Supplementary Materials

Appendix A: Additional details on implementation measures

Fidelity to Intervention Delivery

The Early Start Denver Model (ESDM) fidelity scale (included in the ESDM manual by Rogers & Dawson, 2010) is a clinical tool rating 13 key therapist behaviors according to the quality and accuracy of implementation of treatment techniques. These include: management of child attention, quality of behavioral teaching (i.e., the ability to organize teaching episodes in the context of clear antecedent-behavior-consequence sequences embedded in play-routines), the accurate use of instructional techniques such as *fading*, *shaping* and *prompting*, adult ability to modulate child affect and arousal, management of unwanted behaviors using positive approaches, use of turn-taking, quality of dyadic engagement, optimizing child motivation for participation in activity, use of positive affect, sensitivity and responsivity to child communications, targeting multiple and varied communicative functions (e.g. requesting, commenting, protesting, labelling, greeting), appropriateness of adult language for child's language level, use of joint activity routines (articulated around a set-up stage, the establishment of a theme, a variation on the theme, and a clear closure), and smooth transitions between activities that maximize child interest and engagement.

Ratings for each are on a 5-point Likert-type scale, anchored such that 1 = lack of effective display of the practice, 3 = teaching behavior including strengths and also weaknesses, and 5 = best possible example of the teaching behavior. Individual ratings are then summed and converted to a proportionate level of fidelity. Through the advanced training and certification process, therapists are expected to demonstrate implementation of

the ESDM consistently exceeding >80% on the clinical fidelity tool, across multiple sessions working children with ASD.

Center staff who were ESDM Trainers or Senior Certified therapists (i.e., with ≥ 5 years post-certification experience) conducted initial skills training and maintenance workshops with new staff in each of the specialized and inclusive settings early in the school calendar year, and provided ongoing in-room coaching throughout the year. Using the ESDM fidelity tool, spot checks on staff implementation were conducted during in-room coaching sessions across each school year. Staff implementing the therapy practices below the nominal 80% threshold on the clinical fidelity tool during were provided with additional coaching and supervision.

Quality of Classroom Teaching and Learning

The Sustained Shared Thinking and Emotional Well-Being scale (SSTEW; Siraj, Kingston, & Melhuish, 2015) is a standardized classroom observation measure of the quality of pedagogical practices in early childhood education and care settings (see Howard et al., 2018 for details on its psychometric properties). Not specific to the care of children with ASD, nor to special education settings, the SSTEW includes ratings of the extent to which staff scaffold child learning through play and adult-supported activities across five dimensions:

1. Building trust, confidence and independence (e.g., encouraging choices);
2. Social and emotional well-being (e.g., supporting social interactions, and recognition and response to others' verbal and non-verbal expressions);
3. Supporting and extending language and communication (e.g., encouraging children to talk with others, listening to children and encouraging children to listen to others);

4. Supporting learning and critical thinking (e.g., supporting curiosity and problem solving, encouraging sustained, shared thinking);
5. Assessing learning and language (using assessment to support and extend learning, critical thinking and language development).

Each dimension is rated on a 7-point scale, anchored such that 1 = inadequate, 3 = minimal/adequate, 5 = good, 7 = excellent, and these are then averaged for a total SSTEW score.

SSTEW ratings for this study were made once per classroom per year, toward the end of the calendar year, by researchers who were 1) independent of the study team, 2) certified in use of the measure, and 3) kept blind to the study aims and hypotheses. The assessment of 54 Australian early childhood education classrooms – conducted independently of the current study, but during the same period of time – shows evidence of the reliability of this measure as well as its concurrent and predictive validity for child outcomes around one year later (Howard et al., 2018) and provided a benchmark (i.e., mean overall SSTEW score) against which we were able to evaluate the quality of the teaching and learning environment of the Inclusive and Specialized settings in which children in the current study were placed (see Figure 2).

Appendix B: Additional details on proximal outcome measures

Three measures of proximal child outcomes were derived/coded blind to child randomization group.

Language ENvironment Analysis (LENA) was used as an automated measure of each child's spontaneous vocalization, during unstructured 1:1 interaction with an adult. LENA is an automated speech analysis software that uses a digital language processor and audio

processing algorithms to yield an objective measure of language in the child's natural environment (Gilkerson & Richards 2008; Xu et al. 2008). It is considered particularly well suited for use with young children with ASD (Dykstra et al., 2012) given its capability to quantify dimensions of language not captured by standardized assessments, such as the frequency with which language is used (including pre-verbal vocalizations, known to be associated with later language skill; McGillion et al., 2017).

We took 45-minute LENA recordings with each child at the start and end of the intervention year, during unstructured 1:1 interactions with an adult, for a measure of the frequency of spontaneous child vocalization during situations that provided opportunities for language use but without explicit instruction for the child to speak. The 1:1 context was preferred to a group-context to ensure that the LENA recording was only capturing vocalizations of the target-child. A blinded research assistant later extracted data from 40-minute samples of each recording, beginning 5 minutes into each recording to allow time for the child to habituate to wearing the LENA device. A strong correlation between intake and exit measures of frequency of spontaneous child vocalization generated by the LENA audio processing algorithm ($r = .71$, $p < .001$) suggests good reliability of this measure.

The Modified Classroom Observation Schedule to Measure Intentional Communication (M-COSMIC; Clifford et al., 2010) is a coding scheme developed to quantify aspects of the social interaction behavior of children with ASD within their early childhood settings, with teachers and peers. We took short video samples of spontaneous child behavior within their classrooms – toward the start and end of the school year – capturing 5-minutes of free play footage and 5-minutes of semi-structured snack time footage for each child. Following Clifford et al. (2010), we had these coded off-line by blinded research assistants for two key types of social-communication behaviors –initiations and responses. Each child initiation and response toward another – whether a peer or an adult, and

whether signaled vocally/verbally, through gesture or other intentionally communicative means – was coded, and then we summed these for a total count of intentional communication acts within the classroom across the 10-minute sample. Initiations were coded according to the following procedures: *Code initiation when the child spontaneously initiates an interaction. Initiation should not be coded when the communication partner clearly prompts the interaction verbally, physically, or otherwise. Also code initiation when the child's response is a clear elaboration, contradiction or correction to the communication partner – e.g. communicative says, "There's your coat" and the child responds, "That's not my coat: this is my coat" (pointing to a different coat).* Responses were coded according the following procedures: *Code response when a child responds to an instruction, prompt, question, suggestion, or action (e.g. the child sits after being told to "sit down") of another. This code should be used even if the content of the child's response is incorrect (e.g., during a puzzle the teacher instructs the child to find the blue piece, but the child picks the red piece); or non-complaint (e.g., child says 'No' and slumps in chair).* A proportion of tapes (30%) was double-coded, with excellent inter-rater agreement observed on the total communication score across the combined 10-minute samples (ICC = .93).

Finally, following an experimental paradigm detailed by Vivanti et al. (2016a; Experiment 3), we measured children's propensity for spontaneous imitation by showing each child a series of eight, 10-second videos during which a demonstrator performs a two-step action on one of eight available objects. Identical objects are available to the child who has the opportunity to imitate what s/he sees, but is given no explicit instruction to do so. We filmed child behavior during the task and had blinded research assistants code this footage off-line, giving 2 points for any trial during which a child *imitated* the demonstrator's action, 1 point per trial where a child performed *any action* on the same object used by the demonstrator, and 0 points for *any other response* (i.e., including picking up objects not used

by the demonstrator). Summing across points per trial, provided a total spontaneous imitation score – possible range 0 to 16 points – for each child. A proportion of tapes (20%) was double-coded, with excellent inter-rater agreement evident (ICC = .90).