APPENDIX A: SEARCH STRATEGY

#1 PsycINFO (ProQuest search engine)

(cabs(TEACCH)) OR (cabs("Treatment and Education of Autistic"))

#2 Medline (PubMed search engine)

TEACCH[TIAB] OR 'Treatment and Education of Autistic'[TIAB]

#3 Cochrane Central Register of Controlled Trials

TEACCH[TIAB] OR 'Treatment and Education of Autistic'[TIAB]

Notes. Search date: December 1, 2012. We did not use restrictions by publication year, publication type, population, and language of publication. The following hierarchical order was used for the elimination of duplicates and for the assignment of distinct references to each database: (1) PsycINFO, (2) Medline, and (3) Cochrane Central Register of Controlled Trials.

APPENDIX B: QUALITY ASSESSMENT

Methodological quality was assessed by means of the Downs and Black checklist for randomised and non-randomised studies of heath care interventions (Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions J Epidemiol Community Health 1998;52;377-384). Following the original authors' guidelines, the checklist was adapted to the field of applied behavior analysis for autism adding a list of confounders, adverse effects, and ranges for power assessment. New additions, specifications, and comments to the original checklist are typed in italic.

Reporting

1. Is the hypothesis/aim/objective of the study clearly described?

Yes (1 point), No (0 points)

2. Are the main outcomes to be measured clearly described in the Introduction or Methods section? If the main outcomes are first mentioned in the Results section, the question should be answered no.

Yes (1 point), No (0 points)

3. Are the characteristics of the patients included in the study clearly described? In cohort, within subject studies and trials, inclusion and/or exclusion criteria should be given. In case-control studies, a case-definition and the source for controls should be given. *(Inclusion criteria or at least one other relevant feature apart from sex and age).*

Yes (1 point), No (0 points)

4. Are the interventions of interest clearly described? Treatments and placebo*/control* (where relevant) that are to be compared should be clearly described.

Yes (1 point), No (0 points)

5. Are the distributions of principal confounders in each group of subjects or treatment condition to be compared clearly described? *At least one of the following are described apart from sex and age: past and concurrent interventions, intervention intensity (hours per week), diagnoses, severity of existing illness, intellectual quotient at pretest, treatment fidelity indexes, other relevant confounder. If only sex and/or age are described, answer 1.*

Yes (2 points), Partially (1 point), No (0 points)

6. Are the main findings of the study clearly described? Simple outcome data should be reported for all major findings so that the reader can check the major analyses and conclusions *(provide means or data from all participants).* (This question does not cover statistical tests, which are considered below).

Yes (1 point), No (0 points)

7. Does the study provide estimates of the random variability in the data for the main outcomes? In non normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes (1 point), No (0 points)

8. Have all important adverse events that may be a consequence of the intervention been reported? This should be answered yes if the study demonstrates that there was a comprehensive attempt to measure adverse events *or at least to prevent them on the basis of specific exclusion and inclusion criteria.*

When no detail was provided about the treatment group, No

Yes (1 point), No (0 points)

9. Have the characteristics of patients lost to follow-up been described? This should be answered yes where there were no losses to follow-up or where losses to follow-up were so small that findings would be unaffected by their inclusion *(-10%).* This should be answered no where a study does not report the number of patients lost to follow-up. *Question should be answered 'no' in retrospective studies.*

   Yes (1 point), No (0 points)

10. Have actual probability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?

    Yes (1 point), No (0 points)

External validity

All the following criteria attempt to address the representativeness of the findings of the study and whether they may be generalised to the population from which the study subjects were derived.

11. Were the subjects asked to participate in the study representative of the entire population from which they were recruited? The study must identify the source population for patients and describe how the patients were selected. Patients would be representative if they comprised the entire source population, an unselected sample of consecutive patients, or a random sample. Random sampling is only feasible where a list of all members of the relevant population exists. *Where a study does not report the proportion of the source population from which the patients were selected, the question should be answered as unable to determine.*

    Yes (1 point), Unable to determine (0 points), No (0 points)

12. Were those subjects who were prepared to participate representative of the entire population from which they were recruited? The proportion of those asked who agreed should be stated. Validation that the sample was representative would include demonstrating that the distribution of the main confounding factors was the same in the study sample and the source population.

    Yes (1 point), Unable to determine (0 points), No (0 points)

13. Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive? For the question to be answered yes the study should demonstrate that the intervention was representative of that in use in the source population. The question should be answered no if, for example, the intervention was undertaken in a specialist centre unrepresentative of the hospitals most of the source population would attend. *For interventions that took place at the participants' home question should be answered yes.*

    Yes (1 point), Unable to determine (0 points), No (0 points)

Internal validity - bias

14. Was an attempt made to blind study subjects to the intervention they have received? For studies where the patients would have no way of knowing which intervention they received, this should be answered yes. *It can we assumed that children with developmental disabilities that participated in most studies under scrutiny were unable to describe the type of intervention they were receiving compared to others.*

    Yes (1 point), Unable to determine (0 points), No (0 points)

15. Was an attempt made to blind those measuring the main outcomes of the intervention? In cases were outcome variables were collected through self-administered questionnaires answer should be answered yes. *Question should be*

answer yes if those measuring the main outcomes were whether blind to group status or were independent of treatment delivery.

Yes (1 point), Unable to determine (0 points), No (0 points)

16. If any of the results of the study were based on "data dredging", was this made clear? Any analyses that had not been planned at the outset of the study should be clearly indicated. If no retrospective unplanned subgroup analyses were reported, then answer yes. *Question should be answered 'no' for retrospective studies were cases were not admitted consecutively or were selected in anyway. Note: data dredging is the inappropriate (sometimes deliberately so) search for 'statistically significant' relationships in large quantities of data in spite of previous hypotheses.*

Yes (1 point), Unable to determine (0 points), No (0 points)

17. In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control *and within-subjects* studies, is the time period between the intervention and outcome the same for cases and controls? Where follow-up was the same for all study patients the answer should yes. If different lengths of follow-up were adjusted for by, for example, survival analysis the answer should be yes. Studies where differences in follow-up are ignored should be answered no.

Yes (1 point), Unable to determine (0 points), No (0 points)

18. Were the statistical tests used to assess the main outcomes appropriate? The statistical techniques used must be appropriate to the data. For example nonparametric methods should be used for small sample sizes. Where little statistical analysis has been undertaken but where there is no evidence of bias, the question should be answered yes. If the distribution of the data (normal or not) is

not described it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes (1 point), Unable to determine (0 points), No (0 points)

19. Was compliance with the intervention/s reliable? Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes. *In no measure for treatment fidelity assurance were taken, question should be answered no.*

Yes (1 point), Unable to determine (0 points), No (0 points)

20. Were the main outcome measures used accurate (valid and reliable)? For studies where the outcome measures are clearly described, the question should be answered yes *(e.g., systematic behavioural observation with inter-rater reliability information)*. For studies which refer to other work or that demonstrate the outcome measures are accurate *(e.g., validated psychometric tests)*, the question should be answered as yes.

Yes (1 point), Unable to determine (0 points), No (0 points)

Internal validity - confounding (selection bias)

21. Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies, *or all participants within-subjects designs*) recruited from the same population? For example, patients for all comparison groups should be selected from the same hospital. The question should be answered unable to determine for cohort and case control studies where there is no information concerning the source of patients included in the study.

Yes (1 point), Unable to determine (0 points), No (0 points)

22. Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) *or participants in within-subjects studies* recruited over the same period of time? For a study, which does not specify the time period over which patients were recruited, the question should be answered as unable to determine. *For studies that recruited participants during a time expand longer than one year answer should be answer no.*

    Yes (1 point), Unable to determine (0 points), No (0 points)

23. Were study subjects randomised to intervention groups? Studies that state that subjects were randomised should be answered yes except where method of randomisation would not ensure random allocation. For example alternate allocation would score no because it is predictable. *In studies in which allocation was conducted by means of an unpredictable variable unrelated to subjects characteristics (e.g., staff availability) answer yes. Question will be answered yes if subjects in intervention and comparison groups were matched at least in pre-IQ, sex or age. All non-randomised controlled studies should be answered no. For within-subject studies this item is "not applicable." Therefore it should be ignored and the total score be prorated.*

    Yes (1 point), Unable to determine (0 points), No (0 points)

24. Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable? All non-randomised *controlled* studies should be answered no. If assignment was concealed from patients but not from staff, it should be answered no. *For within-subject studies this item is "not applicable." Therefore it should be ignored and the total score be prorated.*

    Yes (1 point), Unable to determine (0 points), No (0 points)

25. Was there adequate adjustment for confounding in the analyses from which the main findings were drawn? This question should be answered no for trials if: the main conclusions of the study were based on analyses of treatment rather than intention to treat; the distribution of known confounders in the different treatment groups was not described; or the distribution of known confounders differed between the treatment groups but was not taken into account in the analyses. In nonrandomised studies if the effect of the main confounders was not investigated or confounding was demonstrated but no adjustment was made in the final analyses the question should be answered as no.

    Yes (1 point), Unable to determine (0 points), No (0 points)

26. Were losses of patients to follow-up taken into account? If the numbers of patients lost to follow-up are not reported, the question should be answered as unable to determine. If the proportion lost to follow-up was too small to affect the main findings, the question should be answered yes. *(<10%). For retrospective studies question should be answered 'no.'*

    Yes (1 point), Unable to determine (0 points), No (0 points)

Power

27. Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%?

    < 10 (0 points), 10 – 19 (1 point), 20 – 29 (2 points), 30 – 39 (3 points), 40 – 49 (4 points), ≥ 50 (5 points)

Table A.

Quality assessment of TEACCH intervention studies for autism according to Downs and Black checklist (results by item).

	Reporting										External validity			Internal validity–bias							Internal validity–confounding						Power
First author, year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14†**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
Aoyama (1995)	1	1	1	1	0	1	1	0	1	0	0	0	1	1	0	1	1	1	0	0	1	1	–	–	0	0	0
Braiden et al (2012)	1	0	1	1	2	1	1	0	1	0	0	0	1	1	0	1	1	1	0	0	1	0	–	–	0	1	1
McConkey et al. (2010)	1	1	1	1	2	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0	1	3
Ozonoff & Cathcart (1998)	1	0	1	1	2	1	1	0	1	0	0	0	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1
Panerai et al. (2002)	1	1	1	0	2	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0
Panerai et al. (2009)	1	1	1	1	2	1	1	0	1	1	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0	1	1
Persson (2000)	0	1	1	0	2	1	1	0	1	1	0	0	1	1	0	1	1	1	0	1	1	1	–	–	0	1	0
Probst & Leppert (2008)	1	1	1	1	2	0	1	0	1	0	0	0	1	1	0	1	1	1	0	0	1	1	–	–	0	1	0
Siaperas & Brown (2006)	1	0	1	1	2	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	1	1	–	–	0	1	1
Siaperas et al. (2007)	1	1	1	1	2	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	1	1	–	–	0	1	0
Tsang et al. (2007)	1	1	1	0	2	1	1	0	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1
Van Bourgondien et al. (2003)	1	1	1	1	2	0	1	0	1	1	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	0	0
Welterlin et al. (2012)	1	1	1	1	2	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	1	2

Items 23 and 24 were only applicable to controlled studies.

Table B.

Quality assessment of TEACCH intervention studies for autism according to Downs and Black checklist (results by domain).

First author, year	Reporting	External validity	Internal validity		Power	Total (%)
			Bias	Confounding		
Aoyama (1995)	7 (0.6)	1 (0.3)	4 (0.6)	2 (0.5)	0 (0.0)	14 (44)
Braiden et al (2012)	8 (0.7)	1 (0.3)	4 (0.6	2 (0.5)	1 (0.2)	16 (50)
McConkey et al. (2010)	10 (0.9)	1 (0.3)	5 (0.7)	2 (0.3)	3 (0.4)	21 (70)
Ozonoff & Cathcart (1998)	8 (0.7)	1 (0.3)	5 (0.7)	3 (0.5)	1 (0.6)	18 (60)
Panerai et al. (2002)	10 (0.9)	1 (0.3)	6 (0.9)	5 (0.8)	0 (0.0)	22 (73)
Panerai et al. (2009)	10 (0.9)	1 (0.3)	5 (0.7)	2 (0.3)	1 (0.4)	19 (63)
Persson (2000)	8 (0.7)	1 (0.3)	5 (0.7)	3 (0.8)	0 (0.2)	17 (53)
Probst & Leppert (2008)	8 (0.7)	1 (0.3)	4 (0.6)	3 (0.8)	0 (0.2)	16 (50)
Siaperas & Brown (2006)	9 (0.8)	3 (0.3)	4 (0.6)	3 (0.8)	1 (0.0)	20 (63)
Siaperas et al. (2007)	10 (0.9)	3 (1.0)	4 (0.6)	3 (0.8)	0 (0.4)	20 (63)
Tsang et al. (2007)	8 (0.7)	3 (1.0)	5 (0.7)	5 (0.8)	1 (0.2)	22 (73)
Van Bourgondien et al. (2003)	9 (0.8)	0 (0.0)	3 (0.4)	2 (0.3)	0 (0.0)	14 (47)
Welterlin et al. (2012)	10 (0.9)	1 (0.3)	7 (1.0)	3 (0.5)	2 (0.4)	23 (77)

Score and proportion of the maximum achievable score by domain. Total expressed in raw scores and %.

APPENDIX C: POWER ANALYSIS

Full report of prospective power calculation of random-effects meta-analysis

Based on a preliminary review of study abstracts, we expected to have approximately 5 studies meeting the pre-specified exclusion criteria for each one of the selected homogeneous outcomes. Most studies were controlled trials and included similar number of participants in intervention and control groups. The overall number of participants varied substantially across studies with a typical sample size of about 30 participants. We assumed a high common correlation coefficient of 0.80 between pre-test and post-test measurements for all the assessed outcomes. The effect size estimate and variance in each study were assumed to be based on the difference in mean change from pre-test to post-test between intervention and control groups, divided by the pooled pre-test standard deviation (Morris, 2008).

Based on this preliminary information, we conducted an a priori power analysis of the random-effects meta-analysis according to the methods by Hedges and Pigott (2001) assuming different scenarios for the underlying effect size distribution, including (a) a small or moderate mean effect size of 0.30 or 0.50, respectively, and (b) a small, moderate, or large degree of between-study heterogeneity corresponding to an I^2 statistic of 25, 50, or 75%, respectively (Higgins & Thompson, 2002). All tests were two-tailed with a significance level of 0.05.

The power to detect as statistically significant a moderate underlying mean effect size of 0.50 was 0.98 for a small amount of between-study heterogeneity of I^2 = 25%, 0.90 for a moderate amount of heterogeneity of I^2 = 50%, and 0.63 for a large amount of heterogeneity of I^2 = 75%. Similarly, the power to detect a small mean effect size of 0.30

was 0.69 for a small amount of heterogeneity of I^2 = 25%, 0.52 for a moderate amount of heterogeneity of I^2 = 50%, and 0.29 for a large amount of heterogeneity of I^2 = 75%. Thus, combining 5 studies with a typical sample size of 30, we anticipated high power for a moderate effect size with small or moderate heterogeneity and reasonable power for a moderate effect size with large heterogeneity or a small effect size with small heterogeneity, but little power for a small effect size with moderate or large heterogeneity.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11,* 364-386.

APPENDIX D: ISOLATED OUTCOMES INCLUDED IN THE MEAN EFFECT SIZES META-ANALYSIS

Studies included in the meta-analysis that only reported isolated/unmatched outcomes

- *Aoyama (1995).* This author examined the effects of an environmental manipulation inspired by the TEACCH notion of structured teaching on work efficiency (i.e., time to complete an arbitrary work task).

- *Van Bourgoudien et al. (2003).* These authors assessed practitioners' perceptions and program changes prompted by the introduction of TEACCH components into a residential program. We included a range of aspects of the school environment of children as perceived by teachers based on the subscales of the Environmental Rating Scale (Communication, Structure, Socialization, Developmental, Behavior, and Total).

Studies included in the meta-analysis that reported some isolated outcomes

- *McConkey et al. (2010).* We extracted the isolated outcomes: (a) autism quotient and autism percentile evaluated by the Gilliam Autism Rating Scale, and (b) number of parent-reported problems and number of parent-reported problems that are getting better.

- *Persson (2000).* We extracted the isolated outcomes: (a) vocational behavior, (b) vocational skills, and (c) leisure skills. All these measured with the AA-PEP.

- *Siaperas et al. (2006).* We have added a series of observational outcomes including: no activity, other task, no social act, mean of activity, walking, practical task, work, leisure, unclear social act, addressing the observer. All these outcomes were measures as duration of engagement.

- *Siaperas et al. (2007).* We have added a series of observational outcomes including: self-stimulation, self-injury, aggression, and damage to property. All these outcomes were measures as the frequency of responses per observation session.

- *Tsang (2007).* We have added the developmental score of the Chinese version of the PEP-R.

- *Welterlin et al. (2012).* We have added language expression and language comprehension (Scales of Independent Behavior-Revised).