1080/08098131.2015.1117123</p> | No MBT no pre and post measure |
| <p>Strawn, J. R., Bitter, S. M., Weber, W. A., Chu, W., Whitsel, R. M., Adler, C., & DelBello, M. P. (2012). Neurocircuitry of generalized anxiety disorder in adolescents: A pilot functional neuroimaging and functional connectivity study. <i>Depression and Anxiety</i>, 29(11), 939–947. https://doi.org/10.1002/da.21961</p> | No MBT, no pre and post measure |

| Study Reference | Reason |
|---|--|
| <p>Strawn, J. R., John Wegman, C., Dominick, K. C., Swartz, M. S., Wehry, A. M., Patino, L. R., & DelBello, M. P. (2014). Cortical surface anatomy in pediatric patients with generalized anxiety disorder. <i>Journal of Anxiety Disorders</i>, 28(7), 717–723. https://doi.org/10.1016/j.janxdis.2014.07.012</p> | <p>No MBT, no pre and post measure</p> |
| <p>Strawn, J. R., Wehry, A. M., Chu, W., Adler, C. M., Eliassen, J. C., Cerullo, M. A., & DelBello, M. P. (2013). Neuroanatomic abnormalities in adolescents with generalized anxiety disorder: A voxel-based morphometry study. <i>Depression and Anxiety</i>, 30(9), 842–848. https://doi.org/10.1002/da.22089</p> | <p>No MBT, no pre and post measure</p> |
| <p>Taubner, S., Müller, S., & Kotte, S. (2017). Mentalizing Vocational Training increases mentalization interest in professionals and young people with the need for learning support: Results of a pilot study. <i>Mental Health and Prevention</i>, 6, 1–11. https://doi.org/10.1016/j.mhp.2017.02.001</p> | <p>No MBT</p> |
| <p>Taubner, S., White, L. O., Zimmermann, J., Fonagy, P., & Nolte, T. (2012). Mentalization moderates and mediates the link between psychopathy and aggressive behavior in male adolescents. <i>Journal of the American Psychoanalytic Association</i>, 60(3), 605–612.</p> | <p>No MBT and no pre post measures</p> |

| Study Reference | Reason |
|--|--|
| <p>Taubner, S., White, L. O., Zimmermann, J., Fonagy, P., & Nolte, T. (2013). Attachment-related mentalization moderates the relationship between psychopathic traits and proactive aggression in adolescence. <i>Journal of Abnormal Child Psychology</i>, 41(6), 929–938. https://doi.org/10.1007/s10802-013-9736-x</p> | <p>No MBT and no pre post measures</p> |
| <p>Taubner, S., Zimmermann, L., Ramberg, A., & Schröder, P. (2016). Mentalization mediates the relationship between early maltreatment and potential for violence in adolescence. <i>Psychopathology</i>, 49(4), 236–246. https://doi.org/10.1159/000448053</p> | <p>No MBT and no pre post measures</p> |
| <p>Tessier, V. P., Normandin, L., Ensink, K., & Fonagy, P. (2016). Fact or fiction? A longitudinal study of play and the development of reflective functioning. <i>Bulletin of the Menninger Clinic</i>, 80(1), 60–79. https://doi.org/10.1521/bumc.2016.80.1.60</p> | <p>No MBT</p> |
| <p>Twemlow, S. W., Fonagy, P., & Sacco, F. C. (2005). A developmental approach to mentalizing communities: II. The Peaceful Schools experiment. <i>Bulletin of the Menninger Clinic</i>, 69(4), 282–304. https://doi.org/10.1521/bumc.2005.69.4.282</p> | <p>Secondary data</p> |

| Study Reference | Reason |
|---|---|
| Yokota, S., Taki, Y., Hashizume, H., Sassa, Y., Thyreau, B., Tanaka, M., & Kawashima, R. (2013). Neural correlates of deception in social contexts in normally developing children. <i>Frontiers in Human Neuroscience, 7</i> | No MBT, no pre and post measures, no SEND |

Studies excluded at full text screening:

| Study Reference | Reason |
|--|--|
| Bak, P. L., Midgley, N., Zhu, J. L., Wistoft, K., & Obel, C. (2015). The Resilience Program: Preliminary evaluation of a mentalization-based education program. <i>Frontiers in Psychology, 6</i> . | 5. Population sample over 18 years old |
| Sharp, C., Ha, C., Carbone, C., Kim, S., Perry, K., Williams, L., & Fonagy, P. (2013). Hypermentalizing in adolescent inpatients: Treatment effects and association with borderline traits. <i>Journal of Personality Disorders, 27</i> (1), 3–18. | 4. Intervention not based on MBT |

Appendix B - Summary of Included Studies and Effect Sizes

| Study Authors | Participants | Design | Intervention/s | Outcome Measures & Measures of Psychological Distress | Key Findings |
|---------------------|--|--|---|--|---|
| Fonagy et al., 2009 | N=1035 R= 8-11 46% females Education setting in USA | Cluster-level randomized controlled trial with stratified restricted allocation. | School psychiatric consultation (SPC), a systems and mentalization focused whole school intervention (CAPSLE), and Treatment as usual (TAU) | Peer nominations of aggression, victimization, and bystanding; and the Peer experience questionnaire (self-reports of aggression, victimization, and mentalizing). | CAPSLE showed significant improvement relative to TAU across time for: peer-reported aggression ($p < .05$, $ES=.01$, small effect size), peer-reported victimization ($p < .01$, $ES=.12$), aggressive bystanding ($p < .05$, $ES=.05$) and empathic mentalizing ($p < .01$, $ES=1.7$). CAPSLE produced a significant decrease in off-task ($p < .001$) and disruptive classroom behaviors ($p < .01$). Superiority with respect to TAU for victimization ($p < .05$, $ES=0.8$), aggression ($p < .01$, $ES=.2$), and helpful ($p < .05$, $ES=.17$) and aggressive bystanding ($p < .01$, $ES=.21$) were maintained in the follow-up year. |

| Study Authors | Participants | Design | Intervention/s | Outcome Measures & Measures of Psychological Distress | Key Findings |
|-----------------------------------|--|------------|---|---|---|
| Halfon, Bekar & Büşra, 2017 | N=2 M age=6 Females- 100% Clinical sample based in Turkey | Case study | Long-term psychodynamic play therapy informed with mentalization principles. | CBCL (Parent report of problematic behaviours); SCARED (Parent report of anxiety disorders); ERC (Parent report of emotion regulation); TRF (Teacher report of problematic behaviours); CGAS (Clinician report of general functioning) | One participant had significant and reliable change across all measures as determined by RCI, the other participant only had significant and reliable change for the CGAS and adaptive functioning subscale of the TRF |

| Study Authors | Participants | Design | Intervention/s | Outcome Measures & Measures of Psychological Distress | Key Findings |
|--|--|--------------|----------------|--|--|
| Hauber, Boon & Vermeiren, 2017 Clinical sample based in Netherlands | N=62 M=18.2 (SD)=1.6 R=15-22 Females =80.9% Males= 19.1% | Longitudinal | Group MBT | SCID-II (Structured Clinical Interview for DSM personality disorders) and SCL-90 (Self report of General Psychological Distress) | Significant reductions in SCID-II (t = 8.36, p = .000, d=.92, large effect size) and SCL-90 (t = 5.95, p = .000, d=.97, large effect size) |

| Study Authors | Participants | Design | Intervention/s | Outcome Measures & Measures of Psychological Distress | Key Findings |
|---------------------------------|---|-------------------|---|---|---|
| Laurensen et al., 2014 | N=11 R= 14-18 Females 100% Clinical Sample based in the Netherlands | Longitudinal | Group psychotherapy, individual psychotherapy session, Art Therapy, writing therapy and mentalizing cognitive therapy | BSI (Self report of Psychological Distress), SIPP-118 (Self report of Personality Problems), (SIPP-118), EQ-5D (Self report of quality of life) | Significant improvements on BSI ($p < .001$, $d = 1.46$, small effect size); SIPP-118 domains, self-control ($p < .01$, $d = 1.29$, small effect size), social concordance ($p < .05$, $d = .70$), identity integration ($p < .01$, $d = 1.42$, large effect size) and responsibility ($p < .05$, $d = .58$, medium effect size); and EQ-5DF, ($p < .05$, $d = 1.11$, large effect size), |
| Ramires, Schwan & Midgley, 2012 | N=1 Age – 7 Male Community sample in Brazil | Single case study | MBT | CDI (Self report of depression) | Significant improvement in CDI (no reported stats) |

| Study Authors | Participants | Design | Intervention/s | Outcome Measures & Measures of Psychological Distress | Key Findings |
|------------------------|---|--------|---------------------------------|--|--|
| Rossouw & Fonagy, 2012 | N=40 33 Females (82.5%) M = 15.4 (SD=1.3) R=12-17 Clinical sample based in UK | RCT | MBT and family therapy, and TAU | RTSHI (Self report of Self-harm and risk taking, MFQ (self report of depression), CI-BPD (Clinical interview of borderline personality features), CI-BPD (Interview for BPD) | MBT-A led to significantly less self-harm compared with TAU ($r=-.07$, large effect size, $p<.01$); significantly less depression compared with TAU ($r=-3.31$, large effect size, $p<.05$); significantly less likely to have participants meeting BPD criteria compared to TAU after treatment ($b=.072$, 95% CI = 0 to .91, $t(140)=-2.03$, $p<0.042$, $d=0.34$, small effect size). |

| Study Authors | Participants | Design | Intervention/s | Outcome Measures & Measures of Psychological Distress | Key Findings |
|----------------------|--|--------------|---|--|---|
| Twemlow et al., 2011 | R= 7-9 Education setting based in Jamacia | Longitudinal | Whole school approach informed by MBT and Power dynamics focused activities | Peer Experiences Questionnaire, GSAT exam results, Teacher's reports of confiscating weapons or being intimidated by students. | <p>Boys report decreased prevalence of victimization from 70% to 30% ($S = 737, p < .007$); Girls, relational victimization reduced from 70% to 45% ($S = 564, p < .007$); Proportion who volunteered that they took a knife or gun, or other weapon to school, decreased from 40% to 20% ($S = 409, p < .03$); Significant reduction of student reporting other children taking weapons to school ($S = 46, p < .03$).</p> <p>Teacher reports of confiscating a weapon from a student or being intimidated by students, decreased significantly as indicated in the teacher questionnaire ($F = 6.9, df = 1, 21, p < .03$ and $F = 4.1, df = 1, 21, p < .06$ for confiscating weapon and intimidation, respectively).</p> <p>Aggressive bystanding decreased for both genders significantly, ($F = 4.59, df = 3, 21, p < .02$). Overt victimization was reduced across the 3 years in, from 65% to 37%</p> |

($S = 25,818$, $p < .00001$). Relational victimization also reduced substantially particularly for boys in ($S = 1,281$, $p < .002$). significant decrease in relational victimization from 32 to 27%. Both these differences are statistically significant ($S = 7,178$, $p < .05$).

Beliefs that “aggression pays” dramatically declined from an endorsement rate of 75% to the endorsement rate of less than 50% ($S = 23,576$, $p < .0001$).

Note. Fonagy et al., 2009: CAPSLE – Creating a Peaceful School Learning Environment. Halfon, Bekar & Büşra, 2017: CBCL – Child Behaviour Checklist; CGAS – Children’s Global Assessment Scale; CI-BPD - Clinical interview of borderline personality disorder; ERC – Emotion Regulation Checklist; SCARED – Screen for Child Anxiety Related Emotional Disorders; TRF – Teacher Rating Form. Hauber, Boon & Vermeiren, 2017: DSM – Diagnostic and Statistical Manual of Mental Disorders; SCID-II – Structured Clinical Interview for DSM Disorders; SCL-90 – Symptom Check List 90. Rossouw & Fonagy, 2012: RTSHI – RiskTaking and Self-Harm Inventory; MFQ – Mood and Feelings Questionnaire; HIF – How I Feel questionnaire (unpublished data, 2008); ECR – Experience of Close Relationships Inventory; BPD – Borderline Personality Disorder. Laurensen et al, 2014: Brief symptom inventory (BSI), Severity Indices of Personality Disorders (SIPP-118), EuroQol (EQ-5d). Ramiers et al 2012: CDI – Child’s Depression inventory; R= Range; ES= effect size calculated by dividing beta estimate for main effect by the square root of the variance estimate at every time-point and averaging these to obtain overall effect size for the group; RCI= Reliable Change Index

*Appendix C - Weight of Evidence Information***Weight of Evidence A (WoE A) – Methodological Quality**

For the WoE A each of the selected studies were rated on the quality of the research methodology. Five of the included studies were RCTs or Cohort designs, therefore Gersten et al.'s (2005) coding protocol was used for WoE A. Halfon, Bekar & Büşra (2017) and Ramires, Schwan & Midgley (2012) were small N studies. The Horner et al.'s (2005) protocol was used for these small N studies.

Below is the justification for all WoE A with reference to specific features. Further to this, the weighting compares each study with other studies of similar type. Though the coding protocols were not altered, additional criteria were provided in WoE A for RCT related to comparison conditions, whereas the cohort design did not have the relevant data. The scores across the key elements were averaged to calculate an overall value for the methodological quality.

The original indicators were used from Gersten et al. (2005) and Horner et al. (2005). A numerical rating from 0-3 was given in relation to the methodology quality of each study, instead of a yes/no response as specified in Gersten et al. (2005) and in Horner et al. (2005). This is because some studies partially met some of Gersten et al.'s (2005) and Horner et al.'s (2005) criteria, however this was unaccounted for in a yes/no response. Having a numerical rating from 0-3 allowed for more specific and accurate ratings of each study's criteria. Gersten et al. (2005) suggests that for a rating of high quality a study would need to meet all but one of the Essential Quality Indicators, and demonstrate at least four of the quality indicators listed as Desirable. To be considered acceptable quality, Gersten et al. (2005) suggests that the study should meet all but one Essential Quality Indicator and demonstrate at least one

Desirable Indicator. The latter has been factored into the sixth indicator, which has been included to rate the desirable aspects of each study.

An average rating for each study was calculated as follows: Low 1.4 or less, Medium 1.5-2.4, High 2.5 or above.

WoE A Criteria for Group Designs (summarised from Gersten et al. 2005)

| Criteria | Quality Ratings |
|--------------------------------|---|
| I: Description of Participants | <p>3. Sufficient information was provided to confirm participants demonstrated difficulties presented. Sufficient information was given characterizing the interventionists and/or teachers.</p> <p>RCTs - Sufficient information was given characterizing the interventionists and/or teachers and indicated that they were comparable across conditions. Appropriate procedures were used to increase the probability that participants were comparable across conditions.</p> <p>2. Information was provided to confirm participants demonstrated difficulties presented. Information was given characterizing the interventionists and/or teachers.</p> <p>RCTs - Information was given characterizing the interventionists and/or teachers. Procedures were used to increase the probability that participants were comparable across conditions.</p> <p>1. Information was provided to confirm participants.</p> <p>RCTs - Interventionists and/or teachers mentioned. Conditions mentioned.</p> <p>0. No information was provided to confirm participants.</p> <p>RCTs – No mention of conditions, interventionists and/or teachers.</p> |

| Criteria | Quality Ratings |
|--|---|
| II: Implementation | <p>3. The intervention was clearly described and specified. The fidelity of implementation was described and assessed.</p> <p>2. The intervention was not clearly described and specified. The fidelity of implementation was described and assessed.</p> <p>1. The intervention was not clearly described and specified. The fidelity of implementation was not described or assessed.</p> <p>0. The intervention was not described and specified. The fidelity of implementation was not described or assessed.</p> |
| III: Description of Comparison Conditions (for RCTs only) | <p>3. The nature of services provided in comparison conditions was described.</p> <p>2. Comparison conditions were described/detailed.</p> <p>1. Comparison conditions were mentioned.</p> <p>0. No mention of comparison conditions.</p> |
| IV: Outcome Measures | <p>3. Multiple measures were used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance. Outcomes are measured at the appropriate times to capture the intervention's effect.</p> <p>2. Some measures were used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance. Outcomes are measured at pre and post times to capture the intervention's effect.</p> <p>1. No measures were used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance. Outcomes are measured at either pre or post times to capture the intervention's effect.</p> <p>0. No measures were used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance. Outcomes were not measured at either pre or post times to capture the intervention's effect.</p> |

| Criteria | Quality Ratings |
|----------------------------------|--|
| V: Data Analysis | <p>3. Clearly linked data analysis techniques to key research questions and hypotheses. They were also appropriately linked to the unit of analysis in the study. The study included inferential statistics and effect size calculations.</p> <p>2. Clearly linked data analysis techniques to key research questions and hypotheses. They were also appropriately linked to the unit of analysis in the study. The study did not include inferential statistics and effect size calculations.</p> <p>1. Unclear links to data analysis techniques to key research questions and hypotheses. Unclear or no links to the unit of analysis in the study. The study did not include inferential statistics and effect size calculations.</p> <p>0. No links to data analysis techniques to key research questions and hypotheses. No links to the unit of analysis in the study. The study did not include inferential statistics and effect size calculations.</p> |
| VI: Desirable Quality Indicators | <p>3. Includes four elements of desirable qualities evidenced and all but one essential evidenced.</p> <p>2. Considered acceptable with all but one essential evidenced and at least one desirable quality evidenced.</p> <p>1. Not all essential evidenced and at least one desirable quality evidenced.</p> <p>0. Not all essential evidenced and no desirable qualities evidenced.</p> |

Summary of WoE A for Group Designs

| Study Authors | Dimensions | | | | | | WoE A Rating |
|--------------------------------|------------|----|-----|----|---|----|--------------|
| | I | II | III | IV | V | VI | |
| Fonagy et al., 2009 | 3 | 2 | 3 | 3 | 3 | 3 | 2.8 |
| Hauber, Boon & Vermeiren, 2017 | 3 | 2 | -- | 3 | 3 | 3 | 2.8 |
| Laurensen et al., 2014 | 3 | 3 | -- | 3 | 3 | 2 | 2.8 |
| Rossouw & Fonagy, 2012 | 3 | 3 | 3 | 3 | 3 | 2 | 2.8 |
| Twemlow et al., 2011 | 0 | 1 | -- | 3 | 2 | 1 | 1.4 |

Note. The dashes in dimension III denote the lack of comparison groups for the cohort studies.

WoE A Criteria for Small N Studies (Horner et al., 2005)

| Criteria | Quality Ratings |
|--|--|
| I: Description of Participants and Setting | <p>3. Participants were described with sufficient detail that allows for others to select individuals with similar characteristics. The process of participant selection was described with operational precision. Critical features of the physical setting were described with sufficient precision which allows for replication.</p> <p>2. Two out of rating 3 criteria are fulfilled.</p> <p>1. One of rating 3 criterion is fulfilled.</p> <p>0. None of the criteria in rating 3 is fulfilled fully.</p> |

| Criteria | Quality Ratings |
|---------------------------|--|
| II: Dependent Variable | <p>3. (All criteria is fulfilled) Dependent variables were described with operational precision and were measured repeatedly over time. Measurement of the dependent variable is valid and described with replicable precision. Each dependent variable was measured with a procedure that generated a quantifiable index. Data was collected on the reliability or inter-observer agreement and IOA levels met minimal standards (e.g. IOA = 80%;Kappa = 60%).</p> <p>2. Three or more of rating 3 criteria are fulfilled.</p> <p>1. Two or more of rating 3 criteria are fulfilled.</p> <p>0. Only one of rating 3 criterion is fulfilled.</p> |
| III: Independent Variable | <p>3. Independent variable was described with replicable precision, was systematically manipulated and under the control of the experimenter. Overt measurement of the fidelity of implementation for the independent variable is highly desirable.</p> <p>2. Two out of rating 3 criteria are fulfilled.</p> <p>1. One of rating 3 criterion is fulfilled.</p> <p>0. None of the criteria in rating 3 is fulfilled fully.</p> |
| IV: Baseline | <p>3. The study provided repeated measurement of a dependent variable and established a pattern of responding that could be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur. Baseline conditions was described with replicable precision.</p> <p>2. Two out of rating 3 criteria are fulfilled.</p> <p>1. One of rating 3 criterion is fulfilled.</p> <p>0. None of the criteria in rating 3 is fulfilled fully.</p> |

| Criteria | Quality Ratings |
|---|--|
| V: Experimental Control/Internal Validity | <p>3. The design provided at least three demonstrations of experimental effect at three different points in time. The design controlled for common threats to internal validity and the results documented a pattern that demonstrated experimental control.</p> <p>2. Two out of rating 3 criteria are fulfilled.</p> <p>1. One of rating 3 criterion is fulfilled.</p> <p>0. None of the criteria in rating 3 is fulfilled fully.</p> |
| VI: External Validity | <p>3. Experimental effects were replicated across participants, settings, or materials to establish external validity.</p> <p>2. Replicated across two cases for external validity.</p> <p>1. Replicated across one case for external validity.</p> <p>0. Not replicated across any cases for external validity.</p> |
| VII: Social Validity | <p>3. The dependent variable was socially important and the magnitude of change in the dependent variable resulting from the intervention is socially important. Implementation of the independent variable was practical and cost effective. Social validity was enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts.</p> <p>2. Three or more of rating 3 criteria are fulfilled.</p> <p>1. Two or more of rating 3 criteria are fulfilled.</p> <p>0. Only one of rating 3 criterion is fulfilled.</p> |

Summary of WoE A for Small N Studies

| Study Authors | Dimensions | | | | | | | WoE A Rating |
|---------------------------------|------------|----|-----|----|---|----|-----|--------------|
| | I | II | III | IV | V | VI | VII | |
| Halfon, Bekar & Büşra, 2017 | 3 | 3 | 2 | 1 | 3 | 2 | 3 | 2.4 |
| Ramires, Schwan & Midgley, 2012 | 2 | 2 | 1 | 1 | 0 | 0 | 2 | 1.1 |

Weight of Evidence B (WoE B) – Methodological Relevance

WoE B was used to consider the research design of each study and its relevance for evaluating the effectiveness of MBT intervention with children and young people experiencing psychological distress. Guyatt et al. (1995) informed the criteria and descriptors used for WoE B.

A criterion relevant to the use of follow-up measures was included in WoE B, this was reflected in the higher rating. This considered the potential significance in the long-term effects of MBT intervention. Furthermore, this would highlight the evaluation of MBT effectiveness for children and young people experiencing psychological distress.

A rating of ‘High’ (3) studies must:

- Have randomly assigned participants to the intervention or control group or established group equivalence using statistical analyses.
- Have a minimum of one active control group, no wait-list control or alternative treatment group.
- Have completed pre and post intervention and follow-up assessments using a standardised measure.

A rating of 'Medium' (2) studies must:

- Non-randomisation to allocate participants to the intervention and comparison group.
- Used a comparison group which may not be matched to the intervention group.
- Have completed pre or post intervention or follow-up assessments using a standardised measure.

A rating of 'Low' (1) studies must:

- Non-randomisation to allocate participants to the intervention and comparison group.
- Have no comparison group
- Have completed pre and post intervention or follow up assessments using any type of measure.

Summary of WoE B

| Study Authors | WoE B Rating |
|---------------------------------|---|
| Fonagy et al., 2009 | <input checked="" type="checkbox"/> High <input type="checkbox"/> Medium <input type="checkbox"/> Low |
| Halfon, Bekar & Büşra, 2017 | <input type="checkbox"/> High <input type="checkbox"/> Medium <input checked="" type="checkbox"/> Low |
| Hauber, Boon & Vermeiren, 2017 | <input type="checkbox"/> High <input checked="" type="checkbox"/> Medium <input type="checkbox"/> Low |
| Laurensen et al., 2014 | <input type="checkbox"/> High <input checked="" type="checkbox"/> Medium <input type="checkbox"/> Low |
| Ramires, Schwan & Midgley, 2012 | <input type="checkbox"/> High <input type="checkbox"/> Medium <input checked="" type="checkbox"/> Low |
| Rossouw & Fonagy, 2012 | <input checked="" type="checkbox"/> High <input type="checkbox"/> Medium <input type="checkbox"/> Low |
| Twemlow et al., 2011 | <input type="checkbox"/> High <input checked="" type="checkbox"/> Medium <input type="checkbox"/> Low |

The Halfon, Bekar & Büşra, (2017) and Ramires, Schwan & Midgley, (2012) studies were rated an overall WoE B of Low (1). This was because as small N studies there was no randomisation or comparison group.

Weight of Evidence C (WoE C) – Topic Relevance

WoE C reviewed the relevance and appropriateness of each study weighted against the review question. The WoE C criteria detailed: 1) Intervention based solely on MBT; 2) Each studies relevance to an education setting; 3) The generalisability to male and female population.

To receive a rating of ‘High’ (3) studies must:

- All interventions based on MBT
- Education setting
- Sample includes both males and females

To receive a rating of ‘Medium’ (2) studies must:

- All interventions based on MBT
- Any setting
- Sample includes both males and females

To receive a rating of ‘Low’ (1) studies must:

- Includes an intervention based on MBT
- Any setting
- Any sample characteristics

Summary of WoE C

| Study Authors | WoE C Rating |
|---------------------------------|---|
| Fonagy et al., 2009 | <input checked="" type="checkbox"/> High <input type="checkbox"/> Medium <input type="checkbox"/> Low |
| Halfon, Bekar & Büşra, 2017 | <input type="checkbox"/> High <input type="checkbox"/> Medium <input checked="" type="checkbox"/> Low |
| Hauber, Boon & Vermeiren, 2017 | <input type="checkbox"/> High <input checked="" type="checkbox"/> Medium <input type="checkbox"/> Low |
| Laurensen et al., 2014 | <input type="checkbox"/> High <input type="checkbox"/> Medium <input checked="" type="checkbox"/> Low |
| Ramires, Schwan & Midgley, 2012 | <input type="checkbox"/> High <input type="checkbox"/> Medium <input checked="" type="checkbox"/> Low |
| Rossouw & Fonagy, 2012 | <input type="checkbox"/> High <input checked="" type="checkbox"/> Medium <input type="checkbox"/> Low |
| Twemlow et al., 2011 | <input type="checkbox"/> High <input type="checkbox"/> Medium <input checked="" type="checkbox"/> Low |

Weight of Evidence D (WoE D) – Overall

Weights of Evidence A, B and C were totaled and divided by three to provide WoE D. An overall rating of ‘high’, was categorised as 2.5 or above, ‘medium’ rating between 1.5 and 2.4 and ‘low’ rating 1.4 or below.

Summary of WoE D

| Studies | Quality of Methodology (WoE A) | Relevance of Methodology (WoE B) | Relevance of evidence to the review question (WoE C) | Overall Weighting (WoE D) |
|---------------------------------|--------------------------------|----------------------------------|--|---------------------------|
| Fonagy et al., 2009 | High (2.8) | High (3) | High (3) | High (2.9) |
| Halfon, Bekar & Büşra, 2017 | High (2.4) | Low (1) | Low (1) | Medium (1.5) |
| Hauber, Boon & Vermeiren, 2017 | High (2.8) | Medium (2) | Medium (2) | Medium (2.3) |
| Laurensen et al., 2014 | High (2.8) | Medium (2) | Low (1) | Medium (1.9) |
| Ramires, Schwan & Midgley, 2012 | Low (1.1) | Low (1) | Low (1) | Low (1) |
| Rossouw & Fonagy, 2012 | High (2.8) | High (3) | Medium (2) | High (2.6) |
| Twemlow et al., 2011 | Low (1.4) | Medium (2) | Low (1) | Medium (1.5) |

Appendix D - Coding Protocols

Example of Coding Protocol for Group Designs (Gersten et al. 2005)

Full Study Reference: Fonagy, Peter, Stuart W. Twemlow, Eric M. Vernberg, Jennifer Mize Nelson, Edward J. Dill, Todd D. Little, and John A. Sargent. "A Cluster Randomized Controlled Trial of Child-Focused Psychiatric Consultation and a School Systems Focused Intervention to Reduce Aggression." *Journal of Child Psychology and Psychiatry* 50, no. 5 (May 2009): 607–16

Intervention Name (description of study): Creating a Peaceful School Learning Environment (CAPSLE) is a whole school intervention based on mentalization principles. It aims to improve mentalization and relationship between community members. This intervention was assessed in terms of reports of bullying, bystanding, classroom observations of disruptive and off-task behaviours, and mentalizing behaviours.

Research design: cluster-randomized with stratified restricted allocation controlled trial to assess efficacy after two years and effectiveness after an additional year with minimal support.

Type of Publication: Journal article

Essential Quality Indicators**Describing Participants**

Was sufficient information provided to determine / confirm whether the participants demonstrated the disability/ties or difficulties presented?

Yes

No

N/A

Unknown/Unable to Code

Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?

- Yes
- No
- N/A
- Unknown/Unable to Code

Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

- Yes
- No
- N/A
- Unknown/Unable to Code

Implementation of the Intervention and Description of Comparison Conditions

Was the intervention clearly described and specified?

- Yes
- No A clear description was provided, but without manuals therefore not able to replicate.
- N/A
- Unknown/Unable to Code

Was the fidelity of implementation described and assessed?

- Yes
- No
- N/A
- Unknown/Unable to Code

Was the nature of services provided in comparison conditions described?

- Yes
- No
- N/A
- Unknown/Unable to Code

Outcome Measures

Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalised performance?

- Yes
- No
- N/A
- Unknown/Unable to Code

Were outcomes for capturing the intervention's effect measured at the appropriate times?

- Yes
- No
- N/A
- Unknown/Unable to Code

Data Analysis

Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?

- Yes
- No
- N/A
- Unknown/Unable to Code

Did the research report include not only inferential statistics but also effect size calculations?

- Yes
- No
- N/A
- Unknown/Unable to Code

Desirable Quality Indicators

Was data available on attrition rates among intervention samples? Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?

- Yes
- No
- N/A
- Unknown/Unable to Code

Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures? Were data collectors and/or scorers blind to study conditions and equally un/familiar to examinees across study conditions (if appropriate)?

- Yes
- No
- N/A
- Unknown/Unable to Code

Were outcomes for capturing the intervention's effect measured beyond an immediate post-test?

- Yes
- No
- N/A

Unknown/Unable to Code

Was evidence of the criterion-related validity and construct validity of the measures provided?

Yes

No

N/A

Unknown/Unable to Code

Did the research team assess not only surface features of fidelity implementation (e.g. number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?

Yes

No

N/A

Unknown/Unable to Code

Was any documentation of the nature of instruction or series provided in comparison conditions?

Yes

No

N/A

Unknown/Unable to Code

Did the research report include actual audio or videotape excerpts that capture the nature of the intervention?

Yes

No

N/A

Unknown/Unable to Code

Were results presented in a clear, coherent fashion?

Yes

No

N/A

Unknown/Unable to Code

Example of Coding Protocol for Small N Designs (Horner et al. 2005)

Full Study Reference: Halfon, Sibel, Ozlem Bekar, and Büşra Gürleyen. "An Empirical Analysis of Mental State Talk and Affect Regulation in Two Single-Cases of Psychodynamic Child Therapy." *Psychotherapy* 54, no. 2 (June 2017): 207–19.

Intervention Name (description of study): Long term psychodynamic therapy informed by MBT. Outcome measures

Research design: Case study

Type of Publication: Journal article

Description of Participants and Setting

Participants are described with sufficient detail to allow others to select individuals with similar characteristics; (e.g., age, gender, disability, diagnosis).

Yes

No

N/A

Unknown/Unable to Code

The process for selecting participants is described with operational precision.

- Yes
- No
- N/A
- Unknown/Unable to Code

Critical features of the physical setting are described with sufficient precision to allow replication.

- Yes
- No
- N/A
- Unknown/Unable to Code

Dependent Variable

Dependent variables are described with operational precision.

- Yes
- No
- N/A
- Unknown/Unable to Code

Each dependent variable is measured with a procedure that generates a quantifiable index.

- Yes
- No
- N/A
- Unknown/Unable to Code

Measurement of the dependent variable is valid and described with replicable precision.

- Yes
- No
- N/A
- Unknown/Unable to Code

Dependent variables are measured repeatedly over time.

- Yes
- No
- N/A
- Unknown/Unable to Code

Data are collected on the reliability or inter-observer agreement associated with each dependent variable, and IOA levels meet minimal standards

- Yes
- No
- N/A
- Unknown/Unable to Code

Independent Variable

Independent variable is described with replicable precision.

- Yes
- No
- N/A
- Unknown/Unable to Code

Independent variable is systematically manipulated and under the control of the experimenter.

- Yes
- No
- N/A
- Unknown/Unable to Code

Overt measurement of the fidelity of implementation for the independent variable is highly desirable.

- Yes
- No
- N/A
- Unknown/Unable to Code

Baseline

The majority of single-subject research studies will include a baseline phase that provides repeated measurement of a dependent variable and establishes a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur.

- Yes
- No
- N/A
- Unknown/Unable to Code

Baseline conditions are described with replicable precision.

- Yes
- No
- N/A
- Unknown/Unable to Code

Experimental Control/internal Validity

The design provides at least three demonstrations of experimental effect at three different points in time.

- Yes
- No
- N/A
- Unknown/Unable to Code

The design controls for common threats to internal validity (e.g., permits elimination of rival hypotheses).

- Yes
- No
- N/A
- Unknown/Unable to Code

The results document a pattern that demonstrates experimental control.

- Yes
- No
- N/A
- Unknown/Unable to Code

External Validity

Experimental effects are replicated across participants, settings, or materials to establish external validity.

- Yes
- No, almost, 2 cases, but similar, recommended at least 3 for external validity
- N/A
- Unknown/Unable to Code

Social Validity

The dependent variable is socially important.

- Yes
- No
- N/A
- Unknown/Unable to Code

The magnitude of change in the dependent variable resulting from the intervention is socially important.

- Yes
- No
- N/A
- Unknown/Unable to Code

Implementation of the independent variable is practical and cost effective.

- Yes
- No
- N/A
- Unknown/Unable to Code

Social validity is enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts.

- Yes
- No
- N/A
- Unknown/Unable to Code