Research in Autism Spectrum Disorders 59 (2019) 10–21

Contents lists available at ScienceDirect

Research in Autism Spectrum Disorders

journal homepage: www.elsevier.com/locate/rasd

Systematic review of data analyses and reporting in group-based social skills intervention RCTs for youth with ASD

Christopher Lopata*, James P. Donnelly, Jonathan D. Rodgers, Marcus L. Thomeer

Institute for Autism Research, Canisius College, 2001 Main Street, Science Hall 1016, Buffalo, NY, 14208, USA

ARTICLE INFO

Keywords:
Group-based social skills interventions
Youth with ASD
Systematic review

ABSTRACT

Background: Group-based social skills interventions (GSSIs) are commonly used to address the social impairments of youth with ASD. However, the administration of treatments in group formats (i.e., clusters) poses several methodological challenges including accounting for cluster effects. The most recent and comprehensive meta-analysis of RCTs of GSSIs for youth with ASD yielded an overall medium effect ($g = 0.51$; Gates et al., 2017). This suggested a positive effect; however, little is known about the extent to which the studies adhered to standards for conducting and reporting RCTs including standards related to group-based interventions.

Method: The current review assessed the extent to which the study planning, data assessment, and data analytic procedures used in the RCTs ($N = 18$) included in the meta-analysis adhered to established standards for RCTs.

Results: Results were consistent across the three areas assessed and suggested an overall adherence rate of 42% (range 41%–43%). Significant variability was found within each of the three areas and suggested that few facets of the standards were met by a majority of studies. The statistically-oriented aspects were most neglected. None of the studies accounted for the group-based (clustered) design and delivery of the intervention which can negatively impact power, effect size, and precision estimates and overestimate intervention effects. Year of article publication and journal impact factor were predominantly unrelated to adherence rates.

Conclusions: Increased familiarization with standards for RCTs appears necessary to improve the practices of researchers, along with increased requirements for adherence by journal editors and reviewers.

1. Introduction

Social impairment is the most prominent and debilitating feature of youth with autism spectrum disorder (ASD), including those without concomitant intellectual disability (ID; American Psychiatric Association, 2013). This group is of particular concern as the most recent Centers for Disease Control and Prevention prevalence estimates indicated that youth with ASD without ID account for more than two-thirds of those diagnosed (Christensen et al., 2016). For the purposes of this study, the term ASD will be used to describe youth with ASD without ID. The chronic and long-term nature of the social impairments of these youth have been associated with a number of poor life outcomes (e.g., social isolation, family and financial dependency, mental health problems; Portway & Johnson, 2005; Shattuck, Wagner, Narendorf, Sterzing, & Hensley, 2011), leading to widespread recognition of the need for effective

* Corresponding author at: Institute for Autism Research, Canisius College, Science Hall 1016C, 2001 Main Street, Buffalo, NY, 14208, USA.
E-mail addresses: lopatac@canisius.edu (C. Lopata), donnell6@canisius.edu (J.P. Donnelly), rodgers1@canisius.edu (J.D. Rodgers), thomeerm@canisius.edu (M.L. Thomeer).

https://doi.org/10.1016/j.rasd.2018.11.008
Received 4 May 2018; Received in revised form 27 October 2018; Accepted 12 November 2018
Available online 14 December 2018
1750-9467/

treatments. However, development of effective treatments is a challenge given the complex nature of the social impairments. The social deficits of youth with ASD consist of impairments in basic social behaviors (e.g., social approach and response behaviors) and more complex social-cognitive skills (e.g., decoding and understanding facial expressions, vocal expressions, and another's perspective; Bellini, Gardner, & Markoff, 2014); cumulatively, these affect the manner in which these youth approach, understand, and respond to others (Scarpa, Reyes, & Attwood, 2013).

The most common psychosocial treatment used to address the social impairments of youth with ASD is group-based social skills interventions (GSSIs; Gates, Kang, & Lerner, 2017). These generally employ cognitive and behavioral techniques to address the core impairments in social cognition/understanding and skills/behaviors (Kaat & Lecavalier, 2014; Scarpa, White, & Attwood, 2013). Common elements of GSSIs include explicit instruction, modeling, role-play/rehearsal, repeated practice, and performance feedback/reinforcement (McMahon, Lerner, & Britton, 2013; Reichow, Steiner, & Volkmar, 2012). The need to learn, develop, and practice social interaction skills makes group delivery particularly appealing. In fact, the opportunity to learn and rehearse social skills, and receive performance feedback within the context of a peer group is considered an important therapeutic element (Kasari et al., 2016).

Although GSSIs are a common and highly applicable psychosocial intervention for youth with ASD, conclusions about their efficacy have been hindered primarily due to limited testing in rigorously-conducted randomized controlled trials (RCTs). Appropriately designed, executed, and reported RCTs are considered the gold standard for testing the efficacy of treatments (Schulz, Altman, & Moher, 2010). While two previous meta-analyses both reflect limited testing in RCTs, the more recent meta-analysis reflects an increase in the number of RCTs. In the earlier meta-analysis, Reichow et al. (2012) identified only five RCTs of GSSIs for youth with ASD; these yielded a moderate overall effect for improvement of social competence ($g = 0.47$). In the most recent and largest meta-analysis to date (completed per PRISMA guidelines), Gates et al. (2017) identified 18 RCTs of GSSIs for youth with ASD; more than triple the number of studies in the earlier meta-analysis. Results yielded an overall medium effect ($g = 0.51$) supporting the efficacy of GSSIs in improving social competence. Despite the consistency in overall magnitudes of effect, the authors of both meta-analyses noted significant variability in outcomes within and across studies.

Although delivery of social skills interventions in group formats offers a number of potential therapeutic benefits for youth with ASD (e.g., instruction in core skills, in vivo rehearsal, adult and peer feedback, peer relationship development, etc.), group delivery can pose methodological and data analytic challenges in RCTs. One important consideration is the potential influence of individual groups (i.e., clusters) on efficacy conclusions and estimates of effect sizes and precision (confidence intervals; Boutron, Moher, Altman, Schulz, & Ravaud, 2008). This is especially important for group-delivered psychosocial treatments because these are often difficult to manualize and deliver accurately across groups (clusters), and variability in treatment fidelity (implementation accuracy) and provider competency/expertise can affect outcomes. In addition, because multiple subjects are treated by the same clinician(s) at the same time the assumption of independence of each subject cannot be made (Boutron et al., 2008). At present, there is a paucity of research examining how factors such as clustering are addressed and analyses are conducted and reported for GSSIs for youth with ASD.

To foster clear, transparent, and thorough reporting in RCTs, guidelines have been developed for researchers, reviewers, and journal editors by several professional groups/associations such as the Consolidated Standards for Reporting of Trials (CONSORT; Schultz et al., 2010), American Psychological Association's Journal Article Reporting Standards (JARS; Appelbaum et al., 2018; Appelbaum, Cooper, Maxwell, Stone, & Sher, 2008), American Educational Research Association's (AERA's) Standards for Reporting on Empirical Social Science Research (2006), What Works Clearinghouse (WWC) Standards Handbook, Version 4.0 (U. S. Department of Education, Institute of Education Sciences, 2017), and American Statistical Association's (ASA's) statement on *p*-values (Wasserstein & Lazar, 2016). These standards were developed based upon the contributions of experts in clinical trials and methods, statistics, and epidemiology, and they are applicable to both non-pharmacological (e.g., social, behavioral, and educational) and medical trials (Appelbaum et al., 2018; Boutron et al., 2008; Schulz, Altman, & Moher, 2010). Many elements of the standards and guidelines have undergone revisions to accommodate advances in design, measurement, and data analytic approaches, including several extensions and/or modules for different types of designs. There is significant overlap in the content and items of the standards, reflecting substantial agreement around many methodological features. Although the standards do not recommend specific designs or analytic methods, they indirectly do so by setting standards that must be met to publish in many top journals (Schulz et al., 2010). Because of this, they can help researchers to rigorously design and execute studies, analyze data, and report findings (Appelbaum et al., 2018; Campbell, Piaggio, Elbourne, & Altman, 2012; Schulz et al., 2010; U. S. Department of Education, Institute of Education Sciences, 2017). In addition to methodological rigor, adherence to the standards can increase transparency and confidence in the findings and improve the execution of replication studies (Appelbaum et al., 2008, 2018; Schulz et al., 2010). In the absence of complete and transparent reporting, readers are unable to assess the quality of the study, the validity and/or interpretation of the results, and/or distinguish unbiased from questionable results (Hopewell, Dutton, Yu, Chan, & Altman, 2010). In addition, inadequate adherence to standards and reporting has been associated with biased and exaggerated estimates of treatment effects (Hopewell et al., 2010).

Although development and dissemination of standards was intended to improve the design, conduct, and reporting of RCTs, reviews suggest variable and modest effects. For example, Hopewell et al. (2010) assessed methodological reporting adherence for 616 healthcare intervention RCTs published 10 years following publication of the CONSORT statement. Adherence rates were < 50% (range 1%–45%) for 9 of the 12 methodological items (with the other 3 items falling in the 53%–69% range); the authors characterized the overall adherence and reporting quality as unacceptable. Ivers et al. (2011) conducted a similar review of methodological reporting adherence for 161 randomly selected healthcare cluster RCTs published after publication of the CONSORT extension for cluster randomized trials. Results indicated adherence < 50% for 6 of the 14 items (range 9%–49%) and between 53% and 95% for the other 8 items. Consistent with Hopewell et al. (2010), Ivers et al. (2011) also noted some progress in adherence but

emphasized the need for methodological improvements and more comprehensive reporting in these trials. Although these studies provide insight into methodological and reporting trends in healthcare RCTs, the designs of the studies reviewed do not represent the common design used in GSSI RCTs for youth with ASD.

RCTs of GSSIs for youth with ASD have unique design and data analytic challenges because of the manner in which participants are most often randomized to treatment conditions and the interventions are delivered in group format (clusters). This results in a study design and methodology that is neither a true 2-group parallel trial design (in which individual participants are randomized and receive individual treatment, and statistical independence is established; Schulz et al., 2010) nor a true cluster randomized trial (in which existing clusters [e.g., classrooms] are randomly assigned to treatment conditions and cluster effects are assumed; Campbell et al., 2012). In an extension of the main CONSORT statement, Boutron et al. (2008) proposed additional standards for designs in which individuals are randomized to clusters and the intervention is delivered in group-format to clusters.

Of particular relevance to the current review is the manner in which the group-based delivery of treatment is accounted for in the sampling and data analytic methodology, both of which should be adjusted based on the estimated intraclass correlation coefficient (ICC; Boutron et al., 2008). Given these cluster-level implications, statistical power may be reduced based on the ICC and sample size estimates and outcome analyses should account for these (Boutron et al., 2008; Campbell et al., 2012). Publishing the ICC used in these calculations can also help communicate the magnitude of the clustering effect on the outcomes and estimates (Campbell et al., 2012) and with planning future studies and power calculations (Ivers et al., 2011). Failing to account for the effect of clusters falsely assumes independence among the participants and may yield an incorrect, and exaggerated, estimation of intervention effects (Boutron et al., 2008; Ivers et al., 2011). As such, a detailed examination of the sampling and outcome analytic techniques used in GSSI studies for youth with ASD is warranted.

As previously described, Gates et al. (2017) completed the most extensive meta-analysis of GSSIs for youth with ASD to date; it yielded an overall medium effect ($g = .51$) and provided critical information on the state of the RCT evidence. That study was not intended to, and did not review or analyze the study planning, data assessment, or data analytic methodologies used in the studies and/or whether cluster-based effects were taken into account. The present study examined the extent to which the studies included in the Gates et al. (2017) meta-analysis adhered to reporting standards in the methodological areas of study planning, data assessment, and data analysis. It was intended to complement the meta-analysis findings, provide the field with an understanding of the extent to which these RCTs met established standards, and improve design and reporting practices for future RCTs of GSSIs for youth with ASD. The studies were assessed using a comprehensive list of items generated from multiple reporting standards including the CONSORT (and extensions), JARS, AERA, WWC, and ASA. The list was developed to maximize content coverage while minimizing the application of overlapping items (see Measure section). Other factors including the year of article publication and journal impact factor were examined to determine whether they were related to adherence with the standards.

2. Methods

2.1. Studies reviewed

This study reviewed the same 18 RCTs of GSSIs for youth with ASD included in the Gates et al. (2017) meta-analysis. The 18 RCTs used to calculate the overall effect size estimate were retrieved and reviewed for the present study. The search and selection procedures for those 18 RCTs are described in detail in Gates et al. (2017) and will only be briefly reiterated here. Searches were conducted of databases (PubMed, PsychINFO, and Web of Science) using the terms "(ASD OR autism spectrum disorder OR Asperger OR autism OR pervasive developmental disorder) AND (social skills OR peer interaction OR social competence OR social functioning OR friendship OR social interaction OR social play) AND (treatment OR intervention) NOT (early intervention OR toddler OR early intensive behavior intervention) NOT (pharmacological OR medical)" (Gates et al., 2017, p. 167). Inclusion criteria were that the study was empirical, included youth ages 5–21 years diagnosed with autism, Aspergers, or PDDNOS, tested a treatment targeting core social impairments, was peer reviewed or a dissertation, did not include pharmacological or medical intervention or early intervention, and was written in English. Additionally, the studies were required to have used an RCT design, employed a group-delivered social skills intervention, included a comparison group (i.e., no-treatment, wait-list, or treatment-as-usual control), and been published through January 2016. There was a total of 735 participants in the included studies, with sample sizes ranging from 11 to 97 ($M = 41$). The mean age of the participants was 11 years (range 5 to 20 years), with one-third of the studies having > 90% males and the others having 50% to 90% males. The overall cognitive ability and verbal ability of participants was in the average range (overall $M = 102$, range 88 to 113; and verbal $M = 100$, range 86 to 106). Approximately 44% of studies provided comorbidity data on the participants and 50% included participants on medications. Lastly, significant variability was reported in intervention lengths (5 to 97 sessions) across studies (Gates et al., 2017).

2.2. Measure

2.2.1. Development of review checklist

Consistent with prior development and revisions of reporting standards (Appelbaum et al., 2018), the current checklist of items was created taking into account items from the various standards/guidelines (i.e., CONSORT, JARS, AERA, WWC, and ASA). Careful reviews of those standards revealed significant overlap in content and items, reflecting agreement on many features considered important in conducting and reporting RCTs. Given the focus of the current review, attention was directed toward items involving study planning (specifically power, precision, and analyses), data assessment (data quality and management), and data analysis

(procedures and reporting).

Per AERA (2006) recommendations, the following describes how the items were included and the checklist was created and agreed upon. The CONSORT statement (Schulz et al., 2010) and JARS (Appelbaum et al., 2018) were used to create the main headings for the review checklist. One of the study authors initially generated a list of headings which was then reviewed by one of the co-authors; minor clarifications/changes were discussed and made based upon consensus resulting in the headings *power and sample size estimation*, *data analysis plan/statistical methods*, *data quality and diagnostics*, *missing data*, *p-values*, and *inferential statistics, effect sizes and precision*. The same process was then repeated, with the author who created the headings compiling the list of items from each of the reporting standards according to the pertinent heading. A primary goal was to ensure the checklist comprehensively addressed each area using the standards, while minimizing redundancies in the items. A second author then reviewed the items per heading, and minor clarifications in wording were made via consensus between the two authors. Whenever possible, the item wording was preserved to be as close as possible to the item from the reporting standard(s). Agreement between the two authors for the assignment of primary items to a designated heading was 97% (33/34), with the disagreement resolved via discussion and consensus. The checklist was then distributed to, and reviewed and approved by the remaining authors, which confirmed the clarity of items and their inclusion under each of the designated headings. Two of the authors then reviewed the initial headings to further simplify the coding form and combine sections involving similar study features. Agreement between the two authors for the combination of sections assessing similar study features into broader headings was 100%. This process resulted in the creation of three broader headings including *study planning* (comprised of the *power and sample size estimation* and *data analysis plan/statistical methods* items), *data assessment* (comprised of the *data quality and diagnostics* and *missing data* items), and *data analysis* (comprised of the *p-values*, and *inferential statistics, effect sizes and precision* items). The final coding checklist was reviewed and approved by all of the study authors.

This process resulted in a total of 58 items included on the final coding form. Because a number of items utilize branching (i.e., item is only relevant if the prior item was positively endorsed), it was decided to present the 34 primary (core) items in Table 2, with Tables 3–5 containing all of the secondary branching items along with their primary items. The branching items provided additional information to assist with interpretation of results based on the 34 primary items (Table 2). Items in Tables 2–5 also identify the specific source(s)/standard(s) from which it was drawn.

2.2.2. Establishment and assessment of inter-observer agreement during coding

Coding was completed by the two lead authors. In order to initially establish inter-observer agreement (IOA), three articles were randomly selected from the 18 RCTs to be reviewed. To ensure the coding checklist was tested and IOA established on articles across the years encompassed, the 18 articles were listed in chronological order and separated into three equal clusters of six articles; next, one article was randomly selected from each cluster (i.e., selection of one from the earliest cluster of six articles, one from the middle cluster of six articles, and one from the most recent cluster of six articles). These three articles were then independently reviewed by the two reviewers using the coding checklist. After each article was reviewed and coded, the agreements and disagreements were logged and any disagreements were to be resolved by discussion and consensus. Given that the items were categorical, IOA was assessed using Kappa. Overall agreement was 1.0 (100% agreement) for the initial three articles, negating the need for discussion/consensus on these articles. Following establishment of IOA, the remaining 15 articles were independently reviewed and coded by the two reviewers using the same procedure. Upon completion of every one or two articles, the reviewers logged their results (and agreements and disagreements) and resolved disagreements via discussion and consensus. Overall agreement across the 18 articles was 0.963 ($p < 0.001$) and was 100% following consensus.

2.2.3. Study adherence, publication characteristics, and analysis plan

The coding checklist was completed for each article, with the final results indicating the consensus of the two reviewers for each item. Each item was rated on the basis of the presence or absence of the corresponding information reported within the article. Review of the secondary items took place whenever the associated primary item had been checked as present. In calculating adherence rates, both a raw and an adjusted adherence rate were calculated. The adjusted rate was calculated by excluding the items that were not applicable. In addition, information was collected to determine whether year of article publication or journal impact factor were associated with adherence to the standards. Consistent with Ivers et al. (2011), data on impact factors were obtained from *Journal Citation Reports* (Web of Science). Because journal impact factors can change from year to year and this review included studies across a number of years, the 5-year impact factor was used. Other journal-related information was collected from each journal's respective website (see Table 1 for article and journal information).

To address the primary purpose of documenting adherence rates with the standards, descriptive data were generated for each item, cluster, and the overall rate (frequencies and percentages). The secondary goal of assessing whether year of article publication or journal impact factor was associated with adherence rate was tested using Kendall's rank (tau-b) correlations.

3. Results

3.1. Overview and primary standard adherence

As noted, the review checklist focuses on three main areas including Study Planning, Data Assessment, and Data Analysis. It includes two kinds of items in each of these areas: 1) primary items reflecting a specific standard; and, 2) secondary items (i.e., branching items) that include details on a specific standard. Results for the primary items are presented first (Table 2), followed by

Table 1
Journal Names, Article Publication Dates, 5-Year Impact Factors, Manuscript Length Limits, and Presence of Reporting Guidelines by Journal.

Journal	5-Year Impact Factor (2016 Journal Citation Reports)	Word/Page Limit	Reporting Guidelines
Autism Research (n = 2) Begeer et al. (2015) Yoo et al. (2014)	4.310	5,000 words (20.6 pages)	CONSORT + others
JADD (n = 12) Gantman, Kapp, Orenski, and Laugeson, (2012) Solomon, Goodlin-Jones, and Anders, (2004) Begeer et al. (2011) Corbett et al. (2016) Frankel et al. (2010) Kamps et al. (2015) Koenig et al. (2010) Laugeson, Frankel, Mogil, and Dillon, (2009) Laugeson, Gantman, Kapp, Orenski, and Ellingsen, (2015) Lopata et al. (2010) Schohl et al. (2014) White et al. (2013)	4.099	20-23 pages preferred	Instructions but not for analyses
RASD (n = 2) Andrews, Attwood, and Sofronoff, (2013) Koning, Magill-Evans, Volden, and Dick, (2013)	2.165	6,000 words (24.8 pages)	CONSORT
BioPsychoSocial Medicine (n = 1) Ichikawa et al. (2013)	0.993*	None listed	CONSORT for Abstract format
Psychology in the Schools (n = 1) Thomeer et al. (2012)	1.588	None listed	None listed

Note. JADD = Journal of Autism and Developmental Disorders; RASD = Research in Autism Spectrum Disorders.
*BioPsychoSocial Medicine impact factor identified as Source Normalized Impact per Paper (SNIP; https://bpsmedicine.biomedcentral.com/). BioPsychoSocial Medicine accepted to Journal Citation Reports (2015): https://bpsmedicine.biomedcentral.com.

detailed reports on each of the three focal areas (Tables 3–5). The following description of results focuses on the primary items, with the information from the detailed reports (i.e., branching items) used to provide additional information to assist with interpretation.

Table 2 displays the 34 primary items by focal area. The Study Planning items (5 items) show that power analysis had been conducted in just two studies and precision analysis in none. Similarly, the principle of "Intent to treat" was specifically included in only two studies. An opposite result is shown for data analysis planning, as 17 of the 18 studies provided a data analysis plan and all but one of these was in sufficient detail to enable replication. The overall rate of adherence was 41% for planning.

The Data Assessment section (8 items) yielded an adherence rate of 42% on the relevant items. Reports of attrition were most evident, with 15 of the 16 applicable studies meeting this standard. Frequencies of missing data were given in nine of 11 applicable studies and of these nine, three utilized and presented details on procedures for replacing missing data. Similarly, three of the five applicable studies reported details on deleted cases. Seven of the 18 studies reported that they had checked the assumptions of the analysis. Only two of the 18 studies assessed for the presence of outliers, two of the eight applicable studies provided an explanation for data transformations, and none of the studies reported skewness and kurtosis statistics.

The Data Analysis section (21 items) had an overall adherence rate of 43% on the applicable items. All of the studies provided some discussion of clinical significance, but only three described a pre-specified method and threshold for clinical significance determination. All 18 studies reported the smallest unit of analysis but none accounted for groups (clusters) in the analysis or reported an intracluster correlation. When reporting descriptive statistics, 11 of the 18 studies provided these for the primary outcomes and six of 10 studies provided them for secondary outcomes. Of the 17 applicable studies, only three reported exact p-values for all analyses, six gave an a priori alpha error rate, and seven applied adjustments for multiple analyses. Effect sizes were given for each primary and secondary outcome in five of the 18 studies and six of the 17 applicable studies provided an effect size for every p-value reported. Of the 18 studies, only two used binary outcomes and one conducted moderation or mediation analyses however all met the standards when used. Lastly, multivariate and complex analyses were used in only two of the 18 studies.

3.2. Secondary branched item adherence

Table 3 presents additional details on Study Planning adherence. Only one study employed power analysis to plan sample size and one reported power on the obtained sample. These two studies provided an effect size with justification for the estimate utilized but did not give the alpha level, software, or power formula, and neither included an intraclass correlation in their power calculations. When reporting on analytic strategies, 15 of the 17 applicable studies met the standard for testing primary hypotheses and nine of the 10 met it for the secondary hypotheses.

Table 4 presents additional details on Data Assessment. Ten of the 14 applicable studies with attrition provided reasons for loss of

Table 2
Adherence to Reporting Criteria for Individual Items and Clusters for the 18 RCTs.

Reporting Criteria	Adherence *n*		
	Yes	No	NA
Study Planning (*n* = 5 items)			
1)Was a power analysis conducted? (6)	2	16	
2)Was a precision analysis conducted? (4, 6)		18	
3)Were planned analyses described? (5)	17	1	
4)Was enough detail provided that the analysis could be replicated? (5)	16	2	
5)Was the analysis via "intent to treat"? (1)	2	16	
Summary Results	37/90 (41%)		
Data Assessment (*n* = 8 items)			
6)Were the underlying assumptions of the analyses checked? (5, 6)	7	11	
7)Were skewness and kurtosis statistics reported? (6)		18	
8)Were details/criteria for any deleted cases given? (4, 6)	3	2	13
9)Were outliers assessed? (4, 6)	2	16	
10)Were any data transformations explained? (6)	2	6	10
11)Was attrition reported? (1)	15	1	2
12)Were frequencies/percentages of missing data given? (4)	9	2	7
13)If missing data were replaced/imputed, were the procedures detailed? (4, 6)	3		15
Summary Results	41/144 (29%)		
Adjusted Summary Results (accounting for NA)	41/97 (42%)		
Data Analysis (*n* = 21 items)			
14)Were all exact *p*-values reported? (3, 4, 6)	3	14	1
15)Were any adjustments made for multiplicities? (1, 3, 4)	7	10	1
16)Was the *a priori* alpha error rate given? (4)	6	11	1
17)Were results of all inferential statistics reported, including exact *p*-values, and minimally sufficient statistics like *df*, MS effect and MS error values needed to construct the tests given for all analyses (including non-significant tests)? (4, 6)	3	14	1
18)Are descriptions given of each primary outcome including N, n's, M's and SD's? (6)	11	7	
19)Are descriptions given of each secondary outcome including N, n's, M's and SD's? (6)	6	4	8
20)Is a clear differentiation provided for primary, secondary, and exploratory hypothesis tests and their estimates? (6)	4	10	4
21)Are results with effect sizes provided for each primary and secondary outcome? (1, 4, 5, 6)	5	13	
22)Are effect size estimates given for every *p*-value reported? (1, 3, 4)	6	11	1
23)For trials that include group-based (i.e., cluster) intervention, are results with effect sizes and precision given at the cluster level as applicable? (1)		18	
24)If binary outcomes were used, are both absolute and relative effect sizes provided? (1)	2		16
25)Were results of any moderation or mediation analyses described? (6)	1		17
26)Was a discussion of practical/clinical/meaningfulness of effects presented? (5)	18		
27)Was the method of assessing clinical significance given, including whether the threshold was pre-specified? (6)	3	15	
28)For group-based (cluster) intervention trials, is a coefficient of intracluster correlation (ICC or k) provided for each primary outcome? (1, 2)		18	
29)Was smallest unit (individual, group/cluster, class) analyzed described? (6)	18		
30)Was the analytic method used to account for groups (clusters) described? (e.g., multilevel analysis) (6)			18
31)If a multivariate analysis (e.g., MANOVA, SEM, or HLM) was used, was the variance-covariance or correlation matrix provided? (4, 6)	2	5	11
32)If a complex analysis, were details of models estimated given? (6)	2	5	11
33)If a complex analysis, was the statistical software identified? (6)	3	4	11
34)For group-based (clustered) intervention designs, did the report indicate if the analysis was done at the level of individuals or clusters? (1)	18		
Summary Results	118/378 (31%)		
Adjusted Summary Results (accounting for NA)	118/277 (43%)		
TOTAL SUMMARY RESULTS	196/612 (32%)		
TOTAL ADJUSTED SUMMARY RESULTS (accounting for NA)	196/464 (42%)		

Note. The studies included in this review all employed group-delivered social skills interventions. In this study, the term "cluster" refers to the individual groups in which the social skills interventions were provided. The term is not referring to "cluster randomized trials" in which existing clusters are randomly assigned to treatment conditions. Only one study in the current review randomly assigned existing clusters to treatment conditions (i.e., Kamps et al., 2015).
Number(s) in parentheses following the items indicate the standard(s)/source(s) from which it was drawn: 1 = CONSORT; 2 = WWC; 3 = ASA; 4 = JARS (2008); 5 = AERA; 6 = JARS (2018).

participants. Of the 11 applicable studies, only one provided reasons for missing data, and only two provided an assessment of patterns of missingness (e.g., missing at random or not).

Table 5 provides additional details on Data Analysis. Of the seven studies that reported adjustments for multiplicities, all seven applied the adjustments for the set of primary hypothesis tests and three applied them for the secondary hypotheses. This table also shows that only one study met the standard of providing a confidence interval for every effect size.

Table 3
Detailed Summary of Adherence to Study Planning Criteria for the 18 RCTs.

Reporting Criteria	Adherence *n*		
	Yes	No	NA
Study Planning			
1)Was a power analysis conducted? (6)	**2**	**16**	
1a. If yes, did the study specify the intended sample size based on power? (4, 6)	1	17	
1b. If yes, did the study specify the effect size used in power calculation? (4, 6)	2		16
1c. If yes, did the study specify the source of the effect size used in power calculation?	2		16
1d. If yes, was the alpha used in the power calculation stated?		2	16
1e. If yes, was the level of power stated?	2		16
1f. If yes, was the software used in the calculation noted?		2	16
1g. If yes, was the formula(s) used in the calculation noted?		2	16
1h. If the trial included clusters, were ICCs included in the power analysis? (1)		2	16
1ha. If yes, was the source of the ICC estimate reported? (1)			18
1hb. If yes, was an indicator of ICC uncertainty reported (e.g., SE, 95% CI) (1)			18
2)Was a precision analysis conducted? (4, 6)		**18**	
3)Were planned analyses described? (5)	**17**	**1**	
3a. Were analytic strategies for primary hypotheses described? (6)	15	2	1
3b. Were analytic strategies for secondary hypotheses described? (6)	9	1	8
3c. Were analytic strategies for exploratory hypotheses described? (6)		1	17
4)Was enough detail provided that the analysis could be replicated? (5)	**16**	**2**	
5)Was the analysis via “intent to treat”? (1)	**2**	**16**	

Note. Bolded and underlined values constitute primary items (also reported in Table 2).
Number(s) in parentheses following the items indicate the standard(s)/source(s) from which it was drawn: 1 = CONSORT; 2 = WWC; 3 = ASA; 4 = JARS (2008); 5 = AERA; 6 = JARS (2018).

Table 4
Detailed Summary of Adherence to Data Assessment Criteria for the 18 RCTs.

Reporting Criteria	Adherence *n*		
	Yes	No	NA
Data Assessment			
1)Were the underlying assumptions of the analyses checked? (5, 6)	**7**	**11**	
2)Were skewness and kurtosis statistics reported? (6)		**18**	
3)Were details/criteria for any deleted cases given? (4, 6)	**3**	**2**	**13**
4)Were outliers assessed? (4, 6)	**2**	**16**	
5)Were any data transformations explained? (6)	**2**	**6**	**10**
6)Was attrition reported? (1)	**15**	**1**	**2**
6a. If yes, were reasons for attrition given? (1)	10	4	4
7)Were frequencies/percentages of missing data given? (4)	**9**	**2**	**7**
7a. If yes, were reasons for missing data given? (1)	1	10	7
7b. If yes, was there an explanation of missing data patterns? (e.g., missing at random?) (4, 6)	2	9	7
7c. If yes, was there an analysis of characteristics of missing respondents? (2)	1	13	4
7d. If yes, were criteria for deciding when to infer missing data reported? (6)	1	2	15
8)If missing data were replaced/imputed, were the procedures detailed? (4, 6)	**3**		**15**

Note. Bolded and underlined values constitute primary items (also reported in Table 2).
Number(s) in parentheses following the items indicate the standard(s)/source(s) from which it was drawn: 1 = CONSORT; 2 = WWC; 3 = ASA; 4 = JARS (2008); 5 = AERA; 6 = JARS (2018).

3.3. Relationship of adherence to year of article publication and journal impact factor

Adherence rates for Study Planning, Data Assessment, and Data Analysis, as well as an overall adherence rate, were calculated at the study level in order to explore relationships with the year of publication and 5-year journal impact factor. Kendall’s rank (tau-b) correlation was utilized because year of publication and journal impact factor were not normally distributed. As indicated in Table 6, the correlations of the adherence rates with year of publication were low and not statistically significant. Fig. 1 depicts the timeline of standards introductions and overall adherence rates. The correlations of adherence rates with impact factor were also low and not statistically significant, with the exception of the correlation between Data Analysis adherence and impact factor. This coefficient was -.39 with an associated *p*-value of .05, equaling but not surpassing the < 0.05 threshold.

4. Discussion and implications

Youth with ASD exhibit chronic and pervasive social impairments and GSSIs are a commonly used psychosocial treatment to

Table 5
Detailed Summary of Adherence to Data Analysis Criteria for the 18 RCTs.

Reporting Criteria	Adherence *n*		
	Yes	No	NA
Data Analysis			
1)Were all exact *p*-values reported? (3, 4, 6)	**3**	**14**	**1**
2)Were any adjustments made for multiplicities? (1, 3, 4)	**7**	**10**	**1**
2a. If yes, were adjustments made for primary hypotheses? (6)	7		11
2b. If yes, were adjustments made for secondary hypotheses? (6)	3	1	14
2c. If yes, were adjustments made for exploratory hypotheses? (6)			18
3)Was the *a priori* alpha error rate given? (4)	**6**	**11**	**1**
4)Were results of all inferential statistics reported, including exact *p*-values, and minimally sufficient statistics like *df*, MS effect and MS error values needed to construct the tests given for all analyses (including non-significant tests)? (4, 6)	**3**	**14**	**1**
5)Are descriptions given of each primary outcome including N, n's, M's and SD's? (6)	**11**	**7**	
6)Are descriptions given of each secondary outcome including N, n's, M's and SD's? (6)	**6**	**4**	**8**
7)Is a clear differentiation provided for primary, secondary, and exploratory hypothesis tests and their estimates? (6)	**4**	**10**	**4**
8)Are results with effect sizes provided for each primary and secondary outcome? (1, 4, 5, 6)	**5**	**13**	
8a. If yes, are confidence intervals provided for each effect size? (1, 4, 5)	1	5	12
9)Are effect size estimates given for every *p*-value reported? (1, 3, 4)	**6**	**11**	**1**
9a. If yes, are confidence intervals provided for every effect size? (1, 3, 4)	1	5	12
10)For trials that include group-based (i.e., cluster) intervention, are results with effect sizes and precision given at the cluster level as applicable? (1)		**18**	
11)If binary outcomes were used, are both absolute and relative effect sizes provided? (1)	**2**		**16**
12)Were results of any moderation or mediation analyses described? (6)	**1**		**17**
12a. If yes, were the statistical methods used described? (6)	1		17
13)Was a discussion of practical/clinical/meaningfulness of effects presented? (5)	**18**		
14)Was the method of assessing clinical significance given, including whether the threshold was pre-specified? (6)	**3**	**15**	
15)For group-based (cluster) intervention trials, is a coefficient of intracluster correlation (ICC or k) provided for each primary outcome? (1, 2)		**18**	
16)Was smallest unit (individual, group/cluster, class) analyzed described? (6)	**18**		
17)Was the analytic method used to account for groups (clusters) described? (e.g., multilevel analysis) (6)			**18**
18)If a multivariate analysis (e.g., MANOVA, SEM, or HLM) was used, was the variance-covariance or correlation matrix provided? (4, 6)	**2**	**5**	**11**
19)If a complex analysis, were details of models estimated given? (6)	**2**	**5**	**11**
20)If a complex analysis, was the statistical software identified? (6)	**3**	**4**	**11**
21)For group-based (clustered) intervention designs, did the report indicate if the analysis was done at the level of individuals or clusters? (1)	**18**		

Note. Bolded and underlined values constitute primary items (also reported in Table 2).
Number(s) in parentheses following the items indicate the standard(s)/source(s) from which it was drawn: 1 = CONSORT; 2 = WWC; 3 = ASA; 4 = JARS (2008); 5 = AERA; 6 = JARS (2018).

Table 6
Kendall Rank (tau-b) Correlations of Adherence Rates with Year of Publication and Journal Impact Factor.

Adherence Area	Publication Year		Journal Impact Factor	
	τ_b	*p*-value	τ_b	*p*-value
Study Planning	−.24	.24	−.16	.51
Data Assessment	.16	.39	.06	.82
Data Analysis	−.10	.59	−.39	.05
Overall	.00	1.0	−.32	.10

address these deficits. Two meta-analyses of GSSI RCTs for youth with ASD supported the efficacy of these interventions in improving social competency/performance, with both yielding an overall moderate effect (Gates et al., 2017; Reichow et al., 2012). Although both meta-analyses included only studies that utilized RCT designs, neither examined the extent to which the studies met methodological standards for RCTs including for areas especially critical for interventions delivered in group format (i.e., clusters). The current review was conducted to determine the extent to which the study planning, data assessment, and data analyses reported in the studies included in the Gates et al. (2017) meta-analysis adhered to methodological and analytical standards including whether cluster-based factors were taken into account. Adherence to such standards can increase methodological rigor, transparency, and confidence in findings, and improve execution of replication studies and potentially the replicability of results (Appelbaum et al., 2008, 2018; Hopewell et al., 2010; Schulz et al., 2010).

Findings indicated an overall adherence rate of 42% and this rate was consistent across the three methodological and analytic areas. In the area of planning, the overall rate of 41% was largely the result of studies including a description of their analyses (17/18) and enough detail to permit replication of the analyses (16/18). No studies included a precision analysis and only two (11%) conducted a power analysis; neither of the power analyses took into account the group-delivered format of the intervention or

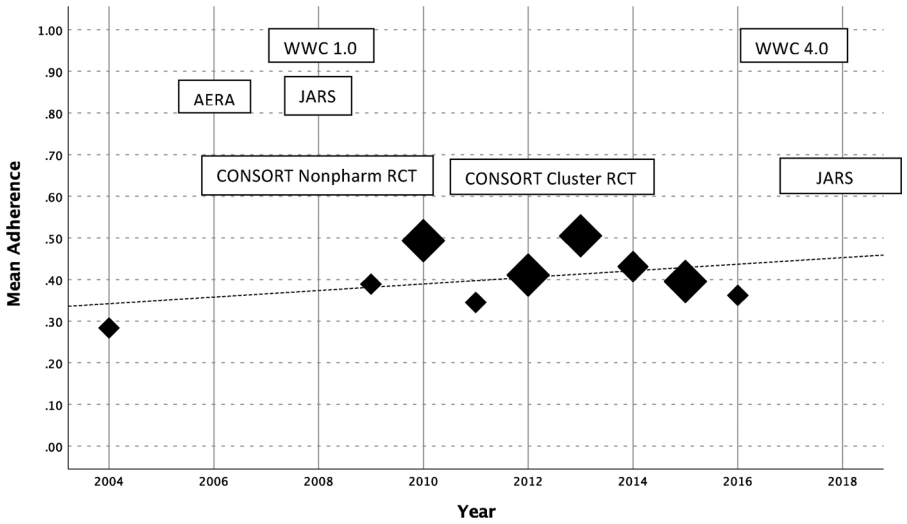

Fig. 1. Timeline of publication dates of the 18 studies, standards/guideline introductions, and mean adherence.

adjusted the power calculation based on the estimated ICC. The lack of a priori power estimation is a concern as insufficient power may result in potentially promising interventions being characterized as ineffective and/or abandoned (Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006). Further, the lack of accounting for clustering when power was estimated is also problematic as statistical power may be reduced based on the ICC and sample size estimates should be adjusted for these (Boutron et al., 2008; Campbell et al., 2012). The reason(s) for the lack of precision and power estimation in RCTs of GSSIs for youth with ASD is unknown but may be influenced by the nature of the clinical population and/or resource availability. For example, identifying and screening participants and executing GSSI RCTs require considerable infrastructure, resources, and expertise that might restrict larger-scale studies. Researchers may be willing to conduct these RCTs with smaller samples despite the potential for under-powered analyses. Lastly, only two studies indicated that their analyses were done via "intent to treat" which may bias outcome estimates and determinations due to selective discontinuation.

In the area of data assessment (i.e., examination and preparation of data for analyses), the adjusted overall adherence rate was 42% however there was significant variability by item/procedure. For the applicable items, higher rates of adherence with the standards were found for reporting attrition (94%) and missing data (82%). Other recommended aspects of the data assessment process occurred at far lower rates on the applicable items (testing of statistical assumptions 39%, reporting skewness and kurtosis statistics 0%, assessment of outliers 11%, and explanations of data transformations 25%). The findings suggested far more diligence and transparency when reporting on attrition and missing data compared to the more statistically-oriented aspects (e.g., statistical assumptions, skewness and kurtosis, and outliers) of the data assessment. The overall pattern suggests that investigators conducting GSSI RCTs should increase their testing of the assumptions underlying the statistical procedures and ensure that they are met so decisions regarding efficacy are not confounded by data-related problems.

The third area assessed involved adherence to data analysis standards. In this area, one of the highest levels of adherence involved reporting of descriptive statistics for the outcome measures (approximately 60% of the applicable studies); however only 29% of the applicable studies clearly differentiated the primary, secondary, and exploratory tests and a sizable minority of studies did not provide descriptive statistics for all the outcome measures. Another area of relative strength involved the discussion of clinical/practical meaningfulness; all of the studies provided narrative descriptions delineating clinical implications but only 17% identified an a priori method and threshold used to establish meaningfulness. This is perhaps not surprising given the lack of a clear method and criteria for testing and establishing clinical/practical meaningfulness but it highlights an important area for study and establishment of a common method(s) and criteria for conducting such tests. Results of the items assessing statistical adjustments and error rate revealed low levels of adherence to the standards, with only 41% of the applicable studies applying statistical adjustments for multiplicities and 35% identifying an a priori error rate. Less than one-fifth of the applicable studies reported all exact *p*-values (18%) and included enough detail to reconstruct the analyses (18%), and only 28% and 35% provided effect sizes for each outcome and *p*-value, respectively. This strongly indicates the need for more comprehensive and detailed reporting of tests, as well as for statistical adjustments given the multiple indicators and tests commonly included in GSSI RCTs for youth with ASD. Relying on non-protected tests as indicators of efficacy can increase the risk of Type I error. This problem is further compounded if results are interpreted in the absence of effect sizes and/or precision estimates. Although there is considerable debate about the use of probability testing to assess treatment outcomes, applying appropriate statistical adjustments and providing effect sizes and confidence intervals can assist reviewers and readers to better assess the accuracy, magnitude, and meaningfulness of results (Wasserstein & Lazar, 2016).

As previously indicated, the delivery of social interventions in group-format requires that the effects of groups (i.e., clusters) be accounted for in study analyses. Despite the fact that all 18 RCTs included in the review utilized a group-delivered intervention (GSSI), all conducted their analyses at the level of the individual participant. None of the studies utilized an analytic method that

accounted for groups (clusters), provided a coefficient of intracluster correlation, or included results with effect sizes or precision at the cluster level. Failure to account for cluster-level effects may be particularly problematic and lead to incorrect determinations of efficacy (Ivers et al., 2011), exaggerated estimates of effects, and errors in precision estimation. These findings are somewhat surprising given the fact that the 2008 extension to the CONSORT guidelines (Boutron et al., 2008) identified the need to account for clusters in these types of psychosocial/behavioral intervention studies (see Fig. 1). Finally, only one study conducted moderation or mediation analyses to help determine factors implicated in treatment efficacy. Despite widespread recognition of the need for such analyses, this continues to be an area in need of investigation. As GSSI RCTs with larger sample sizes become available and/or data are pooled across studies, such analyses will become more feasible.

The current study also assessed whether overall adherence was associated with year of article publication and journal impact factor. Given the increase in the number and sources of RCT standards, along with increased recognition of their importance, it was anticipated that adherence rates would have increased over time and been higher for journals with higher impact factors. Results indicated little relationship between adherence rates and article publication year. Although the regression line (Fig. 1) depicts a slight upward trend in adherence over time, this appears largely due to the low adherence for the single study published in 2004 (this preceded the introduction of many standards). Similarly, adherence rates were generally unrelated to journal impact factors, with the exception of adherence in the area of data analysis ($p = .05$). Interestingly, in this area adherence rate was inversely associated with journal impact factor (i.e., lower adherence associated with higher journal impact factor). This appeared to be due to the clustering of a large proportion of the studies published in one journal (*Journal of Autism and Developmental Disorders* [*JADD*]) with the associated high impact factor and relatively lower adherence rates compared with the four studies with higher adherence rates but lower impact factor scores. This is an area in need of ongoing assessment as the 18 RCTs included in this review were all published in only five journals, with 12 appearing in *JADD*.

The establishment and promulgation of standards appear to have had narrow effects on the design and reporting of RCTs of GSSIs for youth with ASD. The lack of adherence for the GSSI RCTs found in this study does not appear unique. As previously described, Hopewell et al. (2010) and Ivers et al. (2011) also found variable and low-level adherence and reporting for many methodological aspects of healthcare (medical) intervention RCTs. Results of the current study, as well as Hopewell et al. (2010) and Ivers et al. (2011) are consistent with the observation that adoption of standards has been slow (Appelbaum et al., 2018). This is further reinforced by the fact that the standards most applicable to GSSI RCTs for youth with ASD have been out for a decade (Boutron et al., 2008). In general, the lowest levels of adherence for the current study appeared to be for the more statistical aspects of the studies (e.g., power and precision estimation, checking and reporting on statistical assumptions, applying statistical adjustments, comprehensive reporting of results including effect sizes and confidence intervals, etc.). Of particular concern was the absence of accounting for cluster-level effects in power, precision, and outcome analyses which can have significant implications for efficacy decisions.

Given the relatively poor adherence in many areas, efforts appear to be needed to increase awareness of and adherence with established standards. An obvious initial step is for researchers conducting RCTs of GSSIs for youth with ASD to familiarize themselves with the standards. This can assist in developing and executing methodologically-rigorous RCTs that also comply with reporting requirements (Schulz et al., 2010). Journal editors can also publicize and require that RCT manuscripts adhere to a specific standard(s) (Ivers et al., 2011) and reviewers can be required or encouraged to use these when conducting reviews (Schulz et al., 2010). Greater awareness of the standards at all levels of study development, conduct, and review may increase the quality of planning, data assessment, and statistical analyses and reporting (Appelbaum et al., 2018; Campbell et al., 2012). A final but critically-important consideration is how clustering might impact the design, conduct, and outcomes of the studies. Factors such as therapist training, competence, and fidelity and homogeneity of participants might affect (increase or decrease) the influence of clustering and should be considered when designing and executing studies and analyzing outcomes. Although the standards indicate that clustering must be accounted for in such designs, no studies were identified that empirically tested the extent to which clustering impacts efficacy determinations or magnitudes of effects in GSSIs for youth with ASD. Such studies are needed to help researchers understand the impact of clustering and factors that increase or reduce its effects.

This study yielded important information to help improve the planning and execution of data analyses in GSSI RCTs for youth with ASD however several limitations warrant mention. This study was conducted as a complement to a prior meta-analysis (Gates et al., 2017) and included only the studies from that investigation; GSSI RCTs published since that time are not represented in the current results. Although more-recent studies may have been published, the lack of association between adherence and year of publication found in the current study would not predict a significant increase in adherence during that short time period. Another limitation involved the completion of the review checklist based on information reported in the individual articles; this does not guarantee that a particular task was not done, only that it was not reported. This limitation may be relevant for some items (e.g., failure to report checking statistical assumptions or skewness/kurtosis statistics) but was clearly not for other items such as those dealing with accounting for clustering effects. Comprehensive reporting is needed but may be inhibited by page/word limits or other journal parameters that force authors to make difficult choices regarding what to include in a manuscript. For example, authors and reviewers may emphasize the need for more detail in describing the treatment protocol or other study features at the expense of information on the data analyses. This issue may be attenuated by allowing authors to include detailed information about the treatment protocol in online supplemental information available to interested readers (Appelbaum et al., 2018). It is also important to note that, despite substantial overlap, the review checklist included items from several reporting standards which would make 100% adherence more difficult. Another limitation was associated with the limited number of journals in which the 18 RCTs appeared (5 journals); 12 of these were published in one journal (*JADD*) and 16 were published in ASD-specific journals. This indicates that ASD-specific journals are the predominant outlets for dissemination of these RCTs and also highlights the important and influential role that they can play in increasing adherence with planning, assessment, and analysis standards. A final limitation was associated with

the survey which included a different number of items in each of the three areas assessed. The number of items per section were unequal but were given equal weight when calculating the overall rate of adherence.

Despite these limitations, our hope is that this review is seen in the context of efforts to improve experimentation that are rooted in the methodological literature since the time of McCall (1923) and Fisher (1925). This history includes seminal work by methodologists such as Campbell and Stanley (1963); Cook and Campbell (1978), and the various professional groups that have recently produced the checklists and standards that guided this review. All of these authors and standards/guidelines have in common the goal of improving the validity of study outcomes. From Campbell's (1959) evolutionary perspective, scientific wisdom results from better methods, gradually and systematically building cumulative knowledge. In this way, the challenges associated with conducting complex experiments in ASD may become more manageable, ultimately benefitting children and families.

Conflicts of interest

The authors declare no conflicts of interest.

Ethical approval

This article did not include human participants or animals.

Funding

This research was supported by Department of Education, Institute of Education Sciences Grant R324A130216. Findings and conclusions are those of the authors and do not necessarily reflect the views of the funding agency. The sponsor had no role in the design, collection, analysis or interpretation of the data, writing of the report, and decision to submit for publication.

References[1]

American Educational Research Association (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40. https://doi.org/10.3102/0013189X035006033.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed). Arlington, VA: American Psychiatric Association.

*Andrews, L., Attwood, T., & Sofronoff, K. (2013). Increasing the appropriate demonstration of affectionate behavior, in children with Asperger syndrome, high functioning autism, and PDD-NOS: A randomized controlled trial. *Research in Autism Spectrum Disorders, 7*(12), 1568–1578. https://doi.org/10.1016/j.rasd.2013.09.010.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force. *The American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191.

Appelbaum, M., Cooper, H., Maxwell, S., Stone, A., & Sher, K. J. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist, 63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839.

*Begeer, S., Gevers, C., Clifford, P., Verhoeve, M., Kat, K., Hoddenbach, E., et al. (2011). Theory of mind training in children with autism: A randomized controlled trial. *Journal of Autism and Developmental Disorders, 41*(8), 997–1006. https://doi.org/10.1007/s10803-010-1121-9.

*Begeer, S., Howlin, P., Hoddenbach, E., Clauser, C., Lindauer, R., Clifford, P., et al. (2015). Effects and moderators of a short theory of mind intervention for children with autism spectrum disorder: A randomized controlled trial. *Autism Research,* (April), 738–748. https://doi.org/10.1002/aur.1489.

Bellini, S., Gardner, L., & Markoff, K. (2014). Social skill interventions. In (4th ed.). F. R. Volkmar, S. J. Rogers, R. Paul, & K. A. Pelphrey (Vol. Eds.), *Handbook of autism and pervasive developmental disorders: Vol. 2*, (pp. 887–906). Hoboken, NJ: John Wiley & Sons Assessment, interventions, and policy.

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., & Ravaud, P. (2008). Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. *Annals of Internal Medicine, 148*, 295–309. https://doi.org/10.7326/0003-4819-148-4-200802190-00008.

Campbell, D. T. (1959). Methodological suggestions from a comparative psychology of knowledge processes. *Inquiry, 2*, 152–182.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 statement: Extension to cluster randomized trials. *British Medical Journal, 345*, 1–21. https://doi.org/10.1136/bmj.e5661.

Christensen, D. L., Baio, J., Braun, K. V. N., Bilder, D., Charles, J., Constantino, J. N., et al. (2016). *Prevalence and characteristics of autism spectrum disorder among children aged 8 years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. MMWR, 65(No. SS-3)(No. SS-3).* 1–23.

Cook, T. D., & Campbell, D. T. (1978). *Quasi-experimentation.* Boston: Houghton Mifflin.

*Corbett, B. A., Key, A. P., Qualls, L., Fecteau, S., Newsom, C., Coke, C., et al. (2016). Improvement in social competence using a randomized trial of a theatre intervention for children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 46*(2), 658–672. https://doi.org/10.1007/s10803-015-2600-9.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Boyd.

*Frankel, F., Myatt, R., Sugar, C., Whitham, C., Gorospe, C. M., & Laugeson, E. (2010). A randomized controlled study of parent-assisted children's friendship training with children having autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(7), 827–842. https://doi.org/10.1007/s10803-009-0932-z.

*Gantman, A., Kapp, S. K., Orenski, K., & Laugeson, E. A. (2012). Social skills training for young adults with high-functioning autism spectrum disorders: A randomized controlled pilot study. *Journal of Autism and Developmental Disorders, 42*(6), 1094–1103. https://doi.org/10.1007/s10803-011-1350-6.

Gates, J. A., Kang, E., & Lerner, M. D. (2017). Efficacy of group social skills interventions for youth with autism spectrum disorder: A systematic review and meta-analysis. *Clinical Psychology Review, 52*, 164–181. https://doi.org/10.1016/j.cpr.2017.01.006.

Hopewell, S., Dutton, S., Yu, L. M., Chan, A. W., & Altman, D. G. (2010). The quality of reports of randomised trials in 2000 and 2006: Comparative study of articles indexed in PubMed. *British Medical Journal, 340*, 1–8. https://doi.org/10.1136/bmj.c723.

*Ichikawa, K., Takahashi, Y., Ando, M., Anme, T., Ishizaki, T., Yamaguchi, H., et al. (2013). TEACCH-based group social skills training for children with high-functioning autism: A pilot randomized controlled trial. *BioPsychoSocial Medicine, 7*(1), 14. https://doi.org/10.1186/1751-0759-7-14.

Ivers, N. M., Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., et al. (2011). Impact of the CONSORT extension for cluster randomised trials on quality of reporting and study methodology: Review of random sample of 300 trials, 2000-8. *British Medical Journal, 343*, 1–14. https://doi.org/10.1136/bmj.d5886.

Kaat, A. J., & Lecavalier, L. (2014). Group-based social skills treatment: A methodological review. *Research in Autism Spectrum Disorders, 8*(1), 15–24. https://doi.org/

[1] *Indicates RCT included in the review.

10.1016/j.rasd.2013.10.007.
*Kamps, D., Thiemann-Bourque, K., Heitzman-Powell, L., Schwartz, I., Rosenberg, N., Mason, R., et al. (2015). A comprehensive peer network intervention to improve social communication of children with autism spectrum disorders: A randomized trial in kindergarten and first grade. *Journal of Autism and Developmental Disorders, 45*(6), 1809–1824. https://doi.org/10.1007/s10803-014-2340-2.
Kasari, C., Dean, M., Kretzmann, M., Shih, W., Orlich, F., Whitney, R., et al. (2016). Children with autism spectrum disorder and social skills groups at school: A randomized trial comparing intervention approach and peer composition. *Journal of Child Psychology and Psychiatry, 57*(2), 171–179. https://doi.org/10.1111/jcpp.12460.
*Koenig, K., White, S. W., Pachler, M., Lau, M., Lewis, M., Klin, A., et al. (2010). Promoting social skill development in children with pervasive developmental disorders: A feasibility and efficacy study. *Journal of Autism and Developmental Disorders, 40*(10), 1209–1218. https://doi.org/10.1007/s10803-010-0979-x.
*Koning, C., Magill-Evans, J., Volden, J., & Dick, B. (2013). Efficacy of cognitive behavior therapy-based social skills intervention for school-aged boys with autism spectrum disorders. *Research in Autism Spectrum Disorders, 7*(10), 1282–1290. https://doi.org/10.1016/j.rasd.2011.07.011.
Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry, 63*(5), 484–489. https://doi.org/10.1001/archpsyc.63.5.484.
*Laugeson, E. A., Frankel, F., Mogil, C., & Dillon, A. R. (2009). Parent-assisted social skills training to improve friendships in teens with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(4), 596–606. https://doi.org/10.1007/s10803-008-0664-5.
*Laugeson, E. A., Gantman, A., Kapp, S. K., Orenski, K., & Ellingsen, R. (2015). A randomized controlled trial to improve social skills in young adults with autism spectrum disorder: The UCLA PEERS® program. *Journal of Autism and Developmental Disorders, 45*(12), 3978–3989. https://doi.org/10.1007/s10803-015-2504-8.
*Lopata, C., Thomeer, M. L., Volker, M. A., Toomey, J. A., Nida, R. E., Lee, G. K., et al. (2010). RCT of a manualized social treatment for high-functioning autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(11), 1297–1310. https://doi.org/10.1007/s10803-010-0989-8.
McCall, H. W. (1923). *How to experiment in education.* New York: MacMillan.
McMahon, C. M., Lerner, M. D., & Britton, N. (2013). Group-based social skills interventions for adolescents with higher-functioning autism spectrum disorder: A review and looking to the future. *Adolescent Health, Medicine and Therapeutics, 4*, 23–38. https://doi.org/10.2147/AHMT.S25402.
Portway, S. M., & Johnson, B. (2005). Do you know I have Asperger's syndrome? Risks of a non-obvious disability. *Health, Risk & Society, 7*(1), 73–83. https://doi.org/10.1080/09500830500042086.
Reichow, B., Steiner, A. M., & Volkmar, F. (2012). Social skills groups for people aged 6 to 21 with autism spectrum disorders (ASD). *The Cochrane Database of Systematic Reviews*(7), https://doi.org/10.1002/14651858.CD008511.pub2 Art. No.: CD008511.
Scarpa, A., Reyes, N., & Attwood, T. (2013). Cognitive-behavioral therapy for stress and anger management in young children with ASD: The exploring feelings program. In A. Scarpa, S. W. White, & T. Attwood (Eds.). *CBT for children and adolescents with high-functioning autism spectrum disorders* (pp. 147–170). New York, NY: Guilford.
Scarpa, A., White, S. W., & Attwood, T. (2013). *CBT for children and adolescents with high-functioning autism spectrum disorders.* New York: Guilford.
*Schohl, K. A., Van Hecke, A. V., Carson, A. M., Dolan, B., Karst, J., & Stevens, S. (2014). A replication and extension of the PEERS intervention: Examining effects on social skills and social anxiety in adolescents with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 44*(3), 532–545. https://doi.org/10.1007/s10803-013-1900-1.
Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine, 152*(11), 726–732. https://doi.org/10.7326/0003-4819-152-11-201006010-00232.
Shattuck, P. T., Wagner, M., Narendorf, S., Sterzing, P., & Hensley, M. (2011). Post-high school service use among young adults with an autism spectrum disorder. *Archives of Pediatrics & Adolescent Medicine, 165*(2), 141–146. https://doi.org/10.1001/archpediatrics.2010.279.
*Solomon, M., Goodlin-Jones, B. L., & Anders, T. F. (2004). A social adjustment enhancement intervention for high functioning autism, Asperger's syndrome, and pervasive developmental disorder NOS. *Journal of Autism and Developmental Disorders, 34*(6), 649–668. https://doi.org/10.1007/s10803-004-5286-y.
*Thomeer, M. L., Lopata, C., Volker, M. A., Toomey, J. A., Lee, G. K., Smerbeck, A. M., et al. (2012). Randomized clinical trial replication of a psychosocial treatment for children with high-functioning autism spectrum disorders. *Psychology in the Schools, 49*(10), 942–954. https://doi.org/10.1002/pits.21647.
U. S. Department of Education, Institute of Education Sciences (2017). *What works clearinghouse standards handbook, version 4.0.* https://ies.ed.gov/ncee/wwc/Handbooks.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108.
*White, S. W., Ollendick, T., Albano, A. M., Oswald, D., Johnson, C., Southam-Gerow, M. A., et al. (2013). Randomized controlled trial: Multimodal anxiety and social skill intervention for adolescents with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 43*(2), 382–394. https://doi.org/10.1007/s10803-012-1577-x.
*Yoo, H. J., Bahn, G., Cho, I. H., Kim, E. K., Kim, J. H., Min, J. W., et al. (2014). A randomized controlled trial of the Korean version of the PEERS® parent-assisted social skills training program for teens with ASD. *Autism Research, 7*(1), 145–161. https://doi.org/10.1002/aur.1354.