## Table of Contents

- **Header** ........................................ 1
- **Abstract** ......................................... 1
- **Plain Language Summary** ........................ 2
- **Summary of Findings for the Main Comparison** 4
- **Background** ....................................... 8
- **Objectives** ........................................ 9
- **Methods** .......................................... 9
- **Results** ........................................... 14
  - Figure 1. ........................................... 15
  - Figure 2. ........................................... 18
  - Figure 3. ........................................... 19
  - Figure 4. ........................................... 22
  - Figure 5. ........................................... 23
  - Figure 6. ........................................... 24
- **Discussion** ....................................... 25
- **Authors’ Conclusions** ............................ 27
- **Acknowledgements** ................................ 27
- **References** ........................................ 28
- **Characteristics of Studies** ....................... 35
- **Data and Analyses** ............................... 66
  - Analysis 1.1. Comparison 1 Treatment effects in meta-analysis, Outcome 1 Joint engagement in mother-child interaction. 66
  - Analysis 1.2. Comparison 1 Treatment effects in meta-analysis, Outcome 2 Emotion recognition from face photographs, TAU control. 67
  - Analysis 1.3. Comparison 1 Treatment effects in meta-analysis, Outcome 3 Joint attention initiations in standardised assessment. 68
- **Additional Tables** ............................... 68
- **Appendices** ....................................... 73
- **Contributions of Authors** ........................ 78
- **Declarations of Interest** .......................... 79
- **Sources of Support** ............................... 79
- **Differences between Protocol and Review** ..... 80

---

*Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)*

Copyright © 2014 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.
Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD)

Sue Fletcher-Watson¹, Fiona McConnell¹, Eirini Manola², Helen McConachie³

¹Moray House School of Education, University of Edinburgh, Edinburgh, UK. ²Puzzle - School for Children with Autism, Athens, Greece. ³Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

Contact address: Sue Fletcher-Watson, Moray House School of Education, University of Edinburgh, St John's Land, Holyrood Road, Edinburgh, EH8 8AQ, UK. sfwatson@staffmail.ed.ac.uk, sfwatson@staffmail.ed.ac.uk.


Review content assessed as up-to-date: 7 August 2013.


Copyright © 2014 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

ABSTRACT

Background

The 'Theory of Mind' (ToM) model suggests that people with autism spectrum disorder (ASD) have a profound difficulty understanding the minds of other people - their emotions, feelings, beliefs, and thoughts. As an explanation for some of the characteristic social and communication behaviours of people with ASD, this model has had a significant influence on research and practice. It implies that successful interventions to teach ToM could, in turn, have far-reaching effects on behaviours and outcome.

Objectives

To review the efficacy of interventions based on the ToM model for individuals with ASD.

Search methods

In August 2013 we searched CENTRAL, Ovid MEDLINE, Embase, CINAHL, PsycINFO, ERIC, Social Services Abstracts, Autism-Data, and two trials registers. We also searched the reference lists of relevant papers, contacted authors who work in this field, and handsearched a number of journals.

Selection criteria

Review studies were selected on the basis that they reported on an applicable intervention (linked to ToM in one of four clearly-defined ways), presented new randomised controlled trial data, and participants had a confirmed diagnosis of an autism spectrum disorder. Studies were selected by two review authors independently and a third author arbitrated when necessary.

Data collection and analysis

Risk of bias was evaluated and data were extracted by two review authors independently; a third author arbitrated when necessary. Most studies were not eligible for meta-analysis, the principal reason being mis-matching methodologies and outcome measures. Three small meta-analyses were carried out.
Main results

Twenty-two randomised trials were included in the review (N = 695). Studies were highly variable in their country of origin, sample size, participant age, intervention delivery type, and outcome measures. Risk of bias was variable across categories. There were very few studies for which there was adequate blinding of participants and personnel, and some were also judged at high risk of bias in blinding of outcome assessors. There was also evidence of some bias in sequence generation and allocation concealment. Not all studies reported data that fell within the pre-defined primary outcome categories for the review, instead many studies reported measures which were intervention-specific (e.g. emotion recognition). The wide range of measures used within each outcome category and the mixed results from these measures introduced further complexity when interpreting results.

Studies were grouped into four main categories according to intervention target/primary outcome measure. These were: emotion recognition studies, joint attention and social communication studies, imitation studies, and studies teaching ToM itself. Within the first two of these categories, a sub-set of studies were deemed suitable for meta-analysis for a limited number of key outcomes.

There was very low quality evidence of a positive effect on measures of communication based on individual results from three studies. There was low quality evidence from 11 studies reporting mixed results of interventions on measures of social interaction, very low quality evidence from four studies reporting mixed results on measures of general communication, and very low quality evidence from four studies reporting mixed results on measures of ToM ability.

The meta-analysis results we were able to generate showed that interventions targeting emotion recognition across age groups and working with people within the average range of intellectual ability had a positive effect on the target skill, measured by a test using photographs of faces (mean increase of 0.75 points, 95% confidence interval (CI) 0.22 to 1.29 points, Z = 2.75, P < 0.006, four studies, N = 105). Therapist-led joint attention interventions can promote production of more joint attention behaviours within adult-child interaction (mean increase of 0.55 points, 95% CI 0.11 to 0.99 points, Z = 2.45, P value = 0.01, two studies, N = 88). Further analysis undermines this conclusion somewhat by demonstrating that there was no clear evidence that intervention can have an effect on joint attention initiations as measured using a standardised assessment tool (mean increase of 0.23 points, 95% CI -0.48 to 0.94 points, Z = 0.63, P value = 0.53, three studies, N = 92). No adverse effects were apparent.

Authors’ conclusions

While there is some evidence that ToM, or a precursor skill, can be taught to people with ASD, there is little evidence of maintenance of that skill, generalisation to other settings, or developmental effects on related skills. Furthermore, inconsistency in findings and measurement means that evidence has been graded of ‘very low’ or ‘low’ quality and we cannot be confident that suggestions of positive effects will be sustained as high-quality evidence accumulates. Further longitudinal designs and larger samples are needed to help elucidate both the efficacy of ToM-linked interventions and the explanatory value of the ToM model itself. It is possible that the continuing refinement of the ToM model will lead to better interventions which have a greater impact on development than those investigated to date.

PLAIN LANGUAGE SUMMARY

A review of evidence on the use of interventions for people with autism spectrum disorder, based on the psychological model ‘Theory of Mind’

Background

The ‘Theory of Mind’ model suggests that people with autism spectrum disorder (ASD) have a profound difficulty understanding the minds of other people, their emotions, feelings, beliefs, and thoughts. It has been proposed that this may underlie many of the other difficulties experienced by people with ASD, including social and communication problems, and some challenging behaviours. Therefore, a number of studies have attempted to teach theory of mind and related skills to people with ASD.

Review question

This review aimed to explore whether a) it is possible to teach theory of mind skills to people with autism and b) whether or not this evidence supports the theory of mind model. Having a ‘theory of mind’ may depend on developing related basic skills, including joint attention (sharing a focus of interest with another person), recognising other people’s emotions from faces or stories, and imitating other people. Therefore, we included intervention studies that taught not just theory of mind itself, but also related skills.

Study characteristics
We found 22 research studies involving 695 participants, which reported on the efficacy of interventions related to theory of mind. The evidence is current to 7th August 2013.

**Key results and the quality of the evidence**

Despite all studies using a high-quality basic methodology (the randomised controlled trial), there was concern over poor study design and reporting in some aspects. While there is some evidence that theory of mind, or related skills, can be taught to people with ASD, there is currently poor quality evidence that these skills can be maintained, generalised to other settings, or that teaching theory of mind has an impact on developmentally-linked abilities. For example, it was rare for a taught skill to generalise to a new context, such as sharing attention with a new adult who was not the therapist during the intervention. New skills were not necessarily maintained over time. This evidence could imply that the theory of mind model has little relevance for educational and clinical practice in ASD. Further research using longitudinal methods, better outcome measures, and higher standards of reporting is needed to throw light on the issues. This is particularly important as the specific details of the theory of mind model continue to evolve.
### Summary of Findings for the Main Comparison

Theory of Mind based interventions compared with wait-list or treatment-as-usual control for autism spectrum disorder.

**Patient or population:** People with autism spectrum disorder  
**Settings:** Schools, home and clinical settings  
**Intervention:** Based on the Theory of Mind theoretical model of autism  
**Comparison:** Most studies incorporate an ‘empty’ control such as treatment-as-usual or wait-list

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Illustrative comparative risks* (95% CI)</th>
<th>No of Participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
</table>
| **Symptom Level: Communication** | Assumed risk | Wong 2010 and Young 2012 report positive effects of intervention on symptom level in the communication domain, while Hadwin 1996 found no effect on conversational skills (this specific outcome is reported in Hadwin 1997) | ADOS: n = 17 (Wong 2010)  
Conversation: n = 30 (Hadwin 1996)  
SCQ: n = 25 (Young 2012) | ⊕⊕⊕⊕⊕  
very low [1] |
| Various measures, including:  
Autism Diagnostic Observation Schedule (ADOS)  
Conversation Skills  
Social Communication Questionnaire (SCQ) (level of eye-contact) | Corresponding risk | See 'Corresponding Risk' | | |

Symptom Level: Social Interaction  
Various measures, including:  
Autism Diagnostic Observation Schedule (ADOS)  
See 'Corresponding Risk'  
Fewer than half of the relevant included studies report positive effects of intervention on symptom level in the social interaction domain  
ADOS: n = 17 (Wong 2010)  
CSBS: n = 48 (Landa 2011)  
ESCS: n = 200 | ⊕⊕  
low [2] |

Three included studies report outcomes in this area of clinical relevance. Each one uses a different assessment to capture change in this domain. One study uses an unstandardised measure, though it is designed to capture change over time (Hadwin 1996). The other two studies use standardised measures of communication skills but neither of these were designed to capture change over time nor to be used as intervention outcome measures.

Here we include both standardised assessments and direct observations of social behaviours. Eleven included studies report...
Communication and Symbolic Behaviour Scale (CSBS)
Early Social Communication Scales (ESCS)
PDD-BI social approach sub-scale
Precursors of Joint Attention Measure
Social (PJAM) Communication Questionnaire (SCQ)
Social Emotional Scale (SES) (Bayley-III)
Social Skills Rating System (SSRS)
Vineland Adaptive Behaviour Scales (socialisation) (VABS) and
Other social interaction (SI) observations

In addition some studies report mixed findings across methods. For example, Goods 2013 and Kaale 2012 report some positive effects measured in observations, but null findings from the ESCS. Conversely Kim 2009 (outcomes reported in Kim 2008) and Wong 2013 find significant effects measured by the ESCS but not all other measures. In the case of Wong 2013 this is further complicated by a mixed output from the ESCS where a significant effect is found for one scored item but not another. Similar findings are reported by Schertz 2013 and Kasari 2010 who find positive effects of intervention on some observed behaviours but not others. Both studies which report the impact of an emotion recognition intervention on generalised social skills do not find significant effects on their chosen outcomes (Williams 2012; Young 2012).

Outcomes in this area of clinical relevance. There is wide variety in the choice of assessments to capture change in this domain, though most are based on standardised assessments and are often designed to capture change over time.
### General Communication Ability (e.g. vocabulary)
- Mullen Scales of Early Learning (MSEL)
- Reynell Developmental Language Scales

*Schertz 2013* reports significant intervention effects on receptive language and a non-significant but moderate sized effect ($d = 0.78$) for expressive language scores. At a one-year follow-up *Kasari 2006* likewise report intervention effects on expressive language, which were significantly greater for the joint attention intervention compared with both control group and symbolic play interventions. However these effects on expressive vocabulary were not sustained four years later (*Kasari 2012b*). In addition, a methodologically strong study (*Landa 2011*) reports no effects on expressive language.

- MSEL expressive: $n = 71$ (Landa 2011, Schertz 2013)
- MSEL receptive: $n = 23$ (Schertz 2013)
- Reynell: $n = 58$ (Kasari 2006)

Though this has commonly been used as an outcome measure in generalised social skills interventions for children with ASD, only three of the studies included in this review report a general communication ability outcome measure.

### Theory of Mind ability

- Various measures, including:
  - False-belief tasks
  - Happe’s Strange Stories
  - Faux-Pas Recognition Test
  - NEPSY-II ToM tasks
  - The ToM Test

*Two studies report some positive effects of intervention on ToM ability (Begeer 2011; Fisher 2005) one reports no impact on directly-assessed ToM ability (Solomon 2004) and one reports a reduction at follow-up in ToM ability for the intervention group specifically (Williams 2012).*

- False belief: $n = 27$ (Fisher 2005)
- Happe SS, & Faux-Pas RT: $n = 18$ (Solomon 2004)
- NEPSY: $n = 55$ (Williams 2012)
- ToM Test: $n = 36$ (Begeer 2011)

Four included studies report outcomes in this area of principally theoretical relevance. There is wide variety in the choice of assessments to capture change in this domain, though most are based on standardised assessments and are closely linked to the intervention target skill.

---

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).*

**CI:** Confidence interval; **RR:** Risk Ratio
<table>
<thead>
<tr>
<th>GRADE Working Group grades of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High quality:</strong> Further research is very unlikely to change our confidence in the estimate of effect.</td>
</tr>
<tr>
<td><strong>Moderate quality:</strong> Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.</td>
</tr>
<tr>
<td><strong>Low quality:</strong> Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.</td>
</tr>
<tr>
<td><strong>Very low quality:</strong> We are very uncertain about the estimate.</td>
</tr>
</tbody>
</table>

1. Risk of bias (-1); inconsistency (-2): Since the studies included here are of variable methodological quality and report mixed findings this evidence is considered to be of Very Low quality.
2. Risk of bias (-1); inconsistency (-1): The studies included here are of variable methodological quality and report mixed findings from a wide variety of measures. There is a collection of studies reporting on the ESCS (some of which are summarised in Analysis 1.1), but within this group findings are once again mixed. Indeed, even within a single study and measure there may be inconsistency in evidence for intervention efficacy. It is therefore impossible to be confident about the impact of Theory of Mind interventions on social interaction domain symptom level and the evidence quality is rated as Low.
3. Risk of bias (-1); inconsistency (-1); low sample size (-1): These mixed outcomes from only a handful of studies must be judged of Very Low Quality until they can be resolved by additional high-quality evidence.

It is challenging to divide communication and social interaction for measures which tap into both of these qualities. However for the purposes of this table, we have identified measures which are based on observation of an interpersonal interaction as falling into the Social Interaction Domain.

A number of included studies report on measures of emotion recognition and imitation skill. While these are suitable outcomes for the respective interventions, and highly associated with ASD profiles, these cannot be categorised into the domains for this Summary of Findings table, and therefore are not addressed here.
BACKGROUND

Description of the condition

Autism spectrum disorder (ASD) is an umbrella term used to describe all people diagnosed as showing symptoms within two core criteria: communication and social deficits, and fixed or repetitive behaviours (APA 2013). The ASD label replaces former subtypes, including autism, pervasive developmental disorder - not otherwise specified (PDD-NOS), and Asperger's syndrome (AS) (APA 1994). Likewise, the single communication and social interaction cluster is derived from what was originally two separate domains of impairment in communication and social interaction (APA 1994). These difficulties often make it very hard for people with ASD to be successful members of society and can present very serious challenges to parents, teachers, and other professionals.

Prevalence estimates of ASD diagnosis in children have been rising significantly in recent years with an authoritative systematic review estimating global prevalence of pervasive developmental disorders at 62 per 10,000 and autistic disorder at 17 per 10,000 (Elsabbagh 2012). Figures may be higher in more developed countries (e.g. Baird 2006). This represents a more-than-threefold increase on previously published figures, which estimated autism prevalence at about 5 per 10,000 (Fombonne 2001). While there are methodological differences between prevalence studies, the rising prevalence of ASD has been well-documented across Western countries, including Europe, Australia, and the USA (e.g. Yeragin-Allsopp 2003; Williams 2006; Arlottir 2007; Kogan 2009; Nassar 2009). There has been significant debate about the cause of the recent rise in prevalence of ASD, but the influence of increased awareness of the disorder among health professionals and the community at large, and the role of diagnostic substitution, should not be underestimated (Croen 2002; Arlottir 2007). There are other candidate explanations, including the possibility of environmental causes of the rising prevalence estimates, though, as yet, there is no good empirical evidence for these (Rutter 2005). Baird et al (Baird 2006) conclude that “whether the increase is due to better ascertainment, broadening diagnostic criteria, or increased incidence is unclear” (p. 210).

Within the disorder there is a male to female ratio of 4:1 or 5:1 (Baird 2006; Kogan 2009), as noted in the set of case studies, which defined the condition for the first time (Kanner 1943). ASDs have this feature in common with most other neurodevelopmental disorders (such as attention deficit hyperactivity disorder (ADHD), dyslexia, dyspraxia), though to a greater extent. As yet, there is no empirical evidence for systematic differences between male and female individuals with ASD (Hartley 2009).

Theory of Mind

The term 'Theory of Mind' (ToM) describes the ability to understand another's thoughts, beliefs, and other internal states and was originally applied to the study of non-human primate cognition (Premack 1978). The term has since been developed in a number of different directions (e.g. Carruthers 1996), including in research into ASD. The first application of the term in ASD research was in an experiment which used false-belief paradigms to explore ToM in children with autism (Baron-Cohen 1985). In this study, children were presented with a scenario in which a doll, Sally, 'believed' her marble was in the basket where she left it. However, the child and experimenter knew that while Sally was elsewhere, another doll had moved the marble into a box. The key question was “Where will Sally look for her marble?” Typically-developing children from the age of four years, sometimes earlier, can correctly ascertain that Sally will look in the basket; she holds a false belief about the location of the marble (Wellman 2001). Children with ASD are much less likely to give a correct answer to this question at age four years. They normally claim that Sally will look in the box, in accordance with reality, but incompatible with Sally's knowledge of the situation.

Research into ToM in children and adults with ASD has been prolific over the last 25 years (e.g. Baron-Cohen 2000). While the details are subject to debate, it is widely accepted that people with ASD do not possess a fully-functioning theory of mind; even high-functioning adults with ASD may struggle with complex ToM tasks (Ponnet 2004). ToM has been placed in a developmental context, consisting of a range of precursor skills, including following eye-gaze, establishing joint attention, imitation, pretend play, and emotion recognition (Melzoff 1993; Baron-Cohen 1995; Charman 2000; Wellman 2000; Ruffman 2001). ToM then also links to subsequent social and communication skills, including the development of language (Tager-Flusberg 2000; Garfield 2001). As a result, many believe that failures of ToM are central to explaining the difficulties experienced by people with ASD (though not a sufficient explanation). Therefore, ToM and its precursor skills are targets for interventions.

Description of the intervention

A 'Theory of Mind intervention' is a treatment or therapy, which is explicitly or implicitly based on the Theory of Mind (ToM) cognitive model of ASD. ToM interventions target those skills which are either potential components or precursors of ToM (Swettenham 2000). One example of an intervention targeting such skills is using 'thought-bubbles' to teach children with ASD to understand others' thoughts and beliefs by illustrating these in bubbles (as in a cartoon) (Parsons 1999). Specific precursor skills can also be taught such as helping a child to make eye-contact to accompany pointing to an object of interest (joint attention). More detail on which interventions are eligible for inclusion in this review is given in the Methods section, but we will only consider interventions that explicitly target ToM skills.

ToM interventions can be contrasted with other types of treatment for ASD. Many intervention models focus on behaviour...
management and personal skills training, using a basic conditioning model for learning (repetition; rewarding desirable behaviour; 'punishing' or ignoring behaviour that the therapist finds undesirable such as tantrums). In addition, most management strategies for ASD occur within a fairly structured timetable as people with ASD tend to feel more comfortable following familiar routines in a consistent environment, and respond very poorly to change.

The ToM model of autism suggests that the social and communication difficulties that are characteristic of the syndrome stem from a failure to develop an intact ToM. Certainly there is evidence that ToM is correlated with real-life social skills (Frith 1994) and symptomatology (Joseph 2004). Certain ToM precursor skills also have a direct relationship with symptoms (Mundy 1994). Therefore, training in ToM, or in the precursor or component skills of ToM, should alleviate the social and communication difficulties experienced by individuals with the disorder. For example, a targeted joint attention intervention for autism produced improvements in children's responsiveness to joint attention opportunities and also improved sharing and language (Kasari 2006; Kasari 2008), indicating that ToM interventions may have consequences for wider developmental abilities.

It is possible that interventions targeting different ToM skills will produce varied types of change in participants, and the extent of change may vary. The method of delivery of the intervention may also produce different outcomes. For example, one might expect an intervention delivered by a trained therapist to have greater impact than one delivered by parents. An intervention taught in school may have a different impact to one delivered in the home. The duration of the intervention may also be significant. Deficits in ToM and related skills vary with age (Happe 1995), IQ (Ozonoff 1991a; Happe 1994; Bowler 1997), specific diagnosis (Ozonoff 1991b; Bowler 1992) and verbal ability (Happe 1995; Garfield 2001). As a result, the specific skill being targeted, the method of intervention delivery, its duration and individual differences between participants in ToM intervention studies will be important factors for consideration and for statistical analysis in this review.

How the intervention might work

In a chapter reviewing evidence for the possibility of teaching ToM to individuals with autism, Swettenham states (p. 442) that "a successful method for teaching theory of mind may alleviate the impairments in social interaction that are so debilitating in autism" (Swettenham 2000).

The ToM model of autism suggests that the social and communication difficulties that are characteristic of the syndrome stem from a failure to develop an intact ToM. Certainly there is evidence that ToM is correlated with real-life social skills (Frith 1994) and symptomatology (Joseph 2004). Certain ToM precursor skills also have a direct relationship with symptoms (Mundy 1994). Therefore, training in ToM, or in the precursor or component skills of ToM, should alleviate the social and communication difficulties experienced by individuals with the disorder. For example, a targeted joint attention intervention for autism produced improvements in children's responsiveness to joint attention opportunities and also improved sharing and language (Kasari 2006; Kasari 2008), indicating that ToM interventions may have consequences for wider developmental abilities.

It is possible that interventions targeting different ToM skills will produce varied types of change in participants, and the extent of change may vary. The method of delivery of the intervention may also produce different outcomes. For example, one might expect an intervention delivered by a trained therapist to have greater impact than one delivered by parents. An intervention taught in school may have a different impact to one delivered in the home. The duration of the intervention may also be significant. Deficits in ToM and related skills vary with age (Happe 1995), IQ (Ozonoff 1991a; Happe 1994; Bowler 1997), specific diagnosis (Ozonoff 1991b; Bowler 1992) and verbal ability (Happe 1995; Garfield 2001). As a result, the specific skill being targeted, the method of intervention delivery, its duration and individual differences between participants in ToM intervention studies will be important factors for consideration and for statistical analysis in this review.

Why it is important to do this review

To date, there is no comprehensive review of ToM interventions for autism, despite the fact that the first study attempting to teach ToM to individuals with autism was published in 1995 (Ozonoff 1995). This review will be of relevance to both the clinical and academic research communities, since ToM interventions not only have the potential to benefit people with ASD, but also provide a unique and rigorous way to test the theoretical model on which they are based.

OBJECTIVES

To assess the effect of interventions, based on the Theory of Mind (ToM) model, for autism spectrum disorders (ASD), on symptoms in the core diagnostic domains of social and communication impairments in autism, and on language and ToM skills. In addition, in so-doing, to test the applied value of the ToM model of autism.

METHODS

Criteria for considering studies for this review

Types of studies

Randomised and quasi-randomised trials (defined as trials in which allocation was made by, for example, alternate allocation or allocation by date of birth).

Types of participants

Participants of any age with a diagnosis of an ASD, including autism, atypical autism, Asperger's syndrome, and PDD-NOS, according to either ICD-10 (Internal Classification of Diseases), DSM-IV or DSM-V (Diagnostic Statistical Manual of Mental Disorders) criteria. All diagnostic categories could be included since the validity of differentiating between categories on the spectrum is not well established (Klin 2005). Furthermore, the ToM cognitive model does not distinguish, on a qualitative basis, between different forms of ASD. Participants must have received a 'best estimate' clinical diagnosis, confirmed by the study authors. That is, at a minimum, diagnosis by a multidisciplinary clinical team using standard procedures with reference to the international classification systems. Use of a particular diagnostic tool, such as the Autism Diagnostic Observation Schedule (ADOS) (Lord 2000) or the Autism Diagnostic Interview (ADI-R) (Lord 1994), was desirable but not required. Co-morbid cases were also eligible for inclusion since these individuals are just as needful of intervention for their specifically autistic difficulties.
Types of interventions

Interventions eligible for inclusion in this review:

1. explicitly state that they are designed to teach ToM; or
2. explicitly state that they are designed to teach precursor skills of ToM; or
3. explicitly state that they are based on or inspired by ToM models of autism; or
4. explicitly state that they aim to test the ToM model of autism.

We reiterate that ToM (theory of mind) describes the ability to understand another's thoughts, beliefs, and other internal states and is encapsulated in a test of false belief. Prior to the development of false-belief understanding (at about four years old in typical development), associated precursor skills are in evidence such as joint attention, imitation, and emotion recognition. Relevant interventions include those which explicitly teach children to understand others' mental states (e.g. using visual representations of mental states McGregor 1998) and those which use naturalistic teaching to develop imitation skills (Heimann 2006).

The following kinds of interventions are not included in this review:

1. interventions which do not meet the criteria given above;  
2. medical interventions (e.g. risperidone for aggression in ASD);  
3. dietary interventions (e.g. gluten-free and casein-free diets);  
4. interventions which target a particular behaviour rather than a cognitive skill (e.g. over-sensitivity to light modified using colour spectacles; sleep difficulties modified using applied behavioural analysis);  
5. language-focused interventions (e.g. to make requests using the Picture Exchange Communication System or spoken single words);  
6. interventions which have a broad-base both in terms of methods (e.g. combining computerised learning with parent training and social skills groups) and targets (i.e. addressing a range of social communication skills, some which are ToM-linked but also more general skills such as turn-taking, friendship skills, and conversation).

ToM interventions are compared with the following conditions, where these are used:

1. treatment-as-usual/wait-list control;  
2. ‘placebo’ interventions, for example a ‘contact control’ such as watching Thomas the Tank Engine DVDs (e.g. Young 2012);  
3. intervention with no therapeutic content, (e.g. group leisure activities (Baghdadi 2013).  

All ‘doses’ (that is the number and length of treatment sessions per week), durations, and methods were eligible for inclusion.

Types of outcome measures

Outcome measures do not form part of the criteria for inclusion of studies in this review.

Primary outcomes

Primary outcomes at a participant symptom level, measured using standardised diagnostic assessments or clinical report. Outcomes will be in each of two symptom domains that have until recently been used in clinical diagnosis and are followed by most diagnostic tests for autism. These are as follows, with examples of outcomes in each category as measured by the ADOS (Lord 2000) or ADI (Lord 1994).

1. Communication: overall level of non-echoed language; stereotyped or idiosyncratic use of words or phrases; pointing; gestures; conversation.
2. Social function: unusual eye-contact; facial expressions directed to others; spontaneous initiation of joint attention; shared enjoyment in interaction; quality of rapport.

The third diagnostic domain of Restricted and Repetitive Behaviours (imaginative play or creativity; unusual sensory interests; unusually repetitive interests or stereotyped behaviours; compulsions or rituals) is not included as an expected primary outcome.

Secondary outcomes

In addition, the following secondary outcomes will be included.

PARTICIPANT, direct measurement

- Intervention-specific: change in targeted cognitive skill such as false-belief understanding
- Change in participant behaviour or quality of interpersonal interaction, or both, measured by direct observation.

PARENT, teacher (or other individual in caring or educational relationship to the participant) report

- Change in participant behaviour and skills or deficits such as: adaptive skills; school success; challenging behaviours; social participation measured by parent, teacher or other report
- Acceptability of intervention (time, cost)

OTHER

- Intervention process measures e.g. rate of drop-out
- Economic data e.g. financial cost of intervention; time commitment required

Main outcomes for 'Summary of findings' table

The following outcomes measures are specified for a 'Summary of findings' table:

1. symptom level, communication domain;  
2. symptom level, social interaction domain;  
3. general communication ability (e.g. vocabulary);  
4. 'Theory of Mind' ability (e.g. false-belief test score).

Where data were available, we planned to organise outcomes into three time points: immediately post-treatment; medium-term outcome (up to six months post-treatment); and long term (more than six months post-treatment).

The Summary of findings for the main comparison reports on these outcomes and also includes an estimate of the quality of evidence in each category.
Search methods for identification of studies

The complex nature of ToM interventions makes them difficult to capture adequately using search terms. Therefore, to avoid missing relevant studies, we used a highly sensitive search strategy with just two concepts: the condition (ASD) and a search filter to find RCTs. The core search strategy was developed in Ovid MEDLINE and uses the Cochrane highly sensitive search strategy for identifying randomised trials (Lefebvre 2008). The MEDLINE strategy was adapted for other databases using appropriate syntax and controlled vocabulary. The initial searches were run in July 2010 without any date or language restrictions. We last updated the searches on 6 August 2013, apart from ASSIA which was no longer available to us.

Electronic searches

We searched the following databases in August 2013.
- Cochrane Central Register of Controlled Trials (CENTRAL) 2013, Issue 7, part of The Cochrane Library.
- Ovid MEDLINE(R) 1946 to July Week 4 2013.
- CINAHLPlus 1937 to current.
- PsycINFO 1806 to July Week 5 2013.
- ERIC 1966 to current.
- Applied Social Sciences Index and Abstracts: ASSIA (CSA) 1979 to current.
- Social Services Abstracts 1979 to current.
- metaRegister of Controlled Trials (controlled-trials.com/mrct/).
- ICTRP (apps.who.int/trialsearch/).
- UKCRN - UK Clinical Trials Network (public.ukcrn.org.uk/search/).
- ClinicalTrials.gov (clinicaltrials.gov/).
- Autism Data (autism.org.uk/autismdata/).

The search strategies for each source are in Appendix 1.

Searching other resources

In addition to searches of electronic databases, we contacted key authors in the field directly and asked them to provide any relevant published, unpublished or in-progress data, including postgraduate dissertations. We also searched the bibliographies of key articles for citations of papers not found electronically. Searches were made for in-progress, or unpublished clinical trials. Finally, we searched the online databases of journals that regularly publish work on this topic. These journals were the Journal of Autism and Developmental Disorders, Journal of Child Psychology and Psychiatry, and Autism: International Journal of Research and Practice. We also searched the proceedings of the International Meeting for Autism Research.

Data collection and analysis

Selection of studies

All citations sourced from the search strategy were transferred to EndNote, a reference management programme. Initial screening of titles and abstracts by an experienced research assistant (EM or FMcC) eliminated all those citations obviously irrelevant to the topic, for example, prevalence studies, studies unrelated to ASD, and single case studies. Thereafter, two review authors (SFW and either EM or FMcC) assessed and selected studies for inclusion from the group of superficially relevant studies. In the event of a disagreement, resolution was reached in discussion with a third author (HM), if necessary following inspection of the full paper.

Data extraction and management

Two review authors (SFW and either EM or FMcC) independently extracted data from selected trials using a specially designed data extraction form. Extracted data included methods (dose and frequency of intervention); diagnostic description of participants, and type of intervention, including target, intensity, duration, and method of application (parent-mediated, therapist, school-based etc.). Disagreements were resolved in consultation with a third author (HM).

Assessment of risk of bias in included studies

Two review authors (SFW and either EM or FMcC) independently assessed the risk of bias of included studies in the following domains: sequence generation; allocation concealment; blinding of participants and personnel; blinding of outcome assessment; incomplete outcome data; selective outcome reporting; other sources of bias. We used The Cochrane Collaboration tool for assessing risk of bias in these areas. The process involved recording the appropriate information for each study (e.g. describing the method used to conceal allocation in detail) and evaluating whether there was risk of bias in that area (e.g. was allocation adequately concealed?). Any disagreement was resolved by referral to a third author (HM).

We contacted authors to supply missing information from 16 included studies (Bolte 2002; Solomon 2004; Fisher 2005; Golan 2006; Kasari 2006; Kim 2009; Golan 2010; Kasari 2010; Ryan 2010; Wong 2010; Hopkins 2011; Ingersoll 2012; Young 2012; Baghdadli 2013; Schertz 2013; Wong 2013) and information was received from the majority of authors with the exception of (Solomon 2004; Fisher 2005; Golan 2006; Golan 2010; Wong 2010; Baghdadli 2013).

Studies were allocated to categories according to our evaluation of each area or potential risk of bias as follows:
**Random sequence generation**

Low risk of bias: adequate sequence generation as indicated by reference to, e.g., random number table, coin tossing, shuffled cards or envelopes, throwing dice, drawing lots.

Unclear (or moderate) risk of bias: indicates uncertainty about whether the sequence was randomly generated.

High risk of bias: a non-random component is described such as sequence generation by odd or even date of birth, by geographical location or by date of entry to the study.

**Allocation concealment**

Low risk of bias: participants and investigators enrolling participants unable to foresee assignment as indicated by reference to, e.g., central allocation, opaque envelope procedure, allocation by an independent partner outside the research team.

Unclear (or moderate) risk of bias: indicates uncertainty about whether the allocation was concealed.

High risk of bias: participants and investigators enrolling participants may have been able to foresee assignment as indicated by reference to an open random allocation schedule (e.g. random numbers list), unsealed or non-opaque envelopes, alternate allocation, allocation by non-random criteria such as by date of birth.

**Blinding of participants and personnel**

Low risk of bias: participants and personnel blinded to study hypotheses and treatment condition, or incomplete blinding but authors judge that outcome is unlikely to be influenced by lack of blinding. Examples might be when participants are very young and/or low-functioning people with autism and are unlikely to be aware of intervention targets, and where outcome is assessed using a measure resilient to performance bias such as computerised assessment. We note that in behavioural studies, such as those included in this review, it is rarely possible to blind participants and/or personnel.

Unclear (or moderate) risk of bias: indicates uncertainty about whether blinding was consistent, perhaps due to insufficient information being available, or partial blinding (e.g. of participants but not personnel).

High risk of bias: participants and personnel not blind to study hypotheses or treatment condition, and outcome likely to be influenced by this lack of blinding.

**Blinding of outcome assessment**

Low risk of bias: outcome examiners and scorers blinded to participant group membership, or blinding of some outcome assessors with good evidence of agreement between blinded and unblinded raters on outcome measures, or outcome assessors not blind but outcome measurement unlikely to be influenced by this lack of blinding.

Unclear (or moderate) risk of bias: indicates uncertainty about whether blinding was consistent, perhaps due to insufficient information being available.

High risk of bias: outcome examiners and scorers not blind to participant group membership, and outcome likely to be influenced by this lack of blinding.

**Incomplete outcome data**

Low risk: no missing data, or reasons for missing outcome data unlikely to be related to true outcome, or missing data balanced across groups with similar reasons in each case.

Unclear risk of bias: insufficient reporting of attrition or exclusions to permit accurate judgement.

High risk of bias: reasons for missing data likely to be related to true outcome, with imbalance in numbers between groups or different reasons between groups.

**Selective reporting**

Low risk: study protocol available and all pre-specified outcomes are reported in the pre-specified way, or clear from the published reports that all expected outcomes are included.

Unclear (moderate) risk of bias: insufficient information to permit accurate judgement.

High risk of bias: not all of the study’s primary outcomes have been reported, or outcomes which were not pre-specified are reported, or one or more primary outcomes have been reported for only a sub-set of the sample, or one or more outcomes are reported incompletely so that they cannot be entered into a meta-analysis.

**Other bias**

Low risk of bias: the study appears to be free of other sources of bias.

Unclear (or moderate) risk of bias: there may be an additional risk of bias but there is insufficient information to fully assess this risk, or it is unclear whether the risk would introduce bias in study results.

High risk of bias: the study has one important additional risk of bias such as a source of bias related to the study design, or claims of fraudulence.

In each case, only studies where the assessment of overall risk falls into categories ‘Low’ or ‘Unclear/Moderate’ have been included in subsequent analyses.

**Measures of treatment effect**

**Binary and categorical data**

No studies reported binary outcome data in the current review version. Should they be included in future updates, methods for
analysing them appear in the published protocol for the review (Fletcher-Watson 2010).

Continuous data

Where standardised assessment tools generated a continuous score as the outcome measure, and means and standard deviations were reported or provided by the authors, comparisons were made between the means of these scores. When selecting studies for possible meta-analysis the following criteria were of principal importance.

- Similarity of trial design - especially whether the ‘control’ condition had therapeutic content or not
- Similarity of intervention target
- Similarity of outcome measure - the quality being measured, the unit of measurement, and the method of measurement (e.g. parent-report, video coding, standardised assessment)

Where measures were on different scales but those scales were clinically homogeneous, meta-analyses used standardised mean difference with Hedges’ g correction for small sample sizes.

Unit of analysis issues

No cluster-randomised trials were included in the current review version. Methods for analysis are recorded in additional Table 1.

Dealing with missing data

Missing data were assessed for each individual study according to the reports provided by authors. For included studies reporting drop-out, we reported the number of participants included in the final analysis as a proportion of those participants who began the intervention (see Characteristics of included studies). Reasons for missing data are also reported (that is, whether data are missing at random or not). In all cases, we concluded that data were missing at random, and the remaining data were analysed and the missing data ignored.

Where summary data are missing, trial authors were contacted. If no reply was forthcoming or the required summaries were not made available, the study was included in the review and we assessed and discussed the extent to which its absence from meta-analysis affects the review results (e.g. Bolte 2002).

No studies reported the loss of significant quantities of data, without sufficient explanation, and there was no evidence of non-random missing data. Therefore, the review authors agree that the conclusions of individual studies are not compromised by missing data. The extent to which the results of the review may be altered by the missing data is assessed and discussed (Quality of the evidence).

Additional procedures for dealing with non-random missing data in future appear in the published review protocol (Fletcher-Watson 2010) and Table 1.

Assessment of heterogeneity

Consistency of results was assessed visually and by a Chi² test. Where meta-analysis included only a small number of studies, or where studies had small sample sizes, a P value of 0.10 was applied for statistical significance. In addition, since Chi² can have low power when only a few studies or studies of a small sample size are available, we used the I² statistic to measure the amount of observed variability in effect sizes that can be attributed to true heterogeneity (Higgins 2008).

Assessment of reporting biases

Where sufficient studies were found, funnel plots were inspected to investigate any relationship between effect size and sample size. Such a relationship could be due to publication or related biases, or due to systematic differences between small and large studies.

Data synthesis

Data synthesis was performed using RevMan 5.2. Binary data were not reported in any of the included studies but could be assessed in future review versions. Where two or more studies suitable for inclusion were found, and the studies were considered to be homogenous, a meta-analysis was performed on the results. Homogeneity decisions were based on examination of a series of factors identified in the review protocol including the following.

- Similarity in intervention delivery type (e.g. therapist-led, parent training)
- Similarity in intervention target skill (e.g. emotion recognition, imitation, joint attention)
- Similarity in participant populations (e.g. intellectual level in the normal or low range, specific autism diagnostic category, age)

In addition, the following two further factors were developed post hoc in response to the wide variability in study design and outcome measure found.

- Similarity in primary outcome measurement
- Similarity in comparison group status (e.g. did the study compare two different interventions or compare an intervention with a wait-list or treatment-as-usual control)

It is essential to distinguish between measures of primary outcome when assessing intervention efficacy for two main reasons. The first is that there is significant evidence that people with ASD do not generalise skills across contexts. For example, Golan 2010 found differences in outcome measures by close and distant generalisation tasks even though these were all measures of emotion and mental state recognition. Therefore, studies measuring outcome using tasks which differ in complexity and in connection to the teaching context should not be compared directly. The second reason is that the method of measurement can produce widely varying distributions, which are not amendable to combination. For example, Kasari 2010 measured percentage of total time that a mother-
child dyad were jointly engaged, while Goods 2013 reported the number of instances that specific types of joint engagement were observed. It is not possible to combine these two variables, which have also been collected in different settings (laboratory artificial mother-child play versus naturalistic classroom observation) and over different periods of time.

Comparison group status is another key consideration when combining studies. A study that shows an intervention effect compared with a 'placebo' group or compared with another intervention may have a smaller effect size than one comparing intervention and wait-list control. However, the former study has the more powerful design and so this smaller effect should be more influential on conclusions.

A random-effects model analysis was used since we do not assume that each study is estimating exactly the same quantity.

**Subgroup analysis and investigation of heterogeneity**
Subgroup analyses were not possible in this version of the review. Dimensions for possible future subgroup analyses are included in additional Table 1.

**Sensitivity analysis**
Sensitivity analysis was not possible for this version of the review. Details of planned future sensitivity analyses are included in additional Table 1.

### RESULTS

#### Description of studies
See Characteristics of included studies, Characteristics of excluded studies.

#### Results of the search
Searches were carried out in July 2010, and again in July 2012, and August 2013, yielding 18,368 records of potential relevance after de-duplication (July 2010: 11,822 records, July 2012: 4171 records, August 2013: 2375 records). Assessment of titles and abstracts and elimination of duplicates between the searches resulted in a list of 99 records for closer examination (Figure 1). One of these articles is only available in French and is currently awaiting classification pending translation (Baghdadli 2010).
Figure 1. Study flow diagram

24,966 records identified through database searching → 6598 duplicates discarded

18,368 records screened → 18,269 irrelevant records eliminated

68 full-text articles excluded:
- not theory-of-mind linked n=23
- Not RCT n=19
- no data reported n=10
- broad-base intervention n=12
- unsuitable design (e.g., experimental pilot) n=2
- Diagnosis of participants invalid n=1
- Not ASD n=1

99 full-text articles assessed for eligibility → 31 articles potentially eligible for inclusion

31 articles potentially eligible for inclusion → 1 article awaiting classification

22 studies (from 30 reports) included in review
Included studies

We included 22 studies involving 695 participants in this review; in each case the main study is reported in a published journal article and the dates of publication span from 1996 to 2013. There were 17 studies reported in a single published journal article: Solomon 2004; Fisher 2005; Golan 2006; Golan 2010; Kasari 2010; Landa 2011; Ryan 2010; Wong 2010; Begeer 2011; Hopkins 2011; Kaale 2012; Williams 2012; Young 2012; Baghdadi 2013; Goods 2013; Schertz 2013; Wong 2013. In the case of Golan 2006, two studies are reported of which only Experiment One is an RCT, and therefore only this first data set is included in the review. In addition, there are five studies for which data have been reported in multiple outputs. These are a therapist-led theory of mind intervention (Hadwin 1996), a computerised emotion recognition intervention study (Bolte 2002), an imitation intervention study (Ingersoll 2012), a music therapy study (Kim 2009), and a joint attention and symbolic play intervention conducted by Kasari and colleagues (Kasari 2006). In the case of the Kasari study, one output is an unpublished PhD thesis (Arora 2008).

All 22 studies described themselves as randomised controlled trials, and they were conducted in a wide variety of locations: Scandinavia (Bolte 2002; Kaale 2012); mainland Europe (Begeer 2011; Baghdadi 2013); the UK and Ireland (Hadwin 1996; Fisher 2005; Golan 2006; Golan 2010; Ryan 2010); the Far East (Kim 2009; Wong 2010); Australia (Williams 2012; Young 2012); and the USA (Solomon 2004; Kasari 2006; Kasari 2010; Landa 2011; Hopkins 2011; Ingersoll 2012; Goods 2013; Schertz 2013; Wong 2013).

Participant baseline characteristics

Participants varied widely in age-range from preschoolers (e.g. Kasari 2006) to adolescents and adults (e.g. Bolte 2002) but a majority focused on either pre-school or primary-school aged children (see Characteristics of included studies). Almost all studies included both boys and girls, though the proportion of male participants was much higher than females, corresponding to the known greater prevalence of diagnosed ASD in males (Kogan 2009). Four studies reported an all-male sample (Bolte 2002; Solomon 2004; Kim 2009; Baghdadi 2013). For all studies, a diagnosis of an ASD was a requirement for inclusion. A large proportion confirmed diagnosis using a clinical instrument such as the Autism Diagnostic Observation Schedule (ADOS, Lord 1994) or the Childhood Autism Rating Scale (CARS, Schopler 1986). Two studies accepted prior clinical diagnosis as adequate (Hadwin 1996; Fisher 2005), but these also instituted a checklist confirming that all diagnostic criteria were met. Participants were reported as having a range of ASD diagnoses, including autism, autism spectrum disorder, pervasive developmental disorder - not otherwise specified (PDD-NOS), high-functioning autism (HFA), and Asperger’s syndrome (AS). Studies recruiting participants with HFA and/or AS had participants in the adolescent and adult age-range (e.g. Bolte 2002; Golan 2006) or late childhood (Solomon 2004; Begeer 2011). Studies with young children and preschoolers largely described participants as having ‘core’ autism, or ASD.

All studies reported some measure of general intellectual ability such as verbal mental age. Almost half of the included studies included a sample in the normal intellectual range (Bolte 2002; Solomon 2004; Golan 2006; Kim 2009; Golan 2010; Ryan 2010; Begeer 2011; Young 2012; Baghdadi 2013) and the rest reported on a sample with intellectual disability. One study split the participant group into those with and without associated intellectual delay (Hopkins 2011).

Sample sizes varied widely from n = 10 (Bolte 2002; Kim 2009) to n = 61 (Kaale 2012). On the whole, very small proportions of participants failed to complete the interventions. The maximum drop-out rate was 27% from a small sample (Goods 2013), but many studies reported no drop-out at all.

Intervention target types

The reported intervention types can be assigned to the following categories, taken from the review protocol.

1. Interventions that explicitly state that they are designed to teach ToM = (Hadwin 1996; Solomon 2004; Fisher 2005; Begeer 2011; Baghdadi 2013).

2. Interventions that explicitly state that they are designed to teach precursor skills of ToM = (Bolte 2002; Golan 2006; Kasari 2006; Kim 2009; Golan 2010; Kasari 2010; Landa 2011; Ryan 2010; Wong 2010; Hopkins 2011; Ingersoll 2012; Kaale 2012; Williams 2012; Young 2012; Goods 2013; Schertz 2013; Wong 2013).

3. Interventions that explicitly state that they are based on or inspired by ToM models of autism.

4. Interventions that explicitly state that they are designed to test the ToM model of autism.

There were no studies falling into category three or four and the vast majority of studies stated that they were designed to teach precursor skills of ToM. Within this category we could also identify some common intervention targets including the following.

- Emotion recognition (Bolte 2002; Golan 2006; Golan 2010; Ryan 2010; Hopkins 2011; Williams 2012; Young 2012)
- Joint attention and social communication (Kasari 2006; Kim 2009; Kasari 2010; Landa 2011; Wong 2010; Kaale 2012; Goods 2013; Schertz 2013; Wong 2013)
- Imitation skills (Ingersoll 2012)
Delivery methods, durations and dose

Three studies reported on the use of a computer program to deliver the intervention (Bolte 2002; Golan 2006; Hopkins 2011) and all of these studies had emotion recognition as the target skill. Three studies investigated the effect of a set of specially-designed cartoons on emotion recognition (Golan 2010; Williams 2012; Young 2012). Other studies investigated the effects of one-to-one therapist-led interventions (Hadwin 1996; Fisher 2005; Kasari 2006; Landa 2011; Ryan 2010; Wong 2010; Ingersoll 2012; Goods 2013) and two of these used the same manualised treatment program (Kasari 2006; Kaale 2012). Some used a therapist-led approach in a group treatment setting (Solomon 2004; Begeer 2011; Baghdadi 2013) and one was a group music therapy approach (Kim 2009). Non-expert intervention delivery was rare with only four studies reporting a parent-training element (Solomon 2004; Kasari 2010; Begeer 2011; Schertz 2013) and one study reporting on teacher-training for intervention delivery in the classroom (Wong 2013).

Intervention durations varied widely from two or three weeks (Hadwin 1996; Young 2012) to six months (Landa 2011). Dose was more consistent, with most falling within a range of 30 minutes per week (Kim 2009) to 3.5 hours per week (Hadwin 1996; Kasari 2006; Golan 2010), and one outlying intervention which reported therapist contact time of 2.5 hours per day (Landa 2011).

Most studies had wait-list or treatment-as-usual control conditions. Six studies (Kim 2009; Landa 2011; Hopkins 2011; Williams 2012; Young 2012; Baghdadi 2013) included control conditions, which were not expected to have an impact on intervention outcome but were included as a contact control only. These included toy play, non-synchronous one-to-one time, using art software, group leisure activities, and watching a Thomas the Tank Engine DVD.

Outcome measures

On the whole, studies rarely identified a single primary outcome measure. Those that organised outcomes into primary and secondary categories usually had multiple measures in each category. The outcome measures used most commonly included the following.

- Recognition of emotion from a variety of stimuli, including static images of faces, static images of the eyes, film clips, short stories, and cartoons (Bolte 2002; Solomon 2004; Golan 2006; Golan 2010; Ryan 2010; Hopkins 2011; Williams 2012; Young 2012; Baghdadi 2013)
- Direct assessment of ToM abilities (Hadwin 1996; Solomon 2004; Fisher 2005; Begeer 2011; Williams 2012)
- Imitation skills (Landa 2011; Ingersoll 2012)
- Diagnostic outcome (Wong 2010, Young 2012)

The studies below included the following additional outcome measures.

- Caregiver measures such as quality of involvement, adherence to treatment, mental health or satisfaction surveys (Solomon 2004; Kasari 2010; Landa 2011, Wong 2010; Baghdadi 2013; Wong 2013)
- General social skills measures, including rating scales and observation (Kim 2009; Begeer 2011; Hopkins 2011; Ingersoll 2012; Williams 2012)
- Symbolic play measures (Hadwin 1996; Wong 2010) or assessments of play variety (Goods 2013; Wong 2013)
- Language (Kasari 2006; Landa 2011) and conversational skills (Hadwin 1996)
- fMRI (functional magnetic resonance imaging - assessment of brain activity in facial recognition areas) (Bolte 2002)
- Adaptive function (Kasari 2006; Schertz 2013) and general intellectual abilities (Landa 2011; Schertz 2013)

Selection for meta-analyses

Using protocol criteria, three groups of studies were identified as eligible for meta-analysis.

1. Emotion recognition studies, with a treatment-as-usual control, and outcome measures using judgements of emotional expressions from static photographs of faces (Analysis 1.2).
2. Joint attention and social communication studies, with a treatment-as-usual control, and outcome measures using coding of parent-child interaction videos (Analysis 1.1).
3. Joint attention studies, with a treatment-as-usual control, and outcome measures of joint attention initiation frequency within a standardised assessment (the Early Social Communication Scales) (Analysis 1.3).

Excluded studies

Examination of the abstracts and, where necessary, full texts of reports resulted in a number of exclusions, listed in Characteristics of excluded studies for the following reasons.

- Not fitting the ToM-linked criteria for inclusion (23 reports)
- Not presenting any new data (10 reports)
- Not randomised or quasi-randomised controlled trials (18 reports)
- Not reporting on a sample of people with ASD (one report)
- Reporting on a broad-based intervention without a specific ToM-linked focus (12 reports)
- Diagnosis of participants invalid (one report)
- Reporting on an experimental pilot RCT with a very short intervention period (three reports)
Risk of bias in included studies

Further information was requested from the majority of authors as papers were not always complete in their reporting. The summaries of 'Risk of bias' judgements are shown in Figure 2 and Figure 3.

Figure 2. 'Risk of bias' graph: review authors' judgements about each risk of bias item presented as percentages across all included studies.
Figure 3. 'Risk of bias' summary: review authors’ judgements about each risk of bias item for each included study.

<table>
<thead>
<tr>
<th>Study</th>
<th>Random sequence generation (selection bias)</th>
<th>Allocation concealment (selection bias)</th>
<th>Blinding of review assessment (performance bias)</th>
<th>Blinding of outcome assessment (detection bias)</th>
<th>Incomplete outcome data (attrition bias)</th>
<th>Selective reporting (reporting bias)</th>
<th>Other bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baghdadli 2013</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bigger 2011</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Butler 2002</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher 2005</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olian 2006</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olian 2010</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goods 2013</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hadwin 1995</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hopkins 2011</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingersoll 2012</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaale 2012</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kasari 2009</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kasari 2010</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim 2009</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landa 2011</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ryan 2010</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schertz 2013</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solomon 2004</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Williams 2012</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong 2010</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong 2013</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young 2012</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Allocation

All studies in this review described themselves as randomised controlled trials and were included on this basis. One was later revealed to have used a non-random allocation procedure (Young 2012). This study states that participants were “randomly allocated to two groups” (Young 2012, p. 986) but email correspondence to clarify the exact allocation procedure revealed that in fact this study used alternate allocation by study enrolment. In other studies, a similar statement is made but rarely is full detail provided. Thus, only half of the included studies (11 reports) were judged to have ‘low’ risk of bias in terms of the description of the method of randomisation. Only six were judged to have adequately described allocation concealment. Therefore, the majority of included studies have either ‘unclear’ or ‘high’ risk of bias in this category.

In some cases, efforts were made to conceal allocation, for example, using randomisation within blocks to ensure random allocation and smooth delivery of the intervention (Goods 2013). However, the use of blocks of fixed length meant that the final case within each block would be allocated to a known condition.

Blinding

The majority of studies were judged at high risk of bias in this category (19/22 studies, 86%). The three exceptions are Landa 2011 and Hopkins 2011 where partial blinding was achieved, and Young 2012 who created a study design with full blinding as the intervention was delivered not by a therapist but on a DVD. Blinding of participants and personnel was rarely possible in the studies included in this review, as behavioural interventions were being used and these were often therapist-led. Blinding of outcome assessors is easier to achieve and 14 studies (64%) clearly reported blinding at this stage, though in a further five cases it was unclear whether this was completed adequately.

Though risk of bias must be judged as high when blinding is not achieved, a number of mitigating factors might help to reduce the impact of this risk.

1. When working with very young children or those severely affected by autism and/or intellectual disabilities, it is reasonable to judge that participants are relatively oblivious to the intervention content and certainly to the expected outcomes.

2. Likewise, although participants and parents may be aware of their group they may not be apprised of the hypotheses of the study. For example, Golan 2006 worked with able adults with autism who were asked to “help in the evaluation of a piece of new software” (p. 600) rather than being told the software was designed to help them learn to understand emotions.

3. Many studies used automated outcome measures, especially when using a computerised intervention (e.g. Bolte 2002; Golan 2006), which are more resilient to bias than experimenter-led methods.

4. Studies using multiple outcome measures often achieved blinding for a sub-set of those outcomes (e.g. Hopkins 2011).

Incomplete outcome data

There was very little evidence of attrition among the studies reported here, and only three of the included studies were judged to be at high risk of bias. The most extreme case of likely bias was (Goods 2013) who reported 73% retention and analysed outcome data for intervention completers only. Ten studies (45%) reported outcome data for all of the original sample and where there was participant drop-out this was usually described with clear reasons to help the reader judge the impact of this drop-out. Where studies were judged at unclear risk of bias, this was due to either a lack of sufficient detail in the published report (e.g. Begeer 2011) or because it was difficult to evaluate the impact of the drop-out on findings (e.g. Kim 2009; Kasari 2006).

Selective reporting

Selective reporting was not evident among the papers included and 18 studies (82%) were judged at low risk of bias in this category. However, it must be noted that the tendency not to identify a primary outcome measure and to use multiple outcomes does hinder conclusions about intervention efficacy.

One study (Bolte 2002) did not report means for a relevant outcome measure - the International Affective Pictures System or IAPS - for which there was no significant group difference. The study authors were contacted to provide mean scores but were unable to provide these due to the time elapsed since the study and records not having been kept. One further study (Ryan 2010) reported data in graphical form only, but the authors kindly provided accurate means and SDs for these data. Conversely, Goods 2013 reported non-significant findings in a table of results but these were not discussed in the text.

Other potential sources of bias

One study reported a significant difference in ratings provided by mothers and ratings provided by independent examiners (Kim 2009) with professional ratings providing a more positive estimate of intervention efficacy. This bias is judged to be of low risk for two reasons. First, the professional ratings used to construct the primary outcome for the intervention were blind to group. Second, the authors provided a reasonable justification for the under-estimation of intervention efficacy by parent ratings, which is that mothers over-estimate pre-intervention abilities of their children, and thus under-estimate efficacy of the intervention.
Another study combined the wait-list control group with the intervention group to provide a larger sample size for analysis of some intervention efficacy measures (Wong 2010). However, prior to this stage in the analysis, between group comparisons were also made and these provided the primary outcomes for the study. The impact of this bias is judged to be unclear as while a between group comparison was made, the long-term maintenance of intervention gains is disguised by the combination of data sets at the final time point.

Schertz 2013 adopted a variable intervention period which could have weighted findings towards a positive conclusion regarding intervention efficacy. Their strategy was to recruit participants who were demonstrably lacking the target joint attention skills. Participants were then paired and randomly assigned to intervention or treatment-as-usual control groups. Intervention then proceeded until an individual participating child had achieved the target skill. At this point, exit assessment measures were taken for that child and their matched pair. Therefore, within each pair, participants experienced the same interval between baseline and post-intervention assessment. However, when analysed as a group, this system ensured that every child in the intervention group had shown significant gains in the target skill, thus biasing the study towards a positive conclusion.

No other potential sources of bias were identified in the studies selected for inclusion in this review. Also, the authors note some examples of particularly good practice in the prevention of bias, including close measurement of treatment adherence in therapist-led (Baghdadli 2013) and parent-training studies (Schertz 2013).

**Effects of interventions**

See: Summary of findings for the main comparison

Studies in this review used a wide variety of outcome measures, often using measures specific to their intervention target and sometimes designed specifically for that study. In addition, effect sizes, mean differences reported as standardised mean differences (SMD) and confidence intervals (CI) were not always reported, though we include these below where available. Intervention measures are listed in additional Table 2 and discussed below, organised by primary and secondary outcome category. In addition, primary outcome results are collated in the Summary of findings for the main comparison.

**Evaluating primary outcomes: communication**

Primary outcome measures in this section were those using standardised assessments to assess social skills of diagnostic relevance. Six studies used the Early Social Communication Scales (ESCS) to evaluate the outcome, standardised observational assessment of social engagement behaviours, including joint attention (Kasari 2006; Kim 2009; Ingersoll 2012; Kaale 2012; Goods 2013; Wong 2013). Due to differences in measurement (e.g. reporting behaviour frequency versus amount of time; reporting sub-items from the scales versus reporting scale totals) and study design (those with or without a therapeutic control condition), not all studies could be combined for meta-analysis. However, three studies were combined in this way (Analysis 1.3) and the outcome indicated no significant intervention effect on social behaviour measured in this way (SMD 0.23, 95% CI -0.48 to 0.94, Z = 0.63, P value = 0.53, three studies, 92 participants). There was no evidence of heterogeneity in effects ($I^2 = 57\%$, $\text{Chi}^2 (df = 2) = 4.66$, $P$ value = 0.10, $\tau^2 = 0.22$) (Figure 4). Among studies not included in meta-analysis, Kasari 2006 found large treatment gains in the showing (effect size = 1.50) and responding to joint attention (effect size = 1.20) items of the ESCS but not in other relevant items such as pointing and giving. This may indicate a lack of generalisation of skills beyond the specific taught items.

Wong 2010 used a sub-set of ADOS (Lord 2000) items to evaluate communication gains in response to intervention, finding improvements in relevant items (vocalisation directed to others, gestures, pointing) in the intervention group (median difference = 4 points), but not in the control group (median difference = 2.5 points). This finding is weakened by the fact that these analyses compared change from baseline to outcome in each group separately and there was no between-group comparison. Furthermore, the ADOS is not intended as an intervention outcome measure, and it is not usual to analyse a sub-set of items. On the other hand, this finding is strengthened by a comparison that shows no intervention group gains in items pre-identified as non-relevant to the intervention. Young 2012 similarly reported change for individual items of the Social Communication Questionnaire (SCQ) (Rutter 2003). Of specific relevance in this outcome category, they analysed change in eye contact and gaze aversion, and found no intervention effects on these items (effect sizes: $\eta_p^2 = 0.001$ and $\eta_p^2 = 0.002$ respectively).

For participants at a higher level of communicative sophistication, Hadwin 1996 (reported in Hadwin 1997) evaluated the impact of ToM intervention on complex language skills. They found no effect of intervention on conversational skills, and also raised additional evidence that language level may moderate intervention effects when teaching ToM skills (Hadwin 1996) though this is not explored elsewhere.

**Evaluating primary outcomes: social function**

Primary outcome measures in this section were those using standardised assessments to assess social skills of diagnostic relevance. Six studies used the Early Social Communication Scales (ESCS) to evaluate the outcome, standardised observational assessment of social engagement behaviours, including joint attention (Kasari 2006; Kim 2009; Ingersoll 2012; Kaale 2012; Goods 2013; Wong 2013). Due to differences in measurement (e.g. reporting behaviour frequency versus amount of time; reporting sub-items from the scales versus reporting scale totals) and study design (those with or without a therapeutic control condition), not all studies could be combined for meta-analysis. However, three studies were combined in this way (Analysis 1.3) and the outcome indicated no significant intervention effect on social behaviour measured in this way (SMD 0.23, 95% CI -0.48 to 0.94, Z = 0.63, P value = 0.53, three studies, 92 participants). There was no evidence of heterogeneity in effects ($I^2 = 57\%$, $\text{Chi}^2 (df = 2) = 4.66$, $P$ value = 0.10, $\tau^2 = 0.22$) (Figure 4). Among studies not included in meta-analysis, Kasari 2006 found large treatment gains in the showing (effect size = 1.50) and responding to joint attention (effect size = 1.20) items of the ESCS but not in other relevant items such as pointing and giving. This may indicate a lack of generalisation of skills beyond the specific taught items.
As described above, two studies used scores on a diagnostic or screening measure as their primary outcome, principally focusing on change in social function. Wong 2010 reported item-by-item changes rather than algorithm totals for the ADOS (Lord 2000), and Young 2012 analysed intervention effects on a single ‘Social Peer Interest’ item of the SCQ (Rutter 2003). Young 2012 did not find significant effects on this SCQ item ($\eta^2 = 0.06$), despite the intervention having a positive impact on emotion recognition skill; evidence of a lack of generalisation or expansion of a taught skill to related domains. In contrast, Wong 2010 did report intervention effects on items from the ADOS, which are relevant to the intervention content (intervention group median difference = 7 points; control group median difference = 2.5 points), and crucially not for items unrelated to the intervention. However, it should once again be noted in both studies that the measures used were not designed to be analysed on an item-by-item basis. Moreover, Wong 2010 reported no direct comparison of the degree of change between the two groups.

Williams 2012 additionally reported on social abilities measured in a real-world context using standardised measures of life skills (intervention group mean difference = 2.13 points; control group mean difference = 1.59 points). No evidence was found of intervention effects measured in this way, once more indicating a lack of generalisation from specifically-taught skills to wider social abilities.

Further detail on outcomes from observed measures of social engagement is provided below under the heading of Secondary Outcomes: Change in participant behaviour or quality of interpersonal interaction, or both, measured by direct observation.

### Evaluating secondary outcomes

#### Intervention Specific: Change in targeted cognitive skill, Emotion recognition

Seven studies examined the impact of intervention on facial affect recognition skills from photographs, as compared to treatment-as-usual (Solomon 2004; Golan 2006; Ryan 2010; Hopkins 2011; Williams 2012; Young 2012; Baghdadi 2013), which were inspected further for potential inclusion in a meta-analysis. Young 2012 was excluded from the meta-analysis due to use of a non-random allocation procedure. Baghdadi 2013 was eligible for inclusion in a meta-analysis in principle, but it was not possible to extract the relevant data from the paper, which reported median scores and inter-quartile range for it's non-normally distributed data, rather than means and standard deviations. The study found a significant improvement in the intervention group (but not the control group) in recognition of angry emotions only (effect size, Cohen’s $d = -0.8$, $P$ value $= 0.05$), but no significant differences between groups for other emotions. It was also inappropriate to incorporate Williams 2012 into the meta-analysis as this study used a control condition, which was hypothesised to have potential treatment effects.

Hopkins 2011 reported on two separate samples: children with low-functioning autism (LFA) and children with high-functioning autism (HFA). All other studies in this group reported on participant samples with IQ or language ability in the normal range, and therefore it seemed most appropriate to include the HFA sample from Hopkins 2011 in the meta-analysis. Likewise, Solomon 2004 reported separately on younger and older groups of children. The majority of other studies in this group reported on child participants whose age more closely matches the older group of Solomon et al (Ryan 2010; Hopkins 2011), and one remaining study involved an adult sample (Golan 2006). Therefore, it was decided to include the older participant group in the meta-analysis.

All studies in the emotion recognition group reported significant group differences on outcome immediately post-treatment, measured by recognition of facial emotion from static images. The meta-analysis (Analysis 1.2) shows evidence of a positive intervention effect on emotion recognition (SMD 0.75, 95% CI 0.22 to 1.29, $Z = 2.75$, $P$ value $= 0.006$, four studies, 105 participants). There was no evidence of heterogeneity in effects ($I^2 = 36\%$, Chi $^2 (df = 3) = 4.70$, $P$ value $= 0.19$, $\tau^2 = 0.11$) (Figure 5). A study excluded for reasons of bias (Young 2012) also found a positive effect of intervention on emotion recognition skills.

Figure 4. Forest plot of comparison: 1 test, outcome: 1.3 Joint attention initiations in standardised assessment.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Treatment Mean</th>
<th>SD</th>
<th>Total</th>
<th>Control Mean</th>
<th>SD</th>
<th>Total</th>
<th>Std. Mean Difference IV, Random, 95% CI</th>
<th>Std. Mean Difference IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goods 2013</td>
<td>0.4</td>
<td>0.89</td>
<td>5</td>
<td>1</td>
<td>1.73</td>
<td>6</td>
<td>-0.39 [-1.59, 0.82]</td>
<td></td>
</tr>
<tr>
<td>Ingersoll 2012</td>
<td>4.79</td>
<td>5.41</td>
<td>14</td>
<td>0.92</td>
<td>1.44</td>
<td>13</td>
<td>33.7% [0.13, 1.73]</td>
<td></td>
</tr>
<tr>
<td>Kooi 2012</td>
<td>1.6</td>
<td>2.6</td>
<td>27</td>
<td>1.6</td>
<td>2.2</td>
<td>27</td>
<td>44.3% [0.00, 0.53]</td>
<td></td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>46</td>
<td></td>
<td></td>
<td>46</td>
<td></td>
<td></td>
<td>100.0% [0.16, 0.94]</td>
<td></td>
</tr>
</tbody>
</table>

Heterogeneity: $\tau^2 = 0.22$, $\chi^2 = 4.88$, $df = 2 (P = 0.04)$, $I^2 = 57\%$

Test for overall effect: $Z = 0.63$ ($P = 0.53$)
Figure 5. Forest plot of comparison: 1 test, outcome: 1.2 Emotion recognition from face photographs, TAU control.

We note that Bolte 2002 reported a null finding in a related study using emotion recognition training, with recognition of facial emotion from static images as an outcome. However, it was not possible to retrieve these data. This study had a small sample size (n = five per group) and therefore it is unlikely that this null finding would have a large effect on the reported meta-analysis. In addition to the results included in our meta-analyses, many interventions targeting emotion recognition skills were also evaluated by additional measures tapping the same or related constructs. For example, emotion recognition was assessed using stimuli, including audio clips (Golan 2006), film clips (Golan 2006), emotional vocabulary tests (Golan 2010), and matching emotional vignettes to facial expressions (Golan 2010). A consistent finding in these studies using multiple measures of the same construct was an absence of generalisation of the target skill to novel settings or stimuli where this was assessed. For example, Golan 2006 evaluated emotion recognition and found improvements only on ‘close-generalisation’ tasks, which did not extend to other outcome measures (e.g. reading-the-mind-in-the-eyes: intervention group mean difference = + 0.7 points; control group mean difference = -0.9 points).

Intervention Specific: Change in targeted cognitive skill, Theory of Mind (ToM)

Four studies, all aiming to directly teach ToM, used explicit assessment of ToM as their primary outcome (Hadwin 1996; Solomon 2004; Fisher 2005; Begeer 2011). These outcome measures overlap with those described in the Emotion Recognition section above as they sometimes use emotional content (especially complex emotions and mental states). However, they are evaluated independently here as the source authors clearly identify them as assessments of ToM. Once again, positive outcomes were found when assessing progress within, or close to, the taught context, but there was an absence of generalisation of taught skills to novel, or more complex scenarios or to abilities hypothetically built on ToM. For example, Begeer 2011 reported a significant difference between groups in degree of improvement measured by the ToM test (effect size, Cohen’s d = 0.75), but no such effect in measures of self-reported empathy or parent-reported social skills. Williams 2012 reported on an emotion skills intervention but additionally assessed ToM abilities in order to explore the extent of skills learnt. Once more, there were no intervention effects on this extended skill set.

Intervention Specific: Change in targeted cognitive skill, Imitation

Two studies reported gains in imitation skills as their primary outcome in intervention studies with toddlers (Landa 2011; Ingersoll 2012). For example, Ingersoll 2012 (reported in Ingersoll 2010) showed that the treatment group made larger gains in imitation than the control group, though this finding had a small to moderate effect size (elicited imitation, $\eta_p^2 = 0.20$; spontaneous imitation, $\eta_p^2 = 0.29$; object imitation $\eta_p^2 = 0.21$; gesture imitation $\eta_p^2 = 0.38$). Unfortunately, due to differences in measurement it was not possible to combine these two studies in meta-analysis. Their combined sample size is just 69 participants.

Intervention Specific: Change in targeted cognitive skill, Play

Finally, four studies included assessment of play as a secondary outcome (Hadwin 1996; Wong 2010; Goods 2013; Wong 2013). Hadwin 1996 found no effect of teaching ToM understanding on observed symbolic play skills, while the findings of Wong 2010, using an observational symbolic play test are positive, though the larger sample size is reported by Hadwin 1996. Regarding the range of play types observed, Goods 2013 reported positive intervention effects on this variable using a ‘Structured Play Assessment’ while Wong 2013 reported no positive effects on the same measure. These two studies each reported a different output from the ‘Structured Play Assessment’ (play types versus play level) making it illogical to combine these data in a meta-analysis.

Change in participant behaviour or quality of interpersonal interaction, or both, measured by direct observation
A series of studies used joint engagement behaviours during adult-child interactions as an outcome measure. However, subtle but important differences in the operationalisation of these outcomes meant that very few were eligible for combined analysis. For example, some studies reported on parent-child (Kasari 2006; Kasari 2010) and some on teacher-child interactions (Kim 2009; Kaale 2012; Goods 2013). Some studies reported on only one category of joint attention behaviour as a single outcome (Kaale 2012), while some sub-divided these into initiations and responses (Kasari 2006; Kasari 2010; Landa 2011). In addition, there were differences in the outcome unit of measurement with Kasari 2006 reporting total seconds of joint attention behaviours and Kasari 2010 reporting frequency of joint attention behaviours. Finally, some studies reported on behaviour measured in a one-to-one setting (Kaale 2012) while others reported on spontaneous behaviour within a group setting, e.g. the classroom (Wong 2013) or during school recess (Hopkins 2011).

The only outcomes of sufficient similarity to be appropriate for inclusion in meta-analysis were the measures of joint engagement during mother-child play reported in Kasari 2010 and Kaale 2012, both as percentages of total time. In meta-analysis (Analysis 1.1) there was evidence of a positive intervention effect on joint engagement (SMD 0.55, 95% CI 0.11 to 0.99, Z = 2.45, P value = 0.01, two studies, 88 participants). There was no evidence of heterogeneity in effects (I\(^2\) = 5%, Chi\(^2\) (df = 1) = 1.05, P value = 0.30, Tau\(^2\) = 0.01) (Figure 6).

**Figure 6. Forest plot of comparison: 1 test, outcome: 1.1 Joint engagement in mother-child interaction.**

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental Mean</th>
<th>SD</th>
<th>Control Mean</th>
<th>SD</th>
<th>Weight</th>
<th>Std. Mean Difference IV, Random, 95% CI</th>
<th>Std. Mean Difference IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaale 2012</td>
<td>67.3</td>
<td>22.8</td>
<td>49.2</td>
<td>19.9</td>
<td>27</td>
<td>0.37 [0.17, 0.99]</td>
<td></td>
</tr>
<tr>
<td>Kasari 2010</td>
<td>42.85</td>
<td>18.98</td>
<td>19.4</td>
<td>13.41</td>
<td>18</td>
<td>0.84 [0.14, 1.53]</td>
<td></td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>45</td>
<td>43</td>
<td>100.0%</td>
<td></td>
<td></td>
<td>0.55 [0.11, 0.99]</td>
<td></td>
</tr>
<tr>
<td>Heterogeneity</td>
<td>Tau(^2) = 0.01; Chi(^2) = 1.05, df = 1 (P = 0.30); P = 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect</td>
<td>Z = 2.45 (P = 0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here, there was a little evidence that taught skills may generalise to new settings. Kasari 2006 and Kaale 2012 both found that therapist-taught social communication skills were in evidence in a parent-child interaction scenario and when using novel assessment materials (Gulsrud 2007). For example, Kaale 2012 found that children in the intervention group spent on average 12.2% more time in a joint engagement state with their mother compared with children in the control group (95% CI = 2.4% to 22%, effect size Cohen’s d = 0.67).

**Change in participant behaviour or skills measured by adult report**

Other studies reported an array of general social skills measures (Solomon 2004; Fisher 2005; Kim 2009; Wong 2010; Begeer 2011; Hopkins 2011; Ingersoll 2012). These were very different in method of assessment, construct being examined, and data format, and so it is not possible to make a direct comparison between studies. Nevertheless, the overall message from the study authors is of improvement in social skills as a result of intervention.

Three studies reported measures of language and general cognitive or adaptive ability as an outcome (Kasari 2006 reported in Kasari 2008, Landa 2011; Schertz 2013). The studies produced conflicting results. Landa 2011 did not find a significant difference between intervention and control groups on a measure of expressive language despite a moderate effect size of 0.49, while Kasari 2008 and Schertz 2013 reported significant gains in expressive language in the intervention compared with the control group. However, when evaluating participants over the long term (Kasari 2012a), there was no evidence that treatment continued to impact on language and cognitive outcome five years from baseline.

**Acceptability of Intervention**

A range of studies incorporated caregiver measures such as quality of involvement, adherence to treatment, mental health or participant well-being and satisfaction surveys (Solomon 2004; Kasari 2010; Wong 2010; Baghdadli 2013; Wong 2013), but these were very different in method of assessment and specific construct targeted, and so it is not possible to draw solid conclusions from these data.

**Rate of drop-out**

Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)
As discussed previously (Assessment of risk of bias in included studies), rates of drop-out in the included studies were very low with half of the included studies reporting that the full sample remained enrolled for the length of the trial. One study reported a very high rate of drop-out from a small sample in a study taking place within a pre-school setting and it was unclear why this drop-out rate was so high (Goods 2013). Otherwise, the highest rates of drop-out occurred when the intervention was self-directed by an individual with ASD (Golan 2006) or when the intervention took place at a clinic which required significant travelling time for participants (Kim 2009).

Economic data
No studies reported economic data.

Follow-up effects
A minority of studies followed up participants after the immediate intervention period had ended (Hadwin 1996; Fisher 2005; Kasari 2006; Kasari 2010; Landa 2011; Ryan 2010). Follow-up periods ranged from six weeks to five years and findings consistently showed maintenance of treatment gains in a range of measured skills, including ToM (Hadwin 1996; Fisher 2005), social engagement (Kasari 2006; Kasari 2010; Landa 2011), and emotion recognition (Ryan 2010). In some cases there was evidence of continued significant growth during the post-intervention period (Fisher 2005; Kasari 2006; Kasari 2010). The sole exception is Kasari 2006 (reported in Kasari 2012a) who found no long-term impact of intervention on language and cognitive outcome at five years post-baseline.

Discussion
Summary of main results
Twenty-two randomised trials were included in this review. All reported on interventions that targeted either Theory of Mind (ToM), or one of the accepted sub-components of ToM such as shared attention, emotion recognition or imitation. Most involved either wait-list or treatment-as-usual control conditions, or a control condition devised to replicate contact time but without therapeutic content.
Risk of bias in the included studies was variable. Very few studies provided adequate information and in a majority of cases the authors were contacted to provide further details. In particular, details of sequence generation and allocation concealment were lacking in the published articles. Blinding of participants and personnel was judged to be impossible in almost every study reviewed, and as a result risk of bias in this category was high. Blinding of outcome assessors is better achieved and reported on, and concerns about selective reporting or attrition were rare.
In addition, outcomes varied widely, both in the construct being measured and the means of measurement. This made it a challenge to combine studies for meta-analysis and as a result only three, very small meta-analyses were conducted. Outcomes also differed significantly from those predicted in the review protocol. There were very few outcomes reported that fell into the protocol-defined categories of symptom level for the two diagnostic domains of social and communication impairments. Instead, studies mostly reported on intervention-specific measures of constructs such as emotion recognition, joint attention and social communication, and imitation. In these three intervention target areas (emotion recognition, joint attention and social communication, and imitation), there was modest evidence of intervention success.

Emotion recognition
Studies in the emotion recognition field consistently found positive effects of specific training, a conclusion reinforced by combined analysis of four studies (Figure 5). However these skills did not always generalise to testing contexts that moved beyond the trained content. Since the overall goal of emotion recognition training for people with ASD must be to improve real world emotion recognition and, by extension, emotional understanding, this is a significant limitation, which calls into question the value of emotion recognition interventions. These studies are additionally limited by their reliance on both teaching and measuring facial affect recognition as a proxy for emotion recognition. Real world emotion recognition skills require appraisal of emotional valence from a variety of sources (e.g. facial emotion, body language, content and tone of speech, context) and these skills are rarely taught.

Joint attention and social communication
A number of included studies reported on therapist-led, one-to-one or group interventions, normally with young children with autism, targeting a core deficit of joint attention and social communication. These studies often involved a high level of intervention contact hours but could produce sustained effects (Kasari 2010), although the intervention effect may not be maintained in the longer term, i.e. over a period of years (Kasari 2006). The studies in this category have significant real world relevance as there is evidence that therapist training can lead to improvements in interactions with other familiar adults as well (Kasari 2006; Kaale 2012). Since joint attention is a theoretical prerequisite for both language learning and ToM development, more longitudinal studies looking at the effect of joint attention improvement on these linked skills would be of value.

Imitation
One study specifically targeted imitation both as intervention target and method of delivery (Ingersoll 2012), while another used a more general social communicative approach which impacted on imitation skills (Landa 2011). These studies, in common with those described above, require further investigation to explore whether training on this theoretical building block of ToM can subsequently enhance ToM development.

**Theory of Mind**

Perhaps surprisingly, a minority of included studies focused explicitly on training ToM skills. As for emotion recognition, it was found relatively straightforward to teach a core ToM skill and to find a difference in that skill when tested under the same circumstances, but generalisation across props, settings, and partners was a greater challenge.

**General conclusions**

A common theme, therefore, in these studies was the difficulty of creating positive intervention effects recognisable in everyday function. The skills targeted in the interventions included in this review are all developmentally linked, both to each other and to more general abilities such as language and friendship skills. However, it is rare for an intervention to succeed in triggering this developmental chain. For this reason it may be that the broad-based interventions, excluded here because they cannot be clearly categorised according to the protocol criteria, represent a more successful approach to intervention for autism. They may help to promote generalisation by working in a more ecologically valid way to support the development of new skills in combination.

One complicating factor, which hinders the opportunity to draw strong conclusions from the review, is that the studies lack appropriate outcome measures in two different ways. First, there is no consistent, widely-used measure for assessing intervention outcome using diagnosis-relevant domains for ASD research - not even for a specific age-range or ability level. As a result, studies often rely on bespoke measures such as video-coding of parent-child interactions, or multiple measures assessing a range of social, intellectual, and adaptive skills. This is an obstacle to comparison of multiple studies both at a systematic review level and for the average reader of these published reports. Second, the use of multiple measures, combined with an absence of explicit identification of a single primary outcome measure, means that it is often possible to claim intervention success on a sub-set of the measures used, but not all. Without a priori identification of the primary outcome measure it is difficult to judge how these mixed findings should be interpreted.

**Overall completeness and applicability of evidence**

The studies reported in this review cover a good range of ToM and associated skills, though the fact of this array means that more evidence is needed in each category. It was not possible to conduct subgroup analyses for this review, but it is theoretically likely that we would find differences between specific diagnostic categories, age groups, and intervention delivery methods. The evidence, though limited and of generally low quality, suggests that it may be possible to teach ToM-linked abilities.

Research now should explore how different approaches might be applied to different populations, requiring a much larger and more systematic body of evidence. In particular, it is understood that different age and ability levels require support developing different sub-skills and mapping out these relationships would be of value. In the field of emotion recognition, however, this skill has been successfully taught to preschoolers, children and adults, quite often using computer-based methods, and an understanding of why this is a difficulty which persists across the life span would be of value.

**Quality of the evidence**

The evidence included here varies in quality due to the aforementioned difficulty of blinding participants and personnel, and under-reporting of other relevant practices such as sequence generation and allocation concealment. In addition, sample sizes are often very low, with the maximum reported at only n = 61 (Kaale 2012). Due to a lack of strong reporting norms it has not always been possible to accurately judge risk of bias in the studies reported here, which may have unknown impact on conclusions. Further, while attrition rates are low on the whole, all studies analyse only the final sample and do not use an intention-to-treat analysis. Once more it is impossible to judge to what degree this may have affected results, and in what direction. Overall, there is a lack of good quality evidence in this field and a requirement for more randomised controlled trials representing highest standards in methodology, particularly outcome assessment.

One group of studies represented a good example of independent evaluation of the same intervention, *The Transporters* DVD for teaching emotion recognition and understanding (Golan 2010; Williams 2012; Young 2012). Unfortunately, minor but important differences in study design between these reports meant that they could not be combined in a meta-analysis. However this rare example of replication, perhaps because the DVD-based intervention is relatively easy to apply, is a notable and positive exception in a field of diverse intervention strategies. We found no evidence of adverse effects but this partly reflects a lack of attention to these in the source literature. There seems to be a common presumption that behavioural interventions, such as those reviewed here, do not carry a risk of harm but this possibility needs to be addressed explicitly in future studies.
Potential biases in the review process

The review authors were restricted by only evaluating studies available in English language. In addition, some databases searched at the first data collection point were no longer available to us when searches were repeated two years later. It is unclear how these restrictions may have affected the conclusions drawn. Dissertation databases were not searched independently and instead the review authors requested unpublished data, including dissertations, from key authors. Two relevant PhD theses were examined (Arora 2008; Rodgers 2012) one of which is an excluded study (Rodgers 2012). We note that it is extremely rare in this field for postgraduate dissertations to conduct an independent clinical trial because of the challenges surrounding recruitment and design.

Agreements and disagreements with other studies or reviews

As this review appears to be the first to consider ToM interventions as a group of studies, no comparison can be made with the conclusions of other reviews.

AUTHORS’ CONCLUSIONS

Implications for practice

The review suggests it may be possible to teach both Theory of Mind (ToM) and the precursor skills associated with the construct. However, this teaching rarely or never generalises to novel contexts, and it is unclear whether there is long-term maintenance of learnt skills, or developmental progression in learning. Future reviewers should consider investigating the results of multi-modal interventions, which were excluded from this review, and which teach a range of real-world social and communicative skills as well as academic learning. Given the problems with maintenance and generalisation of taught skills, multi-modal approaches may be more successful in improving social and communication outcomes in people with ASD.

Implications for research

The ToM model of autism development proposes that ToM is a core deficit, which links both to precursor skills, such as joint attention and emotion recognition, and to subsequent abilities such as making friends and understanding metaphorical language. Theoretically then, this developmental model implies that a successful intervention at one point in the chain should have impact all the way up the developmental ladder.

More longitudinal studies are required to establish whether teaching a specific skill at one time point will lead not only to maintenance of that skill but also the development of further associated abilities subsequently. Studies systematically linking intervention targets and delivery methods to participant age and ability would support this process by helping to identify the right ‘starting place’ for an intervention for a particular individual. In order to reach larger populations, multi-site studies are required, with co-operation across research centres and national borders. This is challenging not just in a logistical sense but also because countries and even counties vary in the service provision and intervention approaches offered to people with ASD across ages.

There is an urgent need for better outcome measures for autism research across the board. This has been recognised in current funding from the National Institute for Health Research Health Technology Assessment programme for a systematic review of measurement properties of outcome measures for children with autism up to the age of six years (http://www.hta.ac.uk/project/2830.asp). For future studies in the field of ToM interventions, ideally outcome measures should capture symptom severity in core diagnostic domains, without looking for a change in diagnostic status — which is neither feasible nor necessarily to be desired. In the short term, however, the outcome measure issue could be partially resolved by better study reporting, in particular a priori identification of a single primary outcome measure to define intervention success. Additionally, secondary measures should have a clear rationale for inclusion and should be, wherever possible, linked to intervention components. Improvements in reporting could also be made, especially in providing details of sequence generation and allocation concealment, which is rarely described.

The Theory of Mind model of autism was first introduced in 1985 and in 2000 a review of the changes in the model was published (Baron-Cohen 2000). Now, almost another 15 years on, multiple versions (Astington 2011) and alternative explanatory models abound (e.g. Happe 2000; Mottron 2006; Baron-Cohen 2010) and there have been direct attacks both on the model (Hobson 1991) and on the methods used to measure the construct (Bloom 2000). One meaningful way to evaluate the explanatory power and clinical and educational relevance of these competing theories is to explore their impact in an intervention setting. Currently, however, the quality and quantity of evidence needed even for ToM alone is inadequate to do so.

acknowledgements

The authors are grateful for the support of Jo Abbott, Margaret Anderson, Chris Champion, Jane Dennis, Laura MacDonald, Geraldine MacDonald, and Inalgwu Oono for their help in developing this review and for the comments of Michelle Dawson on a previous draft.
References to studies included in this review

Baghdalii 2013 [published data only]

Begeer 2011 [published data only]

Bolte 2002 [published data only]

Fisher 2005 [published data only]

Golan 2006 [published data only]

Golan 2010 [published data only]

Goods 2013 [published data only]

Hadwin 1996 [published data only]

Hopkins 2011 [published data only]

Ingersoll 2012 [published and unpublished data]

Kaaie 2012 [published data only]

Kasari 2006 [published data only]

Kasari 2010 [published data only]

Kim 2009 [published data only]
Kim J, Wigram T, Gold C. The effects of improvisational music therapy on joint attention behaviours in autistic

References

Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)

**Landa 2011** *(published data only)*


**Ryan 2010** *(published and unpublished data)*


**Scherz 2013** *(published data only)*


**Solomon 2004** *(published data only)*


**Williams 2012** *(published data only)*


**Wong 2010** *(published data only)*


**Wong 2013** *(published data only)*


**Young 2012** *(published data only)*


**References to studies excluded from this review**

**Adams 2012** *(published data only)*


**Aldred 2004** *(published data only)*


**Beaumont 2008** *(published data only)*


**Cardon 2011** *(published data only)*


**Carter 2011** *(published data only)*


**Casenbiser 2011** *(published data only)*


**Castorina 2011** *(published data only)*


**Charman 2007** *(published data only)*

Charman T, Hudry K. Interventions targeting joint attention and symbolic play can improve aspects of these skills in young children with autism. *Evidence-Based Mental Health* 2007;10(1):12–21.

**Corbett 2011** *(published data only)*


**Dawson 2010** *(published data only)*


**DeRosier 2011** *(published data only)*


**Drew 2002** *(published data only)*


Estes 2011 [published data only]

Field 2001 [published data only]

Frankel 2010a [published data only]

Frankel 2010b [published data only]

Gantman 2012 [published data only]

Gattino 2011 [published data only]

Geretsegger 2012 [published data only]

Girolametto 1988 [published data only]

Green 2010 [published data only]

Hanley-Hochdorfer 2010 [published data only]

Heimann 2006 [published data only]

Hillier 2012 [published data only]

Howlin 1997 [published data only]

Howlin 2007 [published data only]

Jocelyn 1998 [published data only]

Jones 2010 [published data only]

Kasari 2012a [published data only]

Kern 2011 [published data only]

Koenig 2010 [published data only]

Lang 2010 [published data only]
Lang R. The Autism 1-2-3 Project, a short-duration and low-intensity intervention that targets basic social and communication skills and parent involvement, shows promise and is worthy of future research. Evidence-Based Communication Assessment and Intervention 2010;4(4): 196–9.
Laugeson 2012 [published data only]

Lawton 2012 [published data only]

LeGoff 2004 [published data only]

Lerner 2012 [published data only]

Lopata 2010 [published data only]

Mahoney 2005 [published data only]

McConachie 2004 [published data only]

McGregor 1998 [published data only]

Minshew 2010 [published data only]
Minshew N, Eack S. Cognitive enhancement therapy for adults with autism spectrum disorders. NIH Clinical Trials Register. [:: NCT00902798]

Morgan 2010 [published data only]

Nefdt 2010 [published data only]

Oosterling 2010 [published data only]

Ozonoff 1995 [published data only]

Pajareya 2011 [published data only]

Quirmbach 2009 [published data only]

Roberts 2011 [published data only]

Rodgers 2012 [unpublished data only]

Rogers 2010 [published data only]

Schertz 2007 [published data only]

Schmidt 2012 [published data only]
Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)

Silver 2001 [published data only]

Smith 2000 [published data only]

Smith 2004 [published data only]

Stichter 2001 [published data only]

Strain 2011 [published data only]

Swettenham 1996 [published data only]

Tanaka 2010 [published data only]

Temple 2007 [published data only]

Thomeer 2012 [published data only]

Turner-Brown 2008 [published data only]

Wellman 2002 [published data only]

Whalen 2003 [published data only]

White 2010 [published data only]

Wood 2009 [published data only]

Yoder 2006 [published data only]

Yoder 2006b [published data only]

References to studies awaiting assessment

Baghdadli 2010 [published data only]

Additional references

APA 1994

APA 2013

Arora 2008

Copyright © 2014 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.
Austingon 2011

Atladottir 2007

Baird 2006

Baron-Cohen 1985

Baron-Cohen 1995

Baron-Cohen 2000

Baron-Cohen 2010

Bloom 2000

Bowler 1992

Bowler 1997

Carruthers 1996

Charman 2000

Croen 2002

Elsabbagh 2012

Fombonne 2001

Frith 1994

Garfield 2001

Gulsrud 2007

Hadwin 1997

Happe 1994

Happe 1995

Happe 2006

Hartley 2009

Higgins 2008
Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)

Copyright © 2014 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

Hobson 1991

Ingersoll 2010

Joseph 2004

Kanner 1943

Kasari 2008

Kasari 2012b

Kim 2008

Klin 2005

Kogan 2009

Lefebvre 2008

Lord 1994

Lord 2000

Melzoff 1993

Mottron 2006

Mundy 1994

Nassar 2009

Ozonoff 1991a

Ozonoff 1991b

Parsons 1999

Ponnet 2004

Premack 1978

Ruffman 2001
Rutter 2003

Rutter 2005

Schopler 1986

Swettenham 2000

Tager-Flusberg 2000

Wellman 2000

Wellman 2001

Williams 2006

Yeargin-Allsopp 2003

References to other published versions of this review
Fletcher-Watson 2010
## Characteristics of included studies  
### Baghdadli 2013

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT. Control condition: structured group leisure activities</th>
</tr>
</thead>
</table>
| Participants | High-functioning autism  
Aged eight to 12 years  
All male  
Intellectual level - normal range. Participants matched on verbal IQ  
Diagnosis by clinical best estimate, confirmed by ADOS and ADI-R  
Final n = 13 (93%) |
| Interventions | Designed to teach ToM and precursor skills  
Targets emotion recognition from faces, conversation, social problem solving, stress management, and ToM  
Delivery in group treatment sessions with two therapists. Methods include role play, video-modelling and problem solving skills. Based on manual of treatment goals including techniques, learning stages and tools  
Dose: 90 minutes per week, 20 weeks spread over six months |
| Outcomes | Primary outcome: DANVA-2 short form (emotion recognition assessment)  
Secondary measures: KidScreen, parent-report measure of quality of life |
| Notes | This study was partially funded by the Pfizer Foundation. |

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Computerised random number generator</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Unclear risk</td>
<td>List given to therapists, though allocation did take place after baseline assessment</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>“Until the end of the study, none of the participants (children and parents) were told which group they were assigned to” (p. 436) . However not possible to blind them to group activity content</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>Low risk</td>
<td>Assessors of outcome measures were blinded</td>
</tr>
</tbody>
</table>
### Baghdadli 2013 (Continued)

<table>
<thead>
<tr>
<th>Incomplete outcome data (attrition bias)</th>
<th>Low risk</th>
<th>Use Intention-to-treat analysis as well as assessment of study-completers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td>Used independent fidelity coding system to ensure treatment adherence by therapists</td>
</tr>
</tbody>
</table>

### Begeer 2011

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT; wait-list control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>All forms of ASD</td>
</tr>
<tr>
<td></td>
<td>Aged eight to 13 years old</td>
</tr>
<tr>
<td></td>
<td>Both genders</td>
</tr>
<tr>
<td></td>
<td>Intellectual level: normal range. Participants matched on chronological age, full-scale IQ, verbal IQ, and non-verbal IQ</td>
</tr>
<tr>
<td></td>
<td>Diagnosis by clinical best estimate confirmed by SRS and ASQ</td>
</tr>
<tr>
<td></td>
<td>Final n = 36 (90%)</td>
</tr>
<tr>
<td>Interventions</td>
<td>Designed to teach ToM</td>
</tr>
<tr>
<td></td>
<td>Target ToM understanding</td>
</tr>
<tr>
<td></td>
<td>Delivered by manualised small group training, plus monthly parental training and homework. Sessions supervised by certified therapists. Training progresses through 53 structured settings from precursors of ToM (e.g. recognising emotions and intentions), through elementary skills (e.g. deception, understand others' mental states) to complex ToM skills (e.g. second order mental state reasoning)</td>
</tr>
<tr>
<td></td>
<td>Dose: one and a half hours weekly for 16 weeks</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Primary outcome: The ToM test.</td>
</tr>
<tr>
<td></td>
<td>Secondary measures: Levels of Emotional Awareness Scale - Children; self-reported empathy; parent-report social behaviours</td>
</tr>
<tr>
<td>Notes</td>
<td>Funding source not reported.</td>
</tr>
</tbody>
</table>

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>&quot;An independent researcher made the allocation schedule&quot;, p. 1000</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Unclear risk</td>
<td>Insufficient information available</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Not possible to blind personnel and participants</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Begeer 2011 (Continued)

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Unclear risk</td>
<td>No description of outcome assessors</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Unclear risk</td>
<td>Four families lost to follow-up, three from control group: “refused to return for testing”, p. 1000</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

### Bolte 2002

**Methods**

RCT: wait-list control

**Participants**

HFA and AS  
Aged 16 to 40 years  
All male  
Intellectual level: normal range. Participants matched on non-verbal IQ  
Diagnosis confirmed by ADOS & ADI-R  
Final n = 10 (100%)

**Interventions**

Designed to teach precursor skills of ToM  
Targets affect recognition from face and eyes  
Delivered by “FEFA” computer program: presents six basic emotions plus neutral faces, users match faces or pictures of eyes to the correct written label  
Dose: two hours per week for five weeks

**Outcomes**

Primary outcome measures: Built-in test condition and IAPS (International Affective Picture System) ratings  
Secondary outcome: fMRI

**Notes**

Bolte 2006 - second report of same data set.  
This study was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG)

---

**Risk of bias**

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Participants pooled and numbered randomly, then allocated alternately to each group (information provided by email)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Low risk</td>
<td>Participants assigned to group after initial assessment, using procedure described above</td>
</tr>
</tbody>
</table>
### Bolte 2002 (Continued)

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blinding of participants and personnel</td>
<td>High risk</td>
<td>Participants and personnel not blind</td>
</tr>
<tr>
<td>(performance bias)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection</td>
<td>Unclear risk</td>
<td>Outcome assessors not blind, but primary outcomes measures</td>
</tr>
<tr>
<td>bias)</td>
<td></td>
<td>computer administered (p. 65)</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>No evidence of attrition</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>High risk</td>
<td>IAPS data not reported in either publication; author attempted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>to locate these scores but no longer available fMRI loss of data in 2006 report unexplained.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

### Fisher 2005

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT; executive function training comparison group, also TAU control but non-random selection for this group</th>
</tr>
</thead>
</table>
| Participants | ASD and AS.  
Aged six to 15 years old.  
Both genders.  
Intellectual level: low range. Participants matched on verbal IQ and non-verbal IQ raw scores  
Diagnosis based on clinical best estimate plus DSM-IV checklist  
Final n = 27 (100%). |
| Interventions | Designed to teach ToM.  
Targets false belief-skills.  
One-to-one training by therapists, taught a strategy of thinking about beliefs as "photos in the head", and using illustrative stories. Progress across five stages of increasing complexity existence of mental states to comprehending false belief  
Dose: up to a maximum of 10 x 25mins (p. 763). |
| Outcomes | Primary Outcome: ToM ability, measured by false belief and other tasks, including teacher-report measure |
| Notes | This study was funded by the Medical Research Council, UK. |
### Fisher 2005 (Continued)

<table>
<thead>
<tr>
<th>Bias Type</th>
<th>Risk Level</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>High risk</td>
<td>Participants in control group selected by non-attendance at school during training period. Allocation to ToM and executive function intervention groups may be random, but no method detail provided (p. 759). Also “children from the same school were distributed across groups, to control for any school effects” (p. 759)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>Not addressed in the report, but control group not randomly assigned</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Impossible to conceal intervention type from therapist or participant but participants probably not aware of link between intervention type and different outcome measures</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>High risk</td>
<td>Unclear but seems likely that outcome tests were administered by therapists</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Unclear risk</td>
<td>Five teacher questionnaires are missing (p. 759) attributed to the fact that they had too many to complete each day</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

### Golan 2006

<table>
<thead>
<tr>
<th>Method</th>
<th>RCT; TAU control.</th>
</tr>
</thead>
</table>
| Participants | HFA & AS.  
Aged 17 to 52 years.  
Both genders.  
Intellectual level: normal range. Participants matched on chronological age, full-scale, verbal and non-verbal IQ  
Diagnosis confirmed by AQ.  
\( n = 41 \) (89%).  |
| Interventions | Designed to teach a precursor skill of ToM.  
Targets complex emotion recognition.  
Delivered by MindReading software at home: an emotion library, trainer, and games all promote understanding of how photographs and film clips of facial emotions match on to vocabulary and emotional stories  
Dose: required minimum use of 10 hours over a 10-week period |
Outcomes

Primary outcome: Emotion recognition using a range of computerised tests

Notes

This study was funded by multiple contributors comprising: the National Alliance for Autism Research, the Corob Charitable Trust, the Cambridge Overseas Trust, the B’nai B’rith Leo Baeck scholarships, the Shirley Foundation, the Medical Research Council and the Three Guineas Trust

Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>States random allocation but not described.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>Not described directly but report implies randomisation occurred before initial assessment</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Participants not blind but unaware of intervention goals of the study</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>Low risk</td>
<td>Three assessors were blind to group but one (the lead author) was not. Outcome measures were computer-presented and therefore resilient to bias</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>High risk</td>
<td>Five participants in the intervention group dropped out due to not finding time to complete the required amount of work (p. 597). Study may over-estimate value of intervention</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>One participant’s data not reported for one of many outcome measures (p. 602, Table 2)</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

Golan 2010

Methods

RCT; TAU control.

Participants

ASD.
Aged four to eight years.
Both genders.
Intellectual level: normal range. Participants matched for chronological age and verbal IQ
Diagnosis confirmed by ADI-R and CAST.
Final n = 38 (97%).
### Interventions

<table>
<thead>
<tr>
<th>Interventions</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designed to teach precursor skills of ToM</td>
<td></td>
</tr>
<tr>
<td>Targets emotion comprehension</td>
<td></td>
</tr>
<tr>
<td>Delivered by <em>Transporters</em> cartoon on DVD watched at home: cartoon episodes depict vehicles with</td>
<td></td>
</tr>
<tr>
<td>real human faces who experience emotional situations and display appropriate facial reactions</td>
<td></td>
</tr>
<tr>
<td>Dose: recommend three episodes per day for four weeks.</td>
<td></td>
</tr>
</tbody>
</table>

### Outcomes

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Outcomes: Emotion vocabulary and emotion recognition from complex situations</td>
<td></td>
</tr>
</tbody>
</table>

### Notes

<table>
<thead>
<tr>
<th>Notes</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study was funded by Culture Online and the UK Department for Culture, Media and Sport</td>
<td></td>
</tr>
</tbody>
</table>

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>States random allocation but procedure not described.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Unclear risk</td>
<td>Timing and method not described.</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Not possible to blind participants.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Unclear risk</td>
<td>Blinding of outcome assessors not reported, but computer presentation of outcomes measures resistant to bias</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>Only one participant dropped out (control group) but this is unexplained</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of this.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

### Goods 2013

<table>
<thead>
<tr>
<th>Methods</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT: TAU control (regular school program, 30 hrs per week).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participants</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism diagnosis.</td>
<td></td>
</tr>
<tr>
<td>Aged three to five years.</td>
<td></td>
</tr>
<tr>
<td>No gender information provided.</td>
<td></td>
</tr>
<tr>
<td>Intellectual level: low range. Participants matched on chronological age, full-scale IQ, and</td>
<td></td>
</tr>
<tr>
<td>verbal IQ</td>
<td></td>
</tr>
<tr>
<td>Clinical best estimate diagnosis confirmed by ADOS.</td>
<td></td>
</tr>
<tr>
<td>Final n = 11 (73%).</td>
<td></td>
</tr>
</tbody>
</table>
Interventions

Designed to teach a precursor skill of ToM.
Targets joint engagement and social communication skills.
Delivered by individual therapists sessions: children taken out of regular classroom activities. Approach also described in Kasari 2006 & Kasari 2010.

Dose: Two sessions of 30 minutes each per week, for 12 weeks intervention period

Outcomes

Primary outcomes: Early Social Communication Scales, Structured Play Assessment, classroom observation of joint engagement, and spontaneous communicative gestures

Notes

This study was funded by the Organisation for Autism Research and Autism Speaks

Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Blocked randomisation using SPSS random number generator.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>Allocation post-assessment but concealment not described. As the blocks were of fixed length the final allocation within each block would be known</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Not possible to blind participants or personnel.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>Low risk</td>
<td>Assessors blinded for all outcomes.</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>High risk</td>
<td>Attrition four out of 15 participants. Analysis considered completers only</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>High risk</td>
<td>One measure (Reynell Developmental Language Scales) reported in Table 2 (no group differences at Baseline or Exit) but not discussed</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>
### Hadwin 1996

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT: three intervention groups (emotion, belief, play).</th>
</tr>
</thead>
</table>
| Participants | Autism.  
Aged four to 13 years.  
Both genders.  
Intelectual level: mostly low range, but four children with verbal mental age in normal range (distributed across groups). Participants matched on chronological age, verbal mental age, and expressive language  
Diagnosis confirmed by DSM-III and DSM-IV checklist (see Hadwin 1997, p. 524). Final n = 30 (100%). |
| Interventions | Designed to teach ToM.  
Three intervention groups each targeting a specific ToM component: Understanding Emotion, Understanding Belief, and Pretend Play  
Delivered by one-to-one therapist led intervention: used a question and answer structure with corrective feedback. Play session additional involved modelling of pretend play. Good performance on each level required for progress to the next level of complexity  
Dose: One assessment session, then eight therapy sessions, followed by one post-test assessment session. Eight, consecutive daily half-hour therapeutic sessions |
| Outcomes | Primary outcome: teaching level reached on five-point scale.  
| Notes | also see Hadwin 1997.  
This study was funded by the Bethlem-Maudsley Research Fund and the Mental Health Foundation |

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>States random allocation but procedure not described.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>Baseline assessment conducted after allocation to treatment condition</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Not possible to blind participants or personnel.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>High risk</td>
<td>Not explicitly stated, but seems that both pre- and post-test data were collected by unblinded therapists</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>Low risk</td>
<td>No evidence of attrition.</td>
</tr>
</tbody>
</table>
### Selective reporting (reporting bias)

<table>
<thead>
<tr>
<th>Bias</th>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td></td>
<td>No evidence of selective reporting.</td>
</tr>
</tbody>
</table>

### Hopkins 2011

#### Methods
- RCT: control group with art and design computer program.

#### Participants
- **ASD.**
- Aged six; three to 15; one years.
- Both genders.
- Intellectual level: 24 in the normal range (HFA), 25 in a low-functioning range (LFA).
- Participants matched on chronological age, verbal IQ and non-verbal IQ.
- Diagnosis by clinical best estimate confirmed with CARS.
- Final n = 49 (96%).

#### Interventions
- Teaches precursor skills of ToM.
- Targets emotion recognition, eye contact and facial identity recognition.
- Delivered by FaceSay: social skills training via a computer program: three games to:
  - match gaze direction with an object;
  - select parts of faces which fit with a whole face;
  - and match facial expressions with emotions.
- Dose: 12 x 30 minute sessions over 6 weeks.

#### Outcomes
- **Primary outcome:** Emotion recognition from photos and drawings of faces and Benton Face Recognition Test.
- **Secondary outcomes:** Social Skills Rating System, Social skills observation.

#### Notes
- Authors were emailed for further information, no response.
- This study was partially funded by a grant from Civitan International and with the cooperation of Symbionica LLC.

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>No information provided.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Unclear risk</td>
<td>No information provided.</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>Unclear risk</td>
<td>Parents and participants technically blind but could have guessed from computer program content</td>
</tr>
<tr>
<td>All outcomes</td>
<td>Low risk</td>
<td>SSRS: observed by blinded raters. Other measures resistant to bias.</td>
</tr>
<tr>
<td>Hopkins 2011</td>
<td>(Continued)</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>Two children excluded due to low attendance, and reasons given</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ingersoll 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
</tr>
<tr>
<td>Participants</td>
</tr>
<tr>
<td>Interventions</td>
</tr>
<tr>
<td>Notes</td>
</tr>
</tbody>
</table>

**Risk of bias**

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Randomised by a coin toss.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Low risk</td>
<td>Pairing on expressive language age before assignment by coin toss</td>
</tr>
</tbody>
</table>
Ingersoll 2012  (Continued)

<table>
<thead>
<tr>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
</tr>
<tr>
<td>Other bias</td>
</tr>
</tbody>
</table>

<p>| Risk of bias |</p>
<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Not possible to blind participants and therapists.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>Low risk</td>
<td>In Ingersoll 2010, p. 1155 “Assessments were scored by trained research assistants blind to group” Confirmed in personal communication: although examiners were not blind, coders were</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>Low risk</td>
<td>Two children withdrew, and reasons given. Analysis of completers only</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Unclear risk</td>
<td>No reason given for adding additional subjects to Ingersoll 2010, to randomisation may not be secure</td>
</tr>
</tbody>
</table>

Kaale 2012

<p>| Risk of bias |</p>
<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other bias</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| Risk of bias |</p>
<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>RCT: TAU control.</td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>Autism and ASD. Aged 29 to 60 months. Both genders. Intellectual level: low range. Participants matched on chronological age, mental age, expressive and receptive language ages, and developmental level Diagnosis confirmed by ADOS (or ADI for 49 participants). Final n = 61 (100%).</td>
<td></td>
</tr>
<tr>
<td>Interventions</td>
<td>Designed to teach precursor skills of ToM. Targets joint attention and joint engagement. Delivered as manualised Joint attention intervention, by pre-school teachers: table top training sessions provide multiple opportunities for a child to initiate a targeted skill. Skills are encouraged using verbal and physical prompts, presenting opportunities and positive feedback Dose: two daily sessions, five days per week for eight weeks</td>
<td></td>
</tr>
<tr>
<td>Outcomes</td>
<td>Primary outcome: ESCS. Secondary outcome: Teacher-child play and mother-child, both rated for joint attention and joint engagement</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td>This study was funded by the South-Eastern Norway Regional Health Authority and the Centre for Child and Adolescent Mental Health (Eastern and Southern Norway)</td>
<td></td>
</tr>
</tbody>
</table>
### Kaale 2012

<table>
<thead>
<tr>
<th>Bias</th>
<th>Risk</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low</td>
<td>Concealed randomisation list generated before the study began</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Low</td>
<td>Allocation after baseline assessment.</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High</td>
<td>Parents and pre-school staff were blind to allocation at baseline (p. 98), but therapists were not blind (p. 101). Participants likely to be unaware of intervention</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Low</td>
<td>Blinded coders and outcome assessors (p. 98).</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low</td>
<td>No evidence of attrition.</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low</td>
<td>All pre-specified outcomes were reported.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>

### Kasari 2006

**Methods**

RCT; two control groups: TAU and symbolic play.

**Participants**

ASD and core autism.
Aged 36 to 60 months.
Both genders.
Intellectual level: low range. Participants matched on chronological age, mental age, developmental level and expressive and receptive language ages
Diagnosis confirmed by ADOS and ADI.
Final n = 58 (89%).

**Interventions**

Designed to teach precursor skills of ToM.
Targets joint attention.
Delivered as one to one therapist led intervention: table top training and then floor sessions provide multiple opportunities for a child to initiate a targeted skill. Skills are encouraged using verbal and physical prompts, presenting opportunities and positive feedback. Structure is withdrawn as the training progresses
Dose: 30 minutes daily for five to six weeks.

**Outcomes**

Primary outcomes: ESCS.
Secondary outcome: parent-child play rating, coded for joint engagement and joint attention

**Notes**

See also:
Gulsrud 2007: sub-analysis of n = 35 with age range 33 to 54 months.
Arora 2008: PhD thesis reporting on links between joint attention and perseveration for n = 35
**Risk of bias**

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>“used generated numbers from computer program” (information provided by author)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>“Families who consented were randomised and began treatment right after assessments (by outside independent assessors)” and “allocation to group was done as kids entered the Early Intervention Program and parent consented - next allocation of group number was revealed after eligibility and consenting” (information provided by author). However the published report states that some children were excluded because they did not meet inclusion criteria (p. 612) indicating that assessments were completed but not scored before allocation</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Not possible to blind parents, children or therapists, though children likely to have been unaware</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Low risk</td>
<td>Outcome assessors and coders both blind to group (information provided by author)</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Unclear risk</td>
<td>Seven participants did not meet criteria, refused final assessment or left the program unexpectedly. Distributed across groups (joint attention intervention = two; symbolic play intervention = one; control group = four)</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Unclear risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>
**Kasari 2010**

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT: wait-list control.</th>
</tr>
</thead>
</table>
| Participants | Core autism.  
Aged 21 to 36 months.  
Both genders.  
Intellectual level: low range. Participants matched on chronological age, mental age, and developmental level  
Diagnosis by clinical best estimate and ADI.  
Final n = 35 (83%). |
| Interventions | Designed to teach precursor skills of ToM.  
Targets joint attention and joint engagement.  
Delivered as parent-mediated training adapted from Kasari 2006. Core principles form 10 modules, delivered by parents trained by therapists. Uses aspects of applied behaviour analysis and facilitative and responsive interaction to encourage target behaviours in the child  
Dose: three modules per week for eight weeks. |
| Outcomes | Primary outcome: Videotaped parent-child interaction, coded for engagement, type of play, and frequency of joint attention  
Secondary outcomes: caregiver quality of involvement, adherence to treatment, and service utilisation measure |
| Notes | This study was funded by the National Institute for Mental Health |

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Random numbers list.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Low risk</td>
<td>Children randomised after meeting study criteria.</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Not possible to blind parents and participating children.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>Low risk</td>
<td>Examiners and coders both blind to group status.</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>Low risk</td>
<td>Two children did not meet criteria, two parents declined to participate. Three further children from the control group described as &quot;did not receive allocated intervention&quot;. Though not explicitly stated it appeared from Table 4 that analyses were based on intention-to-treat rather than treatment completers alone.</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting (pp. 1052-1053).</td>
</tr>
</tbody>
</table>
### Other bias

- **Low risk**

---

### Kim 2009

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT with cross-over; toy play session control.</th>
</tr>
</thead>
</table>
| Participants | Core autism.   
| | Aged 39 to 71 months.   
| | Final sample all male.   
| | Intellectual level: normal range. Participants not matched but evaluated for developmental level   
| | Diagnosis confirmed by Korean CARS and four received ADOS as well   
| | Final n = 10 (66%). |
| Interventions | Designed to teach precursor skills of ToM.   
| | Targets joint attention.   
| | Delivered as improvised music sessions based on a semi-flexible treatment manual. Sessions included free play time and directed activities with therapist modelling turn-taking and other activities   
| | Dose: 12 weeks, 30 minutes per week. |
| Outcomes | Primary outcomes: PDD-BI social approach subscale, and ESCS.   
| | Secondary outcomes: observational coding of emotional and motivation responsiveness, and responsiveness to joint attention / joint attention initiation |
| Notes | See also Kim 2008 for details of PDD-BI and ESCS outcome measure.   
| | Email sent to request more information but none received.   
| | This study was funded by Aalborg University, Denmark. |

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>“participants were randomly assigned to two groups” - but no details given</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Unclear risk</td>
<td>No details given.</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Not possible to blind participants and personnel.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Low risk</td>
<td>Coders for PDD-BI and ESCS both blind. PDD-BI administered to both blind teachers and unblinded parents, so in the latter case, susceptible to bias even if the interviewer is blind 30% of observational behaviour measures</td>
</tr>
</tbody>
</table>
### Kim 2009 (Continued)

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Unclear risk</td>
<td>Five children dropped out due to ill health or long distances to travel to receive intervention</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td>Differences in ratings provided by mothers and professionals. Professional ratings support efficacy of intervention and these were also blind. Mothers over-estimate pre-intervention abilities and thus under-estimate efficacy of the intervention, but they were not blind (in Kim, 2008, p. 1763)</td>
</tr>
</tbody>
</table>

### Landa 2011

**Methods**

RCT: contact (non-interpersonal synchrony) control.

**Participants**

Autism and ASD. Aged 21 to 33 months. Both genders. Intellectual level: low range. Participants matched on chronological age, receptive language and visual reception standardised scores. Diagnosis confirmed by ADOS. Final n = 48 (96%).

**Interventions**

Designed to teach precursor skills of ToM. Targets socially synchronous behaviour. Delivered by trained teacher using the Assessment Evaluation Program System (AEPS) for infants and children. Instructional strategies ranged from discrete trial teaching to pivotal response training and routine-based interactions. Intervention incorporated low-tech communication systems. Highly motivating tasks and materials designed to elicit frequent child-initiated communication. Dose: two and a half hours per day, four days per week for six months in classroom. Additional one and a half hours per month parent training and 38 hours parent education.

**Outcomes**

Primary Outcome: Socially engaged imitation in a structured imitation task. Secondary outcomes: Initiations of joint attention & shared positive affect - measured using the CSBS; Mullen Scales of Early Learning expressive language and visual reception subscales.

**Notes**

This study was funded by the National Institute of Mental Health and HRSA.
Random sequence generation (selection bias) | Unclear risk | No information provided.
---|---|---
Allocation concealment (selection bias) | Unclear risk | Participants placed into matched pairs prior to randomisation which could introduce bias
Blinding of participants and personnel (performance bias) | Low risk | Parent, children and teachers all blind to group: not possible to blind intervention trainers
Blinding of outcome assessment (detection bias) | Low risk | Assessment at baseline and follow-up by a independent, blind clinician (p. 16)
Incomplete outcome data (attrition bias) | Low risk | One family withdrew from the study, and one was not analysed due to missing baseline data
Selective reporting (reporting bias) | Low risk | No evidence of selective reporting.
Other bias | Low risk | ---

**Ryan 2010**

**Methods**
RCT: wait-list control.

**Participants**
Autism and ASD.
Aged six to 14 years.
Both genders.
Intellectual level: normal range. Participants matched on verbal IQ and non-verbal IQ
Diagnosis confirmed by ADOS and DISCO.
Final n = 30 (100%).

**Interventions**
Designed to teach precursor skills of ToM.
Targets emotion recognition.
Delivered as therapist-led training using photos of six core facial emotions. Therapists highlighted component parts of emotional expressions (e.g. raised eye-brows) and matched faces with verbal labels. Participants engaged in matching games, role-play, tracing, and drawing of faces to support direct instruction
Dose: one hour per week over four weeks.

**Outcomes**
Primary outcome: Facial emotion recognition test.

**Notes**
This study was funded by the COPE Foundation and supported by the Catherine T MacArthur Foundation Research Network on Early Experience and Brain Development

**Risk of bias**
### Ryan 2010 (Continued)

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>“We assigned sequential numbers to each child (ordered alphabetically) to be included in the study and then used the random number table provided in David Clark-Carter’s book “Doing Quantitative Psychological Research - From Design to Report” (1997) to form two groups” (information provided by author)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>Assessments conducted before allocation, but using open random numbers table could have introduced bias</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Not possible to blind personnel and participants.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Unclear risk</td>
<td>Independent outcome assessor was “blind to the initial scores achieved by each of the children, to the membership of the treatment groups and to our hypotheses” (information provided by author)</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>No evidence of attrition at four weeks, immediately post intervention. However five children lost to follow-up at three months, after all had received the training, without reason given</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting, but only one outcome measure was used</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

### Schertz 2013

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT: TAU control.</th>
</tr>
</thead>
</table>
| Participants             | Autism spectrum disorder and autism.  
Aged under 30 months (mean age in each group: 24.6 months and 27.5 months)  
Gender information not reported.  
Intellectual level: normal range. Participants matched on chronological age, receptive and expressive language  
Diagnosis confirmed by ADOS and M-CHAT (to establish high risk)  
Final n = 23 (100%, but see risk of bias below). |
| Interventions            | Designed to teach precursor skill of ToM.  
Targets joint attention and sub-skills (Focus on Faces, Turn-Taking)  
Delivered as parent-training delivered by intervention co-ordinators at home. Sessions included guidance and reflection based on filmed segments of parent-child play, and a manual for parents |
Dose: Parent training sessions were weekly at home for at least 15 sessions. Parents agree to spend 30 minutes per day in face-to-face interaction with their child. Total intervention period ranged from 4 to 12 months (mean 7 months)

Outcomes
Primary outcomes: Precursors of Joint Attention Measure (PJAM), based on coding of parent-child interaction
Secondary measures: Vineland Adaptive Behaviour Scales and Mullen Scales of Early Learning

Notes
This study was funded by Autism Speaks.

<table>
<thead>
<tr>
<th>Risk of bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Unclear risk</td>
<td>Once entry criteria were met, children were paired and then assigned to intervention or control. Randomisation was by coin toss (confirmed by author email)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>No information given, but coin toss open to abuse if performed by intervention co-ordinators</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Not possible to blind parents and intervention co-ordinators</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Low risk</td>
<td>Intervention co-ordinators and parents carried out assessments but these were coded by raters blind to group (but not time point)</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>There is unreported attrition of one participant (confirmed by author)</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Unclear risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>High risk</td>
<td>The treatment period was of variable length. Treatment was terminated (and outcomes measures taken) once three or more instances of initiation of joint attention were observed across multiple sessions. Within each pair, inter-assessment period was identical, but this variable system means that 100% of participants made gains in responses to treatment. i.e. the treatment effect was weighted by this study design feature</td>
</tr>
</tbody>
</table>
### Methods
RCT, wait-list control.

### Participants
HFA, AS and PDD-NOS.
- Aged eight to 12 years.
- All male.
- Intellectual level: normal range, Participants matched on chronological age and full-scale IQ.
- Diagnosis confirmed by ADOS and ADI.
- Final n = 18 (100%).

### Interventions
Designed to teach ToM.
- Targets social adjustment, emotion recognition, ToM, and executive functions.
- Delivered as group social skills training sessions with parent psychoeducational sessions.
- Children's group sessions followed a strict structure, including welcome, lesson time, games and free social time. The curriculum followed 10 modules focusing on emotional understand and empathy, conversational skills, and lessons about friendship.
- Dose: one and a half hours per week for 20 weeks.

### Outcomes
Primary outcomes: Diagnostic Analysis of Non-Verbal Accuracy - facial expression recognition subscale; Happé's Strange Stories; The Faux-Pas Recognition Test; Test of Problem Solving (TOPS).
Secondary outcomes: Depression Inventories for children and parents; problem behaviour logs.

### Notes
Authors were emailed for further information but no response received.
This study was funded by the MIND Institute, UC Davis, California, USA.

### Risk of bias
<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Participants were matched into pairs and then “Nine boys were randomly selected... to serve as the first intervention group subjects” - but method unclear (p. 654)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Unclear risk</td>
<td>Participants assessed and matched before random allocation, but pairing process could have introduced bias</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>High risk</td>
<td>Not possible to blind participants and personnel.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>High risk</td>
<td>Unlikely to have been blinded as assessments carried out by first author</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>Low risk</td>
<td>No evidence of attrition.</td>
</tr>
</tbody>
</table>
**Solomon 2004 (Continued)**

<table>
<thead>
<tr>
<th>Selective reporting (reporting bias)</th>
<th>Low risk</th>
<th>No evidence of selective reporting.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

**Williams 2012**

<table>
<thead>
<tr>
<th>Methods</th>
<th>RCT; control group Thomas the Tank Engine DVD.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>Autism spectrum disorder, including autism.</td>
</tr>
<tr>
<td></td>
<td>Aged four to seven years.</td>
</tr>
<tr>
<td></td>
<td>Intellectual level: low to normal range, but cognitive level above 30 months minimum.</td>
</tr>
<tr>
<td></td>
<td>Participants matched on chronological age, full-scale IQ, verbal IQ and non-verbal IQ</td>
</tr>
<tr>
<td></td>
<td>Diagnosis based on clinical best estimate confirmed by ADOS.</td>
</tr>
<tr>
<td></td>
<td>n = 55 (92%).</td>
</tr>
<tr>
<td>Interventions</td>
<td>Designed to teach a precursor skill of ToM.</td>
</tr>
<tr>
<td></td>
<td>Targets emotion recognition.</td>
</tr>
<tr>
<td></td>
<td>Delivered by <em>Transporters</em> DVD: cartoon episodes depict vehicles with real human faces who experience emotional situations and display appropriate facial reactions</td>
</tr>
<tr>
<td></td>
<td>Dose: 15 minutes per day for four weeks (mean 11.76 hours total)</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Primary outcomes: NEPSY-II affect recognition test using pictures, and the Pictures of Facial Affect test (using Ekman faces: identify emotions from photos of faces)</td>
</tr>
<tr>
<td></td>
<td>Secondary measures: NEPSY-II ToM tasks; Mindreading tasks; Vineland Adaptive Behaviour Scales (socialisation subscale)</td>
</tr>
<tr>
<td>Notes</td>
<td>This study was funded by the Financial Markets Foundation for Children, Australia</td>
</tr>
</tbody>
</table>

**Risk of bias**

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors’ judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Statistician randomised the allocation.</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Low risk</td>
<td>Each DVD packed in an unmarked numbered envelope - adequate concealment</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Participants not blind to DVD content.</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Low risk</td>
<td>Assessor blind to treatment group.</td>
</tr>
<tr>
<td>All outcomes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Incomplete outcome data (attrition bias)

**All outcomes**

<table>
<thead>
<tr>
<th>Risk of Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>High risk</td>
<td>Attrition of five participants after one month (unbalanced between groups) and only completers analysed. Missing data on emotion identification where children (n = eight) could not complete the task: these may relate to outcome. Effect sizes not given. Attription of nine more participants at three-month follow-up (unbalanced between groups) so follow-up data used last observation carried forward.</td>
<td></td>
</tr>
</tbody>
</table>

Selective reporting (reporting bias)

<table>
<thead>
<tr>
<th>Risk of Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td>No evidence of selective reporting.</td>
<td></td>
</tr>
</tbody>
</table>

Other bias

<table>
<thead>
<tr>
<th>Risk of Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Wong 2010

**Methods**

RCT cross-over: wait-list control.

**Participants**

ASD.  
Aged 17 to 36 months.  
Both genders.  
Intellectual level: low range. Participants matched on chronological age and mental age  
Diagnosis confirmed by ADI-R, ADOS and CARS.  
Final n = 17 (100%).

**Interventions**

Designed to teach precursor skills of ToM.  
Targets eye contact, gesture and vocalisation.  
Delivered as direct one to one therapy: modelling gestures accompanied by verbal labels, eliciting requests using favourite toys  
Dose: 10 x 30 minute sessions over two weeks.

**Outcomes**

Primary outcome: ADOS (reported item by item).  
Secondary outcomes: Ritvo-Freeman Real Life Rating Scale (RFRLRS) - parent rated; Symbolic Play Test; Parenting Stress Index, short form.

**Notes**

Authors emailed for more detail but no reply received.  
Funding source unknown.

### Risk of bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>High risk</td>
<td>&quot;17 children consecutively diagnosed were randomized into the Intervention and Control groups&quot; – assignment may have been alternate rather than random? (p. 679)</td>
</tr>
</tbody>
</table>
### Wong 2010  
*Continued*

<table>
<thead>
<tr>
<th>Risk of bias</th>
<th>High risk</th>
<th>Low risk</th>
<th>Unclear risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Baseline assessment carried out after randomisation to group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias) All outcomes</td>
<td>Not possible to blind participants or personnel.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias) All outcomes</td>
<td>ADOS and SPT both administered and scored by blind assessors. Some outcomes are parent-reported measures and therefore unblinded. (pp. 680-681)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias) All outcomes</td>
<td>No evidence of attrition.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selective reporting (reporting bias) Low risk</td>
<td>No evidence of selective reporting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other bias Unclear risk</td>
<td>Wait-list control group combined with Intervention group to give larger sample size for analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Wong 2013

**Methods**  
Cluster-RCT: repeated measures, nested within children, nested within classrooms. Control conditions: symbolic play intervention, wait-list control

**Participants**  
Autism.  
Aged three to six years.  
Both genders.  
Intellectual level: low range. Participants matched on chronological age, mental age, and expressive and receptive language ages.  
Diagnosis based on clinical best estimate and confirmed by CARS  
Final n = 33 (97%).

**Interventions**  
Designed to teach precursor skills of ToM.  
Targets joint attention and interpersonal engagement behaviours.  
Delivered as teacher training. Teacher then implemented the approach in the classroom according to their own preference (e.g. one-to-one, small group, whole class).  
Dose: teacher training sessions were one hour per week, for four weeks.

**Outcomes**  
Also direct classroom observation of child and teacher behaviours (joint engagement, joint attention)  
Secondary measures: teacher acceptability of intervention.

**Notes**  
Funding source unknown.
### Bias

<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>Low risk</td>
<td>Teacher assignment to one of three groups by random numbers list</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>Low risk</td>
<td>Allocation occurred before initial assessment, but allocation list concealed from baseline assessors (confirm by author email)</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>High risk</td>
<td>Teachers not blinded, as recipients of training.</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Low risk</td>
<td>Research observers for all measures were blinded (confirmed by author email)</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>One teacher (and therefore one child) dropped out; reason given</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>No evidence of selective reporting.</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

### Young 2012

**Methods**
- RCT; control group Thomas the Tank Engine DVD.

**Participants**
- All forms of ASD.
- Aged four to eight years.
- Intellectual level: normal range. Participants matched on verbal and non-verbal IQ
- Diagnosis based on clinical best estimate confirmed by SCQ.
- Final n = 25 (100%).

**Interventions**
- Designed to teach precursor skills of ToM.
- Targets emotion recognition.
- Delivered by Transports DVD: cartoon episodes depict vehicles with real human faces who experience emotional situations and display appropriate facial reactions
- Dose: five to 10 minute episodes, children watch three per day for three weeks

**Outcomes**
- Primary outcome: Social Communication Questionnaire, parent-report
- Secondary outcomes: NEPSY-II affect recognition subscale using pictures; The Faces Task - emotion recognition from photos of faces

**Notes**
- Funding source unknown.

**Risk of bias**
<table>
<thead>
<tr>
<th>Bias</th>
<th>Authors' judgement</th>
<th>Support for judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sequence generation (selection bias)</td>
<td>High risk</td>
<td>“Random allocation was done by simply alternating between the 2 intervention groups for each participant as they responded to the letters which were sent out. .. we tossed a coin to see which group went first and it was Thomas and then the next person who responded was Transporters and so on” (information from the author)</td>
</tr>
<tr>
<td>Allocation concealment (selection bias)</td>
<td>High risk</td>
<td>Published report states that assignment was after baseline assessment (p. 987) but author’s personal communication states that “random allocation occurred before baseline assessment - when the participant registered interest in the study they were allocated a condition”</td>
</tr>
<tr>
<td>Blinding of participants and personnel (performance bias)</td>
<td>Low risk</td>
<td>Parents and children aware of DVD content but “they were naive to the purpose of the study” (information from the author)</td>
</tr>
<tr>
<td>Blinding of outcome assessment (detection bias)</td>
<td>Unclear risk</td>
<td>“assessor was not blind to treatment but the measures didn’t require any subjective interpretation. Parents completed questionnaires and they were naive to the purpose of the study” (information from the author)</td>
</tr>
<tr>
<td>Incomplete outcome data (attrition bias)</td>
<td>Low risk</td>
<td>No evidence of attrition.</td>
</tr>
<tr>
<td>Selective reporting (reporting bias)</td>
<td>Low risk</td>
<td>All pre-specified outcomes are reported. So too are four outcome measures not pre-specified, nor described in the methods</td>
</tr>
<tr>
<td>Other bias</td>
<td>Low risk</td>
<td></td>
</tr>
</tbody>
</table>

ADI-R: Autism Diagnostic Interview - Revised  
ADOS: Autism Diagnostic Observation Schedule  
AS: Asperger’s syndrome  
ASD: autism spectrum disorder  
ASQ: autism spectrum quotient  
CARS: Childhood Autism Rating Scale  
DANVA-2: Diagnostic Analysis of Non-Verbal Accuracy  
DISCO: Diagnostic Interview for Social and Communication Disorders  
DSM: Diagnostic and Statistical Manual of Mental Disorders  
ESCS: Early Social Communication Scales
Characteristics of excluded studies  
[ordered by study ID]

<table>
<thead>
<tr>
<th>Study</th>
<th>Reason for exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams 2012</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Aldred 2004</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Beaumont 2008</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Cardon 2011</td>
<td>Used a multiple baseline case-control design</td>
</tr>
<tr>
<td>Carter 2011</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Casenhiser 2011</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Castorina 2011</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Charman 2007</td>
<td>No data reported</td>
</tr>
<tr>
<td>Corbett 2011</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Dawson 2010</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>DeRosier 2011</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Drew 2002</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Estes 2011</td>
<td>No data reported</td>
</tr>
<tr>
<td>Field 2001</td>
<td>Experimental pilot study: not a full intervention trial</td>
</tr>
<tr>
<td>Frankel 2010a</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Last Name</td>
<td>Notes</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Frankel 2010b</td>
<td>No data reported</td>
</tr>
<tr>
<td>Gantman 2012</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Gattino 2011</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Geretsegger 2012</td>
<td>No data reported</td>
</tr>
<tr>
<td>Girolametto 1988</td>
<td>Not an ASD sample</td>
</tr>
<tr>
<td>Green 2010</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Hanley-Hochdorfer 2010</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Heimann 2006</td>
<td>Experimental pilot study: not a full intervention trial</td>
</tr>
<tr>
<td>Hillier 2012</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Howlin 1997</td>
<td>No data reported</td>
</tr>
<tr>
<td>Howlin 2007</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Jocelyn 1998</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Jones 2010</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Kasari 2012a</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Kern 2011</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Koenig 2010</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Lang 2010</td>
<td>No data reported</td>
</tr>
<tr>
<td>Laugeson 2012</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Lawton 2012</td>
<td>Broad-base intervention, not ToM-specific, teaches play skills as well as joint attention</td>
</tr>
<tr>
<td>LeGoff 2004</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Lerner 2012</td>
<td>Broad-base intervention, not ToM-specific, teaches a wide range of complex social skills</td>
</tr>
<tr>
<td>Lopata 2010</td>
<td>Broad-base intervention, not ToM-specific, teaches a wide range of complex social skills</td>
</tr>
<tr>
<td>Mahoney 2005</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>McConachie 2004</td>
<td>No data reported</td>
</tr>
<tr>
<td>Study</td>
<td>Notes</td>
</tr>
<tr>
<td>----------------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>McGregor 1998</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Minshew 2010</td>
<td>No data reported</td>
</tr>
<tr>
<td>Morgan 2010</td>
<td>No data reported</td>
</tr>
<tr>
<td>Nefdt 2010</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Oosterling 2010</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Ozonoff 1995</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Pajareya 2011</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Quirmbach 2009</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Roberts 2011</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Rodgers 2012</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Rogers 2010</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Schertz 2007</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Schmidt 2012</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Silver 2001</td>
<td>ASD diagnosis unconfirmed</td>
</tr>
<tr>
<td>Smith 2000</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Smith 2004</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Stichter 2001</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Strain 2011</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Swettenham 1996</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Tanaka 2010</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Temple 2007</td>
<td>No data reported</td>
</tr>
<tr>
<td>Thomeer 2012</td>
<td>Broad-base intervention, not ToM-specific, teaches a wide range of complex social skills</td>
</tr>
<tr>
<td>Turner-Brown 2008</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>Wellman 2002</td>
<td>Not an RCT/qRCT</td>
</tr>
</tbody>
</table>
Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whalen 2003</td>
<td>Not an RCT/qRCT</td>
</tr>
<tr>
<td>White 2010</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Wood 2009</td>
<td>Not ToM-linked</td>
</tr>
<tr>
<td>Yoder 2006</td>
<td>Broad-base intervention, not ToM-specific</td>
</tr>
<tr>
<td>Yoder 2006b</td>
<td>Not ToM-linked</td>
</tr>
</tbody>
</table>

ASD: autism spectrum disorder  
RCT: randomised controlled trial  
qRCT: quasi-randomised controlled trial  
ToM: Theory of Mind

Characteristics of studies awaiting assessment  [ordered by study ID]

Baghdadi 2010

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>RCT</td>
</tr>
<tr>
<td>Participants</td>
<td>Fourteen boys aged between eight to 12 years old</td>
</tr>
<tr>
<td>Interventions</td>
<td>Comparison of the effect of two prosocial interventions (social skill entertainment versus general educative intervention)</td>
</tr>
<tr>
<td>Outcomes</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td>Awaiting translation</td>
</tr>
</tbody>
</table>

RCT: randomised controlled trial
## DATA AND ANALYSES

### Comparison 1. Treatment effects in meta-analysis

<table>
<thead>
<tr>
<th>Outcome or subgroup title</th>
<th>No. of studies</th>
<th>No. of participants</th>
<th>Statistical method</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Joint engagement in mother-child interaction</td>
<td>2</td>
<td>88</td>
<td>Std. Mean Difference (IV, Random, 95% CI)</td>
<td>0.55 [0.11, 0.99]</td>
</tr>
<tr>
<td>2 Emotion recognition from face photographs, TAU control</td>
<td>4</td>
<td>105</td>
<td>Std. Mean Difference (IV, Random, 95% CI)</td>
<td>0.75 [0.22, 1.29]</td>
</tr>
<tr>
<td>3 Joint attention initiations in standardised assessment</td>
<td>3</td>
<td>92</td>
<td>Std. Mean Difference (IV, Random, 95% CI)</td>
<td>0.23 [-0.48, 0.94]</td>
</tr>
</tbody>
</table>

### Analysis 1.1. Comparison 1 Treatment effects in meta-analysis, Outcome 1 Joint engagement in mother-child interaction.

Review: Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD)

Comparison: Treatment effects in meta-analysis

Outcome: Joint engagement in mother-child interaction

<table>
<thead>
<tr>
<th>Study or subgroup</th>
<th>Experimental</th>
<th>Control</th>
<th>Std. Mean Difference</th>
<th>Weight</th>
<th>Std. Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaale 2012</td>
<td>26</td>
<td>27</td>
<td>57.3 (22.8)</td>
<td>61.6 %</td>
<td>0.37 [-0.17, 0.92]</td>
</tr>
<tr>
<td>Kasari 2010</td>
<td>19</td>
<td>16</td>
<td>42.85 (19.96)</td>
<td>38.4 %</td>
<td>0.84 [0.14, 1.53]</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>45</td>
<td>43</td>
<td>100.0 %</td>
<td>0.55 [0.11, 0.99]</td>
<td></td>
</tr>
</tbody>
</table>

Heterogeneity: $\tau^2 = 0.01$; $\chi^2 = 1.05$, df = 1 ($P = 0.30$); $I^2 = 5$

Test for overall effect: $Z = 2.45$ ($P = 0.014$)

Test for subgroup differences: Not applicable
Analysis 1.2. Comparison 1 Treatment effects in meta-analysis, Outcome 2 Emotion recognition from face photographs, TAU control.

Review: Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD)

Comparison: 1 Treatment effects in meta-analysis

Outcome: 2 Emotion recognition from face photographs, TAU control

<table>
<thead>
<tr>
<th>Study or subgroup</th>
<th>Experimental</th>
<th>N Mean(SD)</th>
<th>Control</th>
<th>N Mean(SD)</th>
<th>Std. Mean Difference</th>
<th>Weight</th>
<th>Std. Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IV, Random, 95% CI</td>
<td></td>
<td>IV, Random, 95% CI</td>
</tr>
<tr>
<td>Golan 2006</td>
<td>19</td>
<td>37.5 (7.8)</td>
<td>22</td>
<td>34.8 (8.2)</td>
<td>-36.1 %</td>
<td>0.33</td>
<td>[ -0.29, 0.95 ]</td>
</tr>
<tr>
<td>Hopkins 2011</td>
<td>13</td>
<td>5 (1)</td>
<td>11</td>
<td>3.31 (1.18)</td>
<td>-22.5 %</td>
<td>1.50</td>
<td>[ 0.58, 2.43 ]</td>
</tr>
<tr>
<td>Ryan 2010</td>
<td>20</td>
<td>19 (2.49)</td>
<td>12</td>
<td>16.83 (2.37)</td>
<td>-29.4 %</td>
<td>0.86</td>
<td>[ 0.11, 1.62 ]</td>
</tr>
<tr>
<td>Solomon 2004</td>
<td>4</td>
<td>12.8 (2.9)</td>
<td>4</td>
<td>11.8 (2.4)</td>
<td>-12.0 %</td>
<td>0.33</td>
<td>[ -1.08, 1.73 ]</td>
</tr>
<tr>
<td><strong>Total (95% CI)</strong></td>
<td><strong>56</strong></td>
<td><strong>49</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>100.0 %</strong></td>
<td><strong>0.75 [ 0.22, 1.29 ]</strong></td>
</tr>
</tbody>
</table>

Heterogeneity: Tau² = 0.11; Chi² = 4.70, df = 3 (P = 0.19); I² = 36%
Test for overall effect: Z = 2.75 (P = 0.0060)
Test for subgroup differences: Not applicable
Analysis 1.3. Comparison 1 Treatment effects in meta-analysis, Outcome 3 Joint attention initiations in standardised assessment.

Review: Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD)

Comparison: 1 Treatment effects in meta-analysis

Outcome: 3 Joint attention initiations in standardised assessment

<table>
<thead>
<tr>
<th>Study or subgroup</th>
<th>Treatment</th>
<th>Control</th>
<th>Std. Mean Difference</th>
<th>Weight</th>
<th>Std. Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N Mean(SD)</td>
<td>N Mean(SD)</td>
<td>IV,Random,95% CI</td>
<td></td>
<td>IV,Random,95% CI</td>
</tr>
<tr>
<td>Goods 2013</td>
<td>5 0.4 (0.89)</td>
<td>6 1 (1.73)</td>
<td></td>
<td>21.9 %</td>
<td>-0.39 [-1.59, 0.82]</td>
</tr>
<tr>
<td>Ingersoll 2012</td>
<td>14 4.79 (5.41)</td>
<td>13 0.92 (1.44)</td>
<td></td>
<td>33.7 %</td>
<td>0.93 [0.13, 1.73]</td>
</tr>
<tr>
<td>Kaale 2012</td>
<td>27 1.6 (2.6)</td>
<td>27 1.6 (2.2)</td>
<td></td>
<td>44.3 %</td>
<td>0.00 [-0.53, 0.53]</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>46</td>
<td>46</td>
<td>100.0 %</td>
<td>0.23</td>
<td>[-0.48, 0.94]</td>
</tr>
</tbody>
</table>

Heterogeneity: Tau² = 0.22; Chi² = 4.66, df = 2 (P = 0.10); I² = 57%

Test for overall effect: Z = 0.63 (P = 0.53)

Test for subgroup differences: Not applicable

### ADDITIONAL TABLES

Table 1. Additional methods

<table>
<thead>
<tr>
<th>Review section</th>
<th>Item</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit of Analysis</td>
<td>Cluster-randomised trials</td>
<td>Authors will use a summary measure from each cluster and conduct the analysis at the level of allocation (that is sample size = number of clusters). However, if there are very few clusters this would significantly reduce the power of the trial, in which case the authors will attempt to extract a direct estimate of the risk ratio using an analysis that accounts for the cluster design, such as a multilevel model, a variance components analysis or generalised estimating equations (GEEs). Statistical advice will be sought to determine which method is appropriate for the particular trials to be included</td>
</tr>
<tr>
<td>Subgroup Analysis</td>
<td>Identification of dimensions for subgroup analysis</td>
<td>In future updates the following clinically-relevant differences may be the focus of subgroup analyses: 1. intervention delivery type (e.g. therapist, parent-mediated, school-based) and length</td>
</tr>
</tbody>
</table>
### Table 1. Additional methods (Continued)

<table>
<thead>
<tr>
<th>Sensitivity Analysis</th>
<th>Identification of variables for sensitivity analysis</th>
<th>In future updates the impact of factors such as high rates of loss to follow-up or inadequate blinding on outcomes will be explored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dealing with Missing Data</td>
<td>Procedures for imputation in the event of issues with missing data</td>
<td>Should unacceptable levels and/or non-random missing data be found in future studies for inclusion in the review, the authors will attempt to impute missing values. Imputation may use individual data (where available from the original report authors) OR group-level summary statistics (which are normally included in published reports). Mean imputation will be used where variables are normally distributed, and the median will be used for non-normal distributions. In either case the review will report how the imputed values appear to change the outcome of the study/meta-analysis and use this variability to inform the strength of our conclusions</td>
</tr>
</tbody>
</table>

### Table 2. Outcome measures used

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Category</th>
<th>Measure</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>Communication (standardised measure)</td>
<td>Semi-structured conversation task: telling a story from a picture book</td>
<td>Hadwin 1996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Social Communication Questionnaire (SCQ)</td>
<td>Young 2012</td>
</tr>
<tr>
<td>Social Function</td>
<td></td>
<td>Autism Diagnostic Observation Schedule (ADOS): Language and Communication</td>
<td>Wong 2010</td>
</tr>
</tbody>
</table>
| Table 2. Outcome measures used  
(Continued) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication and Symbolic Behaviour Scales developmental profile</td>
</tr>
<tr>
<td>Autism Diagnostic Observation Schedule (ADOS): Reciprocal Social Interaction</td>
</tr>
<tr>
<td>Social Communication Questionnaire (SCQ)</td>
</tr>
<tr>
<td>Social Emotional Scale (SES), Bayley Scales of Infant Development, 3rd Edition</td>
</tr>
<tr>
<td>Vineland Adaptive Behaviour Scales (VABS), Socialisation subscale</td>
</tr>
<tr>
<td>Flexibility &amp; imagination (standardised measure)</td>
</tr>
<tr>
<td>Secondary Intervention specific: ToM</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
### Table 2. Outcome measures used (Continued)

<table>
<thead>
<tr>
<th>Intervention specific: emotion recognition</th>
<th>NEPSY-II ToM task</th>
<th>Williams 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEFA test module: computerised facial emotion recognition test</td>
<td>Bolte 2002</td>
<td></td>
</tr>
<tr>
<td>International Affective Picture System (IAPS) facial emotion recognition test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fMRI evidence of change in neural response to emotional stimuli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnostic Analysis of Nonverbal Accuracy2 (DANVA2) Faces test: emotion recognition</td>
<td>Solomon 2004</td>
<td></td>
</tr>
<tr>
<td>Baghdadli 2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading-the-Mind-in-the-Eyes task: interpreting mental states from images of eyes</td>
<td>Fisher 2005</td>
<td></td>
</tr>
<tr>
<td>Golan 2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambridge MindReading face-voice battery: computerised emotion recognition test (close generalisation)</td>
<td>Golan 2006</td>
<td></td>
</tr>
<tr>
<td>Emotion recognition from novel film clips (holistic distant generalisation)</td>
<td>Williams 2012</td>
<td></td>
</tr>
<tr>
<td>Matching familiar emotional situations to familiar facial expressions (close generalisation)</td>
<td>Golan 2010</td>
<td></td>
</tr>
<tr>
<td>Matching novel emotional situations to novel facial expressions but familiar characters (unfamiliar close generalisation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matching novel emotion situations to novel facial expressions on novel faces (distant generalisation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotion Recognition Test: photographs of faces</td>
<td>Ryan 2010</td>
<td></td>
</tr>
<tr>
<td>Emotion Vocabulary Comprehension Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition of emotional expressions from photographs</td>
<td>Hopkins 2011</td>
<td></td>
</tr>
<tr>
<td>Recognition of emotional expressions from line drawings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benton Facial Recognition Test (short form)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEPSY: Affection Recognition subtest (recognising emotions from photos of faces)</td>
<td>Williams 2012</td>
<td></td>
</tr>
<tr>
<td>Young 2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention specific: imitation</td>
<td>The Faces Task: recognising emotions from photos of faces</td>
<td>Ingersoll 2012</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>Motor Imitation Scale: performance measure of object and gesture imitation</td>
<td>Ingersoll 2012</td>
</tr>
<tr>
<td></td>
<td>Unstructured Imitation Assessment</td>
<td>Ingersoll 2012</td>
</tr>
<tr>
<td></td>
<td>Socially engaged imitation: observed during examiner/child play session</td>
<td>Landa 2011</td>
</tr>
<tr>
<td>Participant behaviour: observation</td>
<td>Social Skills Observation: two x 5 minutes, during recess or free time in school</td>
<td>Hopkins 2011</td>
</tr>
<tr>
<td></td>
<td>Joint attention and joint engagement during teacher-child or therapist-child play</td>
<td>Kaale 2012, Kim 2009</td>
</tr>
<tr>
<td></td>
<td>Structured Play Assessment</td>
<td>Goods 2013, Wong 2013</td>
</tr>
<tr>
<td></td>
<td>Symbolic Play Test</td>
<td>Wong 2010</td>
</tr>
<tr>
<td></td>
<td>Social skills during classroom observation</td>
<td>Goods 2013, Wong 2013</td>
</tr>
<tr>
<td></td>
<td>Children’s Social Behaviour Questionnaire (CSBQ): parent report</td>
<td>Begeer 2011</td>
</tr>
<tr>
<td></td>
<td>ToM Questionnaire: teacher report</td>
<td>Fisher 2005</td>
</tr>
<tr>
<td></td>
<td>Social Skills Rating System: parent report</td>
<td>Hopkins 2011</td>
</tr>
<tr>
<td></td>
<td>KidScreen, parent-report quality of life measure</td>
<td>Baghdadli 2013</td>
</tr>
<tr>
<td></td>
<td>Problem Behaviour Logs: parent report</td>
<td>Solomon 2004</td>
</tr>
</tbody>
</table>
Table 2. Outcome measures used  

<table>
<thead>
<tr>
<th>Participant behaviour: direct assessment</th>
<th>Ritvo-Freeman Real Life Rating Scale (RFRLRS): parent report</th>
<th>Wong 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PDD-BI social approach, rated by both parent and teacher</td>
<td>Kim 2009</td>
</tr>
<tr>
<td></td>
<td>Vineland Adaptive Behaviour Scales (VABS)</td>
<td>Schertz 2013</td>
</tr>
<tr>
<td></td>
<td>Participant behaviour: direct assessment</td>
<td>Reynell Developmental Language Scales (post-test; six-month follow-up, 12-month follow-up)</td>
</tr>
<tr>
<td></td>
<td>Mullen Scales of Early Learning (MSEL)</td>
<td>Landa 2011 Schertz 2013</td>
</tr>
<tr>
<td></td>
<td>Expressive vocabulary test (five-year follow-up)</td>
<td>Kasari 2006</td>
</tr>
<tr>
<td></td>
<td>Differential Abilities Scale (five-year follow-up)</td>
<td>Kasari 2006</td>
</tr>
<tr>
<td></td>
<td>Test of Problem Solving (executive function)</td>
<td>Solomon 2004</td>
</tr>
<tr>
<td>Acceptibility of Intervention</td>
<td>Parent Adherence to Treatment &amp; Competence: parent report Caregiver Quality of Involvement Scale: observational measure during parent-child play</td>
<td>Kasari 2010</td>
</tr>
<tr>
<td></td>
<td>Teacher acceptability of intervention report</td>
<td>Wong 2013</td>
</tr>
<tr>
<td></td>
<td>Children's Depression Inventory: self-report Beck Depression Inventory (BDI) (to assess parent depression) self-report</td>
<td>Solomon 2004</td>
</tr>
<tr>
<td>Economic data</td>
<td>Parenting Stress Index short form</td>
<td>Wong 2010</td>
</tr>
</tbody>
</table>

Unlike the 'Summary of findings' table and the discussion of intervention effects in the main text, the principal organising element for this table is the methodology of each outcome measure. This underscores the great difficulty in comparing findings across studies due to wide variety in assessment scoring systems.
APPENDICES

Appendix 1. Search strategies

CENTRAL 2013, Issue 7, last searched on 6 August 2013 (35 records)
Previous searches
CENTRAL 2012 (6), 10 July 2012 (126 records)
CENTRAL 2010 (3), 2 July 2010 (939 records)
#1 MeSH descriptor Child Development Disorders, Pervasive explore all trees
#2 autis*
#3 asperger*
#4 kanner*
#5 childhood schizophrenia
#6 pervasive developmental disorder*
#7 "PDD"
#8 language near/3 delay*
#9 speech near/3 disorder*
#10 (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9)

Ovid MEDLINE(R), 1946 to July Week 4 2013, last searched on 8 August 2013 (701 records)
Previous searches:
Ovid MEDLINE(R), 1946 to June Week 4 2012, 9 July 2012 (962 records)
Ovid MEDLINE(R) 1950 to July Week 1 2010, 19 July 2010 (4553 records)
1 exp child development disorders, pervasive/
2 autis$.tw.
3 pervasive developmental disorder$.tw.
4 pdd.tw.
5 (language adj3 delay$).tw.
6 (speech adj3 disorder$).tw.
7 childhood schizophrenia.tw.
8 kanner$.tw.
9 asperger$.tw.
10 or/1-9
11 randomized controlled trial.pt.
12 controlled clinical trial.pt.
13 randomi#ed.ab.
14 placebo$.ab.
15 drug therapy.fs.
16 randomly.ab.
17 trial.ab.
18 groups.ab.
19 or/11-18
20 exp animals/ not humans.sh.
21 19 not 20
22 10 and 21

EMBASE (Ovid), 1980 to 2013 Week 31, last searched 6 August 2013 (992 records)
Previous searches:
Embase, 1980 to 2012 Week 27, 9 July 2012 (1979 records)
Embase, 1980 to 2010 Week 28, 19 July 2010 (3922 records)
1 exp autism/
2 autis$.tw.
3 asperger$.tw.
Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)

CINAHLP (EBSCOhost) last searched 6 August 2013 (214 records)

Previous searches:
CINAHL 1937 to current, 9 July 2012 (211 records)
CINAHL 1937 to current, 23 July 2010 (847 records)
S26 S15 and 25
S25 S16 or S17 or S18 or S19 or S20 or S21 or S22 or S23 or S24
S24 TI(speech N3 disorder*) OR AB(speech N3 disorder
S23 TI(language N3 delay*) OR AB(language N3 delay*)
S22 TI(childhood schizophrenia*) OR AB(childhood schizophrenia)
S21 TI(kanner*) or AB(kanner*)
S20 TI(asperger*) or AB(asperger*)
S19 TI(autis*) or AB(autis
S18 TI("PDD") or AB("PDD")
S17 TI(pervasive developmental disorder*) or AB(pervasive developmental disorder*)
S16 (MH "Child Development Disorders, Pervasive+")
S15 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9 or S10 or S11 or
S12 or S13 or S14
S14 TI(effectiv* study or effectiv* research) or AB(effectiv* study or effectiv* research)
S13 “cross over”
S12 crossover*
S11 (MH "Crossover Design")
S10 (tripl* N3 mask*) or (tripl* N3 blind*)
S9 (trebl* N3 mask*) or (trebl* N3 blind*)
S8 (doubl* N3 mask*) or (doubl* N3 blind*)
S7 (singl* N3 mask*) or (singl* N3 blind
S6 (clinic* N3 trial*) or (control* N3 trial*)
S5 (random* N3 allocat*) or (random* N3 assign*)
S4 randomis* or randomiz*
S3 (MH "Meta Analysis")
S2 (MH "Clinical Trials")
S1 MH random assignment

PsycINFO (Ovid) 1806 to July Week 5 2013 , last searched 6 August 2013 (284 records)
Previous searches:
PsycINFO 1806 to July Week 1 2012, searched 10 July 2012 (1361 records)
PsycINFO searched 2010 via EBSCOhost
1 autism/ or pervasive developmental disorders/ or aspergers syndrome/
2 Autistic Thinking/
3 pervasive developmental disorder$.tw.
4 "pdd".tw.
5 autis$.tw.
6 asperger$.tw.
7 kanner$.tw.
8 childhood schizophrenia$.tw.
9 (language adj3 delay$).tw.
10 (speech adj3 disorder$).tw.
11 or/1-10
12 clinical trials/
13 (randomis* or randomiz*).tw.
14 (random$ adj3 (allocat$ or assign$)).tw.
15 ((clinic$ or control$) adj trial$).tw.
16 ((singl$ or doubl$ or trebl$ or tripl$) adj3 (blind$ or mask$)).tw.
17 (crossover$ or "cross over$").tw.
18 random sampling/
19 Experiment Controls/
20 Placebo/
21 placebo$.tw.
22 exp program evaluation/
23 treatment effectiveness evaluation/
24 ((effectiveness or evaluat$) adj3 (stud$ or research$)).tw.
25 or/12-24
26 11 and 25

PsycINFO (EBSCOhost), searched 23 July 2010 (891 records)
S29 S17 and S28
S28 S18 or S19 or S20 or S21 or S22 or S23 or S24 or S25 or S26 or S27
S27 TI(speech N3 disorder*) OR AB(speech N3 disorder*)
S26 TI(languag$ N3 delay$) OR AB(languag$ N3 delay)
S25 TI(childhood schizophrenia*) or AB(childhood schizophrenia*)
S24 TI(kanner*) or AB(kanner*)
S23 TI(asperger*) or AB(asperger*)
S22 TI(autis*) or AB(autis*)
S21 TI("PDD") or AB("PDD")
S20 TI (pervasive developmental disorder*) or AB (pervasive developmental disorder*)
Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD) (Review)

Copyright © 2014 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.
schizophren*”) AND (SU.EXACT(“Experimental Groups”) OR SU.EXACT(“Control Groups”) OR SU.EXACT(“Followup Studies”) OR SU.EXACT(“Longitudinal Studies”) OR SU.EXACT(“Program Effectiveness”) OR ((prospective* OR “follow up” OR evaluat* OR compar* OR blind*) NEAR/3 stud*) OR (evaluat* NEAR/3 research*) OR ((compar* OR control*) NEAR/3 group*) OR random* OR intervention* OR experiment* OR trial*)

**ERIC (Dialog Datastar)** 1966 to current, searched 21 July 2010 (2982 records)
1.SEARCH:Pervasive-Developmental-Disorders$.DE.
2.SEARCH:pervasive ADJ developmental ADJ disorder$
3.SEARCH:pdd
4.SEARCH:speech NEAR disorder$
5.SEARCH:language NEAR delay$
6.SEARCH:autis$
7.SEARCH:asperger$
8.SEARCH:kanner$
9.SEARCH:childhood ADJ schizophrenia
10.SEARCH:1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9
11.SEARCH:Control-Groups.DE.
12.SEARCH:EXPERIMENTAL-GROUPS.DE.
13.SEARCH:LONGITUDINAL-STUDIES.DE.
14.SEARCH:FOLLOW-UP-STUDIES.DE.
15.SEARCH:PROGRAM-EFFECTIVENESS.DE.
16.SEARCH:((prospective$ OR follow ADJ up OR evaluat$ OR compar$ OR blind$) NEAR study).TI,AB.
17.SEARCH:(evaluat$ NEAR research$).TI,AB.
18.SEARCH:((compar$ OR control$) NEAR group$).TI,AB.
19.SEARCH:random$.TI,AB.
20.SEARCH:intervention$.TI,AB.
21.SEARCH:experiment$.TI,AB.
22.SEARCH:trial$.TI,AB.
23.SEARCH:11 OR 12 OR 13 OR 14 OR 15 OR 16 OR 17 OR 18 OR 19 OR 20 OR 21 OR 22
24.SEARCH:10 AND 23

**AutismData** (http://www.autism.org.uk/autismdata/) last searched 7 August 2013 (59 records)
Search in 2013 using the phrase “theory of mind”
Previous searches:
10 July 2012 (151 records)
27 July 2010 (29 records)
Search in 2010 and 2012 using the keywords random* or RCT or control* or trial*

**ICTRP** searched 7 August 2013 (85 records)
Previous searches:
10 July 2012 (234 records)
Simple search using terms autis* OR asperg* OR pervasive developmental disorder*

**metaRegister of Controlled Trials (mRCT)** searched 7 August 2013 (19 records)
Search in 2013 using the phrase “theory of mind”
Previous searches:
25 July 2010 (297 records)
autis* OR asperg* OR pervasive developmental disorder*
CONTRIBUTIONS OF AUTHORS

<table>
<thead>
<tr>
<th>Task</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draft the protocol</td>
<td>S Fletcher-Watson</td>
</tr>
<tr>
<td>Develop a search strategy</td>
<td>S Fletcher-Watson</td>
</tr>
<tr>
<td>Select which trials to include (2 people + 1 arbiter in the event of dispute)</td>
<td>S Fletcher-Watson, E Manola, F McConnell, Arbiter: H McConachie</td>
</tr>
<tr>
<td>Extract data from trials (2 people)</td>
<td>S Fletcher-Watson, E Manola, F McConnell</td>
</tr>
<tr>
<td>Enter data into RevMan (Cochrane software)</td>
<td>S Fletcher-Watson</td>
</tr>
<tr>
<td>Carry out the analysis</td>
<td>S Fletcher-Watson</td>
</tr>
<tr>
<td>Interpret the analysis</td>
<td>S Fletcher-Watson, H McConachie</td>
</tr>
<tr>
<td>Draft the final review</td>
<td>S Fletcher-Watson, H McConachie</td>
</tr>
<tr>
<td>Keep the review up to date</td>
<td>S Fletcher-Watson</td>
</tr>
</tbody>
</table>

DECLARATIONS OF INTEREST

A £1000 grant was provided by the charity Research Autism, to pay for some research assistance.

Sue Fletcher-Watson and Helen McConachie - both of these authors have been involved in the development of a therapeutic iPad app for preschoolers with autism, targeting skills relevant to the current review. An RCT of this app has recently been completed and will be published in due course. This study may be deemed eligible for inclusion in a future version of this review. The app is now available on the commercial market and the authors could both receive royalties from sales of the app if these exceed a certain threshold.

Fiona McConnell - none known.

Eirini Manola - none known.

SOURCES OF SUPPORT

Internal sources
- No sources of support supplied
External sources

- Research Autism, UK.
Funding for training and to employ a research associate

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

Influenced by changes to diagnostic criteria in DSM-5 (APA 2013), and the direction of theory which suggests Theory of Mind (ToM) primarily underlies social and communication impairments in autism, we removed the domain of FLEXIBILITY AND IMAGINATION from the expected Primary Outcomes.

A further criterion under “Types of Interventions” was implemented to exclude complex, broad-based interventions addressing an array of social and communication behaviours. While the authors note the efficacy and appropriateness of this approach, these interventions need to be excluded from our review because they cannot be described to relate directly to the ToM model of autism.