A Sensitivity Analysis of Three Nonparametric Treatment Effect Scores for Single-Case Research for Participants with Autism

Monica E. Carr

Received: 14 September 2014 / Accepted: 15 October 2014 / Published online: 29 October 2014
© Springer Science+Business Media New York 2014

Abstract The widely used percentage of nonoverlapping data (PND) treatment effect calculation was compared to more recently developed methods which, it has been argued, better account for outlying variables and trend in single-case design (SCD) intervention studies. Percentage of all nonoverlapping data (PAND) and nonoverlap of all pairs (NAP) were selected for comparison as both are amenable to hand calculation, making them widely accessible to clinicians and teachers as well as researchers. A data set was developed through a systematic search of peer-reviewed literature on self-management interventions conducted with participants with autism spectrum disorder (ASD). Treatment effect sizes derived from each method were compared for studies that provided sufficient data. Results indicated that PND provided a conservative measure of strength of treatment effect when compared to PAND and NAP scores. Interpretation scales for treatment effect scores derived from each method were reviewed. Implications for selecting a calculation method for participants with ASD are discussed.

Keywords Autism · Treatment effect · Nonparametric · Hand calculation · PND

Introduction

In 1994, the American Psychological Association (APA) publication manual encouraged the inclusion of a treatment effect size in research reports. Since that time, the APA Taskforce on Statistical Inference (1999) argued that a treatment effect size permits the evaluation of the stability of findings across samples and is important to future meta-analyses. Subsequently, in

some quarters, treatment effect scores have been considered a requirement for research publication (Leland Wilkinson and the Taskforce on Statistical Inference 1999).

The initial report developed by the APA Taskforce on Statistical Inference (1996) warned that, with advances in state-of-the-art statistical analysis software, statistics are commonly reported without understanding of the computational methods or necessarily even an understanding of what the statistics mean. Parker and colleagues have also argued that treatment effect calculations reported in meta-analysis of single-case design (SCD) research should be interpretable by various different stakeholder groups (Parker et al. 2005), a point well illustrated with autism intervention-related research where parents, educators and policy makers as well as clinicians and researchers all need to understand reports on treatment effects.

Given the increasing demand to develop an evidence base in educational psychology, quality assessment guidelines have been developed by the US Department of Education What Works Clearinghouse (WWC) (Kratochwill et al. 2010; Kratochwill et al. 2013a; Kratochwill et al. 2013b). Various methods for determining treatment effects have been proposed for use in meta-analyses of SCD research, although the merits of these different computation methods remain a matter of debate (Horner and Kratochwill 2012; Horner et al. 2012; Kratochwill et al. 2013b; Scruggs and Mastropieri 2013). Unlike group research designs, a generally accepted method for the calculation of treatment effect size for SCD research has yet to be established. The initial version of the WWC SCD guidelines indicated a preference for regression based procedures for calculating effect sizes though the WWC panel subsequently suggested conducting a sensitivity analysis of treatment effect scores using several indices. Most recently, the WWC has moved away from the use of a treatment effect score and reverted to visual analyses until a general consensus on the most appropriate method has been reached (Kratochwill et al. 2013b).

M. E. Carr (✉)
Monash University, Melbourne, VIC 3800, Australia
e-mail: mebar4@student.monash.edu

There is an extensive body of literature examining approaches to the evaluation of the strength of treatment effects for SCD research. Shadish (2014) has reported that a number of new methods are currently in development including standardised mean difference approaches, multilevel analysis and Bayesian generalised least squares procedures. In a recent review of SCD research conducted with students with a broad array of disabilities, Maggin et al. (2011a) reported percentage of nonoverlapping data (PND) as the most frequently used treatment effect score appearing in 55 % of the 84 effect sizes garnered from 68 literature syntheses. Maggin and colleagues also reported that nearly 40 % of authors did not describe the method they used for comparing data from various phases within studies when estimating effect sizes. Of the studies that did include these details, several approaches were described. These included taking an arithmetic mean across all baseline and treatment phases, selecting only certain phases (i.e., A1B2) and consolidating baselines and treatment phases (i.e., A1A2B1B2).

The PND (Scruggs et al. 1987) was developed as a method to synthesise SCD literature which could be easily calculated and readily as well as meaningfully interpreted. Several positive features of PND have been described and include ease of calculation from graphical rather than raw data, high degree of inter-rater reliability, applicability to any SCD design type and ease of interpretation (Campbell 2013; Parker et al. 2007). The continued utility of the procedure has recently been argued by the original developers (Scruggs and Mastropieri 2013). However, PND is not without its critics, and major limitations of the procedure include the following: (i) PND requires its own interpretation guidelines as it does not correlate to an accepted effect size; (ii) it lacks sensitivity in discriminating treatment effectiveness as the calculated score approaches 100 %; (iii) PND is reliant on a single extreme data point in baseline, and all other baseline data are excluded from the calculation; and (iv) as PND has no known sampling distribution, confidence intervals cannot be calculated (Parker et al. 2007).

All current alternative procedures have their own limitations however. Maggin et al. (2011b) compared 11 commonly used effect size measures including three parametric methods: interrupted time series analysis procedure (ITSACORR), piecewise regression, hierarchical linear modelling; and seven nonparametric methods: PND, percentage of all nonoverlapping data (PAND), percent of zero data points (PZD), pairwise overlap squared (PDO²), percentage exceeding mean (PEM), percentage exceeding mean trend line (PEM-T), improvement rate difference (IRD) and the standardised mean difference. Of the nonparametric methods that were assessed, PAND received the most favourable assessment.

PAND (Parker et al. 2007) has been presented as an alternative to PND, the developers recommending it for documentation and accountability purposes in schools and clinics, in

addition to applications in meta-analyses and academic research reports. Their method was illustrated with sample data that typically contained between 60 and 80 data points, and the authors noted that the method was not well suited for data series that contained fewer than 20–25 data points. Parker et al. (2011) reported that PAND has been adopted in two meta-analyses along with a *phi* correlation coefficient, which is analogous to an R^2 score, that is frequently reported in large N studies (Parker et al. 2011).

Recently, Parker and Vannest (2009) developed the nonoverlap of all pairs (NAP) procedure, suggesting that this method offers an improvement on both PND and PAND. Several anticipated advantages have been suggested by these researchers, notably that the calculation uses all data points and as such should yield a more representative treatment effect score. Unlike other nonparametric indices, NAP is not based on means or medians, and it has been suggested that the calculation should relate more closely to the regression term R^2 . Importantly for stakeholders within the autism spectrum disorder (ASD) community, NAP can be calculated by hand. NAP was omitted from the effect size comparison conducted by Maggin and colleagues (2011b) however, Parker and colleagues (2011) reported that NAP has also been used in several recent meta-analyses.

Although greater consensus is evident between researchers regarding how to calculate treatment effects in group design research (Kratowchwill et al. 2013b), it has been noted that the interpretation of these treatment effect scores can also be problematic (Brossart et al. 2006). Brossart and colleagues observed that a basis for comparison of treatment effect sizes obtained using different calculations is lacking in the literature, arguably making this task challenging for clinicians and other stakeholders. These researchers noted that simple methods tend to yield different effect sizes than regression-based methods and that even ballpark interpretation guidelines for R^2 —e.g., “large” ($R^2=.25$), medium ($R^2=.09$) and small ($R^2=.01$)—drawn from large N group research in social science vary depending on the field of investigation.

Guidelines for interpretation of derived scores have been clearly defined for PND. However, other than the original developers describing a *phi* correlation coefficient for PAND, based upon a Pearson R 2×2 contingency table (Parker et al. 2007), it appears that an interpretation scale for PAND is not available in the literature to date. Parker and Vannest (2009) did provide a tentative interpretation scale for NAP, analogous to that used in PND, based on a process of expert judgements of 200 data sets.

In a recent exploration of the characteristics of SCD data for participants with ASD, Carr et al. (2014) reported a declining trend over time in the volume of data gathered in both baseline and treatment phases in SCD studies in both an *established treatment* (self-management) and an *emerging treatment* field (physical exercise) as classified by the

National Standards Report (2009). Only 23.7 % of the studies included in the review reported a sufficient volume of data for a regression-based calculation. Carr and colleagues also explored the applicability of three readily hand-calculated non-parametric procedures for calculating effect sizes. PND was selected because of the frequency with which the procedure is reported and both PAND and NAP because they were developed to address limitations evident in PND. The authors concluded that a NAP calculation, which is not restricted by either volume of data points or the presence of ceiling or floor points in baseline, appeared appropriate for all studies that were included in the review and that PND was applicable with 90 % of the studies sampled. Conversely, PAND, which can only be applied when a minimum of 20 data points are presented, appeared applicable for only 54 % of the studies.

The purpose of this current study was to conduct a sensitivity analysis on treatment effect scores for use by the variety of stakeholders working with the ASD community. Accordingly, a primary requirement of the procedures included was the ability to perform all calculations by hand. PND has been selected as the basis of comparison in the sensitivity analysis, as it has been widely adopted in published literature. Burns et al. (2012) have recently recommended that additional research on new overlap approaches, particularly PAND and NAP, is warranted. The literature review for this current study has also found support for the suggestion made by Burns and colleagues, with NAP identified on the basis of anticipated improvements and PAND on the basis of prior favourable review. The current study sought to explore the advantages and limitations of PND, PAND and NAP. In addition, it was noted in the literature that previous research on newer calculations has been limited to AB designs (Brossart et al. 2006; Parker and Vannest 2009). As such, this study has included data from all phases.

The following research questions were investigated:

1. Do estimated effect sizes calculated using PND, PAND and NAP differ significantly from each other?
2. What benefits or limitations are evident in estimating treatment effect size using PAND, or NAP, when compared to the PND method?
3. How do calculated treatment effect scores compare with each other using available interpretation scales?

Method

Data Set Creation

Studies were located by conducting a systematic search of peer-reviewed literature prior to November 2013. Both PsycINFO and ERIC databases were queried using the

keywords “autism*” and “Asperger’s syndrome” which were combined with the following terms typically associated with self-management: “self-management”, “self-regulation”, “self-regulate”, “self-monitoring”, “self-recording”, “self-reinforcement”, “self-evaluation”, “self-advocacy”, “self-observation”, “self-instruction”, “empowerment”, “self-determination” and “self-control”. In addition, a hand search of the reference lists of existing systematic reviews of self-management studies was undertaken.

The abstract of each article was examined to determine whether the article was likely to meet inclusion criteria for further review. The original article was retrieved and reviewed when further clarification appeared necessary. No age limits were placed upon participants. Inclusion criteria required that

1. Participants had an existing diagnosis of ASD or AS (for studies that included participants with differing conditions, only participants with ASD or AS were included for further review).
2. The study utilized a single-subject research design such as a multiple baseline, reversal, changing criteria or alternating treatment design.
3. Data for each phase and for each participant was presented in graphical format thus enabling calculation of PND, PAND and/or NAP.
4. Components of self-management were included throughout the intervention.
5. Articles were published in an English language peer-reviewed journal.

This search procedure identified 38 articles that were included for further review.

Calculating Treatment Effect

A treatment effect score was calculated for each participant included in each study as described for the following three methods.

PND (Scruggs et al. 1987) was calculated by counting the number of treatment data points that exceed the most extreme baseline data point, in the expected direction determined by whether an increase or decrease in target behaviour was desired. This number was then divided by the total number of treatment phase data points. Scruggs and colleagues have advised against coding a study when baseline stability has not been established and additionally noted that for cases including ceiling or floor baseline data points that yield a 0 % PND, the variation between treatment effect score and original research findings should be described.

PAND (Parker et al. 2007) was calculated by determining the minimum number of data points that need to be removed from either the baseline and/or treatment phases to eliminate all overlap. The number of remaining data points was then

divided by the total number of data points across baseline and treatment phases. This number represents the overlap, which is then subtracted from 100 to derive the nonoverlap and finally multiplied by 100 to express this value as a percentage.

NAP (Parker and Vannest 2009) was calculated by counting all nonoverlapping pairs. Often, this is achieved most quickly by counting overlapping pairs and subtracting from the total possible pairs to obtain the nonoverlap count. The total possible pairs are determined by multiplying the number of data points in the baseline phase with the number of data points in the treatment phase. Scores are assigned for each pairwise comparison and totalled. An overlap counts as one point, a tie counts as a half point and a nonoverlap receives a zero. Each overlap score is summed, and the total subtracted from the total possible pairs. The result is in turn divided by the total possible pairs and then multiplied by 100 to derive the percentage of all nonoverlapping pair treatment effect score.

Various approaches to determining treatment effect scores beyond an initial AB phase comparison were identified in previous research of treatment effect calculation methods. Skiba et al. (1985) argued in favour of an effect size based solely on the first AB phase comparison, claiming that treatment effects beyond the first treatment tested may be confounded with multiple treatment interference or that failure to revert to baseline levels in subsequent baseline phases may be attributed to lack of experimental control or powerful treatment effects. Other approaches were based on a combination of comparable phases prior to calculation of an arithmetic mean (Scruggs et al. 1987) and a comparison of first A with last B phase (Allison and Gorman 1993). The methodology adopted by Scruggs and colleagues was selected as preferable, based on consistency with the widely published PND metric.

Interpretation of Treatment Effect Scores

The scales provided by the respective original authors of each method have been adopted to interpret treatment effect scores and are summarised in Table 1. Scruggs and colleagues have suggested the following ranges for the interpretation of PND scores: 0–50 % ineffective, 50–70 % questionable, 70–90 % effective and 90 % or greater very effective (Scruggs and Mastropieri 1998). Parker and colleagues (2007) presented

PAND as an alternative to PND; however, their original paper does not describe an interpretation scale analogous to that of PND. While a *phi* correlation coefficient can be derived using a 2×2 table of proportions, an interpretation scale for the output of this computation has not been described by the developers. Consequently, an interpretation scale for PAND or *phi* has been omitted from Table 1. In their more recent research that compared treatment effect scores with expert visual judgements made on 200 published AB phase comparisons, Parker and colleagues have proposed the following tentative ranges for the interpretation of NAP: 0–65 % weak effect, 66–92 % medium effect and 93–100 % strong effect (Parker and Vannest 2009).

Inter-Observer Agreement

Reliability of computations was verified by conducting inter-observer checks, and an initial trial coding was performed using four randomly selected studies. Both the author and a Senior Professor within the Faculty of Education separately hand counted data points for each phase of each data series and recorded tallies on a coding sheet. PND, PAND and NAP were then calculated independently by hand by each coder and recorded in the coding sheet for all studies that reported sufficient data. Both the coders then met to discuss any variations in results, and a 100 % agreement for these four studies was achieved.

Subsequently, a further 14 articles (36.8 %) were selected at random, and each coder independently calculated tallies for data points for each phase of the data series and calculated PND, PAND and NAP for each AB phase comparison. Both coders then met again to compare the scores that each had calculated independently. Overall, 172 agreements were achieved from a total of 179 treatment effect calculations, and an overall inter-coder agreement was calculated at 96.7 for 47 % of the studies included in the total data set. When calculated separately for computational procedure, an inter-observer agreement (IOA) of 98.3 % was calculated for PND, 96.6 % for PAND and 93.3 % for NAP.

Subjective assessments were made on the consistency of interpretation of the treatment effect scores that were reported in Table 2. Both assessors met to discuss a method for determining consistency between interpretation scales, and consistency was operationalised using the scales provided in Table 1 as follows:

1. One rating of “ineffective” and the other rating as any other considered a disagreement
2. Ratings on the same band as each other considered an agreement
3. Ratings plus/minus one or more band considered a disagreement.

Table 1 Interpretation scale for strength of treatment effect score

PND		NAP	
0–50 %	Ineffective		
50–70 %	Questionable	0–65 %	Weak effect
70–90 %	Effective	66–92 %	Medium effect
90 %+	Very effective	93–100 %	Strong effect

PND percentage of nonoverlapping data, NAP nonoverlap of all pairs

Table 2 Participant treatment effect scores

HEADING Author (Year)		Treatment effect score		Score variances		PAND		NAP		PAND cf PND		NAP cf PND		Interpretation		NAP	Consistent
		PND		PAND		PAND		PAND		PAND		PAND		PND			
*Koegel and Koegel (1990)	Student 1	100.0 %		100.0 %		100.0 %		100.0 %		0.0 %		0.0 %		Very effective		Strong effect	Agree
	Student 2	98.3 %		98.4 %		99.6 %		99.6 %		1.3 %		1.2 %		Very effective		Strong effect	Agree
	Student 3	40.8 %		95.1 %		84.9 %		84.9 %		44.1 %		-10.3 %		Ineffective		Strong effect	Disagree
	Student 4	94.1 %		82.5 %		83.2 %		83.2 %		-10.9 %		0.8 %		Very effective		Medium effect	Disagree
*Sainato et al. (1990)	C1	12.5 %		77.9 %		75.2 %		75.2 %		65.4 %		-2.7 %		Ineffective		Medium effect	Disagree
	C3	0.0 %		68.8 %		61.3 %		61.3 %		68.8 %		-7.5 %		Ineffective		Weak effect	Disagree
	C4	62.5 %		94.4 %		96.3 %		96.3 %		31.9 %		1.8 %		Questionable		Strong effect	Disagree
	Tony	53.9 %		83.9 %		92.8 %		92.8 %		30.0 %		8.9 %		Questionable		Medium effect	Disagree
*Koegel et al. (1992)	Adam	89.5 %		88.3 %		96.0 %		96.0 %		-1.2 %		6.5 %		Effective		Strong effect	Disagree
	Howard	50.0 %		81.4 %		77.7 %		77.7 %		31.4 %		-3.7 %		Effective		Medium effect	Agree
	Ian	62.2 %		97.8 %		98.0 %		98.0 %		35.6 %		0.3 %		Very effective		Strong effect	Agree
	Bruce	100.0 %		100.0 %		100.0 %		100.0 %		0.0 %		0.0 %		Very effective		Strong effect	Agree
*Stahmer and Schreibman (1992)	Justin	73.2 %		83.8 %		82.9 %		82.9 %		10.6 %		-0.9 %		Effective		Medium effect	Agree
	Claire	100.0 %		100.0 %		100.0 %		100.0 %		0.0 %		0.0 %		Very effective		Strong effect	Agree
	Chris	Not coded		90.8 %		92.4 %		92.4 %		n/a		1.6 %		Not coded		Medium effect	n/a
	Andre	Not coded		73.3 %		80.9 %		80.9 %		n/a		7.6 %		Not coded		Medium effect	n/a
Pierce and Schreibman (1994)	Jon	100.0 %		n/a		100.0 %		100.0 %		0.0 %		n/a		Very effective		Strong effect	Agree
	Howard	100.0 %		n/a		100.0 %		100.0 %		0.0 %		n/a		Very effective		Strong effect	Agree
	Robby	100.0 %		n/a		100.0 %		100.0 %		0.0 %		n/a		Very effective		Strong effect	Agree
	Barret	33.7 %		93.8 %		97.3 %		97.3 %		49.4 %		-5.2 %		Ineffective		Strong effect	Disagree
*Strain et al. (1994)	Aubrey	96.1 %		99.0 %		100.0 %		100.0 %		4.0 %		1.9 %		Very effective		Strong effect	Agree
	Sidney	100.0 %		100.0 %		100.0 %		100.0 %		0.0 %		0.0 %		Very effective		Strong effect	Agree
	Scott	96.7 %		94.3 %		99.3 %		99.3 %		-2.4 %		2.7 %		Very effective		Strong effect	Agree
	Peter	90.5 %		81.3 %		97.4 %		97.4 %		-9.2 %		6.9 %		Very effective		Strong effect	Agree
*Strain and Danko (1995)	Alex	50.0 %		77.8 %		75.0 %		75.0 %		27.8 %		-2.8 %		Ineffective		Medium effect	Disagree
	North	80.0 %		87.5 %		94.0 %		94.0 %		7.5 %		6.5 %		Effective		Strong effect	Disagree
	Broderick	100.0 %		100.0 %		100.0 %		100.0 %		0.0 %		0.0 %		Very effective		Strong effect	Agree
	Jarred	50.0 %		84.2 %		98.1 %		98.1 %		34.2 %		13.8 %		Ineffective		Strong effect	Disagree
Dixon et al. (1995)	Student 1 F	28.6 %		n/a		51.2 %		51.2 %		32.9 %		n/a		Ineffective		Weak effect	Disagree
*Newman et al. (1996)	Marv	89.5 %		93.5 %		94.0 %		94.0 %		4.1 %		0.5 %		Effective		Strong effect	Disagree
	James	78.9 %		90.3 %		95.2 %		95.2 %		11.4 %		4.9 %		Effective		Strong effect	Disagree
	Andy	83.3 %		93.1 %		97.9 %		97.9 %		9.8 %		4.8 %		Effective		Strong effect	Disagree
	Child 1	50.0 %		n/a		71.9 %		71.9 %		12.5 %		n/a		Ineffective		Medium effect	Disagree
Shearer et al. (1996)																	

Table 2 (continued)

HEADING		Treatment effect score			Score variances		Interpretation		NAP	Consistent
Author (Year)	PND	PAND	NAP	PAND cf PND	NAP cf PND	NAP	PND			
Kern et al. (1997) *Callahan and Rademacher (1999) *Reinecke et al. (1999)	Child 2	25.0 %	n/a	98.3 %	47.7 %	73.3 %	n/a	Ineffective	Strong effect	Disagree
	Child 3	25.0 %	n/a	60.0 %	44.2 %	35.0 %	n/a	Ineffective	Weak effect	Disagree
	Charlie	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Seth	48.8 %	85.0 %	61.2 %	36.2 %	12.3 %	-23.8 %	Ineffective	Weak effect	Disagree
	Jack-school	58.3 %	92.3 %	98.5 %	34.0 %	40.2 %	6.2 %	Questionable	Strong effect	Disagree
	Paul-school	60.0 %	88.5 %	97.6 %	28.5 %	37.6 %	9.2 %	Questionable	Strong effect	Disagree
	Stan-home	94.7 %	96.7 %	99.5 %	1.9 %	4.8 %	2.9 %	Very effective	Strong effect	Agree
	Keri Study 1	94.4 %	96.8 %	98.8 %	2.4 %	4.4 %	2.0 %	Very effective	Strong effect	Agree
	Keri Study 2	92.1 %	99.2 %	99.4 %	7.1 %	7.3 %	0.2 %	Very effective	Strong effect	Agree
	Evan	92.9 %	95.0 %	97.0 %	2.1 %	4.2 %	2.0 %	Very effective	Strong effect	Agree
*Newman et al. (2000)	Dan	85.7 %	92.9 %	96.9 %	7.1 %	11.2 %	4.1 %	Effective	Strong effect	Agree
	Nancy	88.9 %	96.3 %	94.4 %	7.4 %	5.6 %	-1.9 %	Effective	Strong effect	Agree
	Larry	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Karen	85.1 %	93.7 %	99.8 %	8.6 %	14.7 %	6.1 %	Effective	Strong effect	Disagree
	Rick	84.3 %	99.0 %	99.8 %	14.7 %	15.5 %	0.8 %	Effective	Strong effect	Disagree
	Jason	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	Stewart	95.3 %	97.3 %	100.0 %	1.9 %	4.6 %	2.7 %	Very effective	Strong effect	Agree
	Alice	50.0 %	73.1 %	89.3 %	23.1 %	39.3 %	16.2 %	Ineffective	Medium effect	Disagree
	Bob	50.0 %	76.9 %	71.9 %	26.9 %	21.9 %	-5.0 %	Ineffective	Medium effect	Disagree
	Keith	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
*Wehmeyer et al. (2003) Apple et al. (2005)	Roger	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Abby	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Alex	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Jim	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	Jacob	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
*Newman and Ten Eyck (2005)	Booker	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	Alex	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	GS	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	AH	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	Tom	83.3 %	96.4 %	97.9 %	13.1 %	14.6 %	1.5 %	Very effective	Strong effect	Agree
*Todd and Reid (2006)	Mike	82.6 %	96.3 %	97.3 %	13.7 %	14.7 %	1.0 %	Effective	Strong effect	Disagree
	Robert	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Effective	Strong effect	Disagree
	Alan	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Delano (2007)							Very effective	Strong effect	Agree

Table 2 (continued)

HEADING		Treatment effect score			Score variances			Interpretation		Consistent
Author (Year)	PND	PAND	NAP	PAND cf PND	NAP cf PND	NAP cf PAND	PND	NAP		
Loflin et al. (2008)	Peter	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Justin	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Stuart	75.0 %	n/a	97.5 %	20.0 %	30.0 %	n/a	Effective	Strong effect	Disagree
	Anthony	72.7 %	n/a	97.5 %	16.0 %	35.6 %	n/a	Effective	Strong effect	Disagree
	William	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
Palmen et al. (2008)	Dick	33.3 %	n/a	55.6 %	33.3 %	22.2 %	n/a	Ineffective	Weak effect	Disagree
	Tom	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Kees	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Rob	66.7 %	n/a	88.9 %	16.7 %	22.2 %	n/a	Questionable	Medium effect	Disagree
	Linda	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
Dorminy et al. (2009)	Piet	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Sarah	66.7 %	n/a	94.4 %	16.7 %	27.8 %	n/a	Questionable	Strong effect	Disagree
	Jan	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Klaas	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Jake	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
Soares et al. (2009)	Ben	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Evan	84.6 %	n/a	99.0 %	11.8 %	14.4 %	n/a	Effective	Strong effect	Disagree
	Patrick	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Barry	92.3 %	95.7 %	98.5 %	3.3 %	6.2 %	2.8 %	Very effective	Strong effect	Agree
	Charlie	88.9 %	n/a	93.1 %	3.4 %	4.2 %	n/a	Effective	Strong effect	Disagree
Deitchman et al. (2010)	Jason	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Trevor	83.3 %	n/a	99.1 %	10.0 %	15.7 %	n/a	Effective	Strong effect	Disagree
	Mike	88.9 %	n/a	97.5 %	5.6 %	8.6 %	n/a	Effective	Strong effect	Disagree
	Justin	88.9 %	n/a	98.1 %	6.3 %	9.3 %	n/a	Effective	Strong effect	Disagree
	George	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
Cihak et al. (2010)	Adam	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Jordan	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Richard	100.0 %	n/a	100.0 %	0.0 %	0.0 %	n/a	Very effective	Strong effect	Agree
	Tony	73.7 %	n/a	98.6 %	19.6 %	24.9 %	n/a	Effective	Strong effect	Disagree
	Graham	63.6 %	n/a	99.0 %	32.2 %	35.3 %	n/a	Questionable	Strong effect	Disagree
*Todd et al. (2010)	Lisa	84.6 %	90.0 %	94.5 %	5.4 %	9.9 %	4.5 %	Effective	Strong effect	Disagree
	Daniel	85.0 %	95.8 %	94.7 %	10.8 %	9.7 %	−1.1 %	Effective	Strong effect	Disagree
	Mark	11.1 %	85.7 %	31.5 %	74.6 %	20.4 %	−54.2 %	Ineffective	Weak effect	Disagree

Table 2 (continued)

HEADING Author (Year)	Treatment effect score			Score variances			Interpretation		Consistent
	PND	PAND	NAP	PAND of PND	NAP of PND	NAP of PAND	PND	NAP	
Parker and Kamps (2010)	95.5 %	n/a	99.8 %	2.8 %	4.4 %	n/a	Very effective	Strong effect	Agree
	98.0 %	n/a	99.8 %	0.5 %	1.8 %	n/a	Very effective	Strong effect	Agree
Shogren et al. (2011)	100.0 %	100.0 %	100.0 %	0.0 %	0.0 %	0.0 %	Very effective	Strong effect	Agree
	87.5 %	94.4 %	98.8 %	6.9 %	11.3 %	4.3 %	Effective	Strong effect	Disagree

*Studies that met minimum data requirements for all three calculation methods

PND percentage of nonoverlapping data, PAND percentage of all nonoverlapping data, NAP nonoverlap of all pairs

Both assessors separately rated each participant treatment effect score then compared their findings. A total of 58 agreements and 43 disagreements were recorded by both assessors, and a further two scores were described as not applicable for this procedure. An IOA of 100 % was obtained for this process.

Results

A data set based on 38 articles that reported treatment data gathered across a variety of behaviours and settings for 103 participants was developed. Hand counts of the number of data points reported in each baseline and treatment phase were conducted for each of the 215 data series included in 177 graphs. These tallies were entered manually into an excel spreadsheet and subsequently compared to calculation guidelines to determine the suitability of applying PND, PAND and NAP calculations.

Baseline variability that included numerous ceiling data points was identified in one study (Koegel and Frea 1993), for which a PND result could not be calculated. Another three studies included ceiling data points in baseline that resulted in a 0 % PND for the respective participants (Koegel et al. 1992; Shearer et al. 1996; Stahmer and Schreibman 1992). As a result of too few data points noted in 15 studies, a PAND treatment effect score was calculated in 23 studies. A NAP treatment effect score was calculated for all 38 studies.

Treatment effect scores were calculated for all 103 participants for whom sufficient data was provided (see Table 2). Variations between PND, PAND and NAP scores are summarised, and treatment effect scores interpreted for PND and NAP.

To examine the variation in treatment effect score methods, the data set was reduced to include only studies that met the requirements for all three treatment effect score calculations. Twenty two studies were identified, and data for 57 participants was included. These 22 studies are indicated by an asterisk in Table 2.

Mean treatment effect scores for the 57 participants were PND 78.8 %, PAND 92.7 % and NAP 93.2 %. NAP indicated that treatment effect was on average 14.4 % greater than the product of the PND calculation and 0.5 % greater than that produced by PAND. PAND indicated strength of treatment effect that was 13.9 % greater when compared to PND.

A single-factor ANOVA was used to test the null hypothesis, that the variances between the mean treatment effect scores are equal. H_0 : the mean variances of treatment effect scores are equal. A summary of the mean variances for 57 participants and the ANOVA data is presented in Table 3.

The single-factor ANOVA resulted in an F_{calc} value that was greater than the F_{crit} value, and the conclusion drawn that

Table 3 Analysis of variance: $n=22$ studies, 57 participants

Summary						
Groups	Count	Sum	Average	Variance		
PND	57	44.92883	78.8 %	0.063873		
PAND	57	52.83922	92.7 %	0.006321		
NAP	57	53.13582	93.2 %	0.015822		
ANOVA						
Source of variation	SS	df	MS	F	p value	F _{crit}
Between groups	0.760334	2	0.380167	13.25908	4.5012E-06	3.049792
Within groups	4.816929	168	0.028672			
Total	5.577263	170				

PND percentage of nonoverlapping data, PAND percentage of all nonoverlapping data, NAP nonoverlap of all pairs

the mean variances are not equal. The difference was statistically significant ($F_{2,168}=13.259$, $p<.005$). The mean PND score of 78.8 % for 22 studies indicates an effective treatment. By comparison, the mean NAP score of 93.2 % indicates a strong treatment effect for the same 22 studies.

PND and NAP treatment effect scores were interpreted using the scales provided by the original authors and subjectively assessed for consistency in interpretation for all 103 participant scores. It appeared that these scales yielded a consistent strength of treatment effect for 58 scores (56.3 %) but appeared inconsistent for 43 scores (41.7 %), and a comparison was not applicable for the remaining two scores (1.9 %).

Discussion

The purpose of this research was to explore the suitability of three nonparametric calculation methods to estimate treatment effect size with SCD studies, with a specific focus on the needs of participants with ASD. In particular, PND, PAND and NAP hand calculation techniques were selected, and published data from self-management interventions conducted with participants diagnosed on the autism spectrum was used to test the calculations.

For each calculation method, a mean of all participant scores was calculated. Results from the ANOVA test of differences between the three methods suggested that the mean treatment effect scores derived using these three treatment effect calculation methods differ significantly with PND producing the most conservative estimates of effect size.

Benefits and limitations are apparent for each scoring method. A main criticism of PND is the weighting it places on extreme, possibly outlier data points in baseline phases (Parker et al. 2007). Scruggs and colleagues have defended their procedure, stating that these potential problems have rarely been encountered in the research literature, and when they are encountered, they can be easily addressed, as was

noted in their original conventions, by acknowledging such discrepancies in the research report (Scruggs and Mastropieri 2013). Consistent with their conclusion, the current data set included relatively few instances in which outlying data points in base line skewed the resultant treatment effect score.

Both PAND and NAP address the weakness acknowledged in PND by integrating additional baseline data points in the algorithm. PAND incorporates additional baseline data although eliminates all overlapping data across baseline and treatment phases. This study identified a significant proportion of participant data that does not meet the minimal threshold of 20 data points. Of further concern, the original developers proposed a minimum range of data points of 20–25 data points. Had the upper level of their suggested threshold, 25 data points, been adopted, it is likely that an even greater number of studies would have been deemed as not suitable for PAND methodology.

The NAP calculation appeared to offer the greatest advantage in this regard as the algorithm incorporates a pairwise comparison of all data points included in the data set, thus utilising every data point recorded. NAP is not restricted by a minimum number of data points and, in that sense, is preferable to PAND as a treatment effect score can be calculated for all studies with even greater precision. However, as NAP requires a more complex calculation than PND, or PAND, it is more error prone and potentially problematic with longer data series when calculated by hand.

The treatment effect scores calculated for both PND and NAP have been interpreted using rating scales in published literature, however no such scale is available for PAND. Using PND, the mean effect size for self-management interventions was described as *effective*, the second highest category under this scale. By contrast, using NAP, the mean effect size for self-management interventions was described as *strong*, the highest category on the scale. The difference in interpretative guidelines, in addition to the observed difference in scores derived by the procedures, suggests that PND reports a more conservative strength of treatment effect than that calculated

using NAP. This discrepancy underscores the warning by Brossart and colleagues (2006) that clinicians, as well as other interested stakeholders, face difficulty when interpreting studies that use various treatment effect calculation methods.

Conclusion

Fifteen years ago, the Wilkinson and the Taskforce on Statistical Inference (1999) emphasised the importance of understanding how a given statistical measure is calculated and how to interpret the statistic. Findings from this study suggest that both issues remain of concern in SCD research conducted with participants with ASD.

Of the three treatment effect scores that were reviewed, PAND may cautiously be considered the least applicable to stakeholders in the autism community for two reasons. First, a significant percentage of the articles included in this review did not include sufficient data points to permit a PAND-based analysis, and as observed elsewhere (Carr et al. 2014), evidence suggests that researchers are not collecting more data as would be required were PAND to be adopted in the future. Second, interpretation of the PAND score, or its associated ϕ or ϕ^2 correlation coefficient, is difficult in the absence of a conversion scale like that available for both PND and NAP. Further, as reported by Brossart and colleagues (2006), differences across research fields in the interpretation of a ϕ or ϕ^2 term may further compromise interpretation of these statistics. Further research into the development of a scale to interpret a PAND/ ϕ calculation, with a particular focus on participants with ASD, is justified.

Few studies in this sample included lengthy baselines and consequently the NAP calculation was relatively straightforward to apply to the studies included in this review. Importantly, NAP utilised all data points reported for each participant. Given that a scale for interpreting strength of treatment effect has also been proposed, these factors arguably add support to the adoption of NAP as a potential improvement over PND for incorporation in research for participants with ASD. However, given the greater treatment effect scores calculated by NAP compared to those of PND, the adoption of NAP as a new standard should be treated cautiously.

The PND metric currently dominates SCD literature, and the present data show it to yield a relatively conservative result with strength of treatment effect for self-management intervention procedures described as *effective* under PND and *strong* under NAP. Such calibration differences across methods for calculating treatment effects, were they found to be generalizable, are unlikely to contribute positively to our understanding of the relative effectiveness of our intervention procedures. Arguably, this issue is relevant to the ASD community, and research reports using newer alternate treatment effect scores should be treated with caution to avoid

presenting potentially misleading information to the ASD research stakeholders. Importantly, this study has indicated that PND is widely applicable to the data that is gathered for participants with ASD, and its continued use appears justified.

References

*Studies included in the review

- *Agran, M., Alper, S., Cavin, M., Sinclair, T., Wehmeyer, M., & Hughes, C. (2005). Using self-monitoring to increase following-direction skills of students with moderate to severe disabilities in general education. *Education and Training in Developmental Disabilities*, 40, 3–13.
- Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: the case of the single case. *Behaviour Research and Therapy*, 31, 621–631. doi:10.1016/0005-7967(93)90115-B.
- *Apple, A. L., Billingsley, F., & Schwartz, I. S. (2005). Effects of video modeling alone and with self-management on compliment-giving behaviors of children with high-functioning ASD. *Journal of Positive Behavior Interventions*, 7, 33–46. doi:10.1177/10983007050070010401.
- *Asaro-Saddler, K., & Saddler, B. (2010). Planning instruction and self-regulation training: effects on writers with autism spectrum disorders. *Exceptional Children*, 77, 107–124. doi:10.1177/1053451208330895.
- Brossart, D. F., Parker, R. I., Olson, E., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, 30(5), 531–563. doi:10.1177/0145445503261167.
- Burns, M. K., Zaslofsky, A. F., Kanive, R., & Parker, D. C. (2012). Meta-analysis of incremental rehearsal using phi coefficients to compare single-case and group designs. *Journal of Behavioral Education*, 21, 185–202. doi:10.1007/s10864-012-9160-2.
- *Callahan, K., & Rademacher, J. A. (1999). Using self-management strategies to increase the on-task behavior of a student with autism. *Journal of Positive Behavior Interventions*, 1, 117–122. doi:10.1177/109830079900100206.
- Campbell, J. M. (2013). Commentary on PND at 25. *Remedial and Special Education*. doi:10.1177/0741932512454725.
- Carr, M. E., Anderson, A., Moore, D. W., & Evans, W. H. (2014). How should we determine treatment effectiveness with single-case design research for participants with autism spectrum disorder? *Review Journal of Autism and Developmental Disorders*. doi:10.1007/s40489-014-0030-9.
- *Cihak, D. F., Wright, R., & Ayres, K. M. (2010). Use of self-modeling static-picture prompts via a handheld computer to facilitate self-monitoring in the general education classroom. *Education and Training in Autism and Developmental Disabilities*, 45, 136–149.
- *Deitchman, C., Reeve, S. A., Reeve, K. F., & Progar, P. R. (2010). Incorporating video feedback into self-management training to promote generalization of social initiations by children with autism. *Education and Treatment of Children*, 33(3), 475–488. doi:10.1353/etc.0.0102.
- *Delano, M. E. (2007). Improving written language performance of adolescents with Asperger syndrome. *Journal of Applied Behavior Analysis*, 40, 345–351. doi:10.1901/jaba.2007.50-06.
- *Dixon, R. S., Moore, D. W., Hartnett, N., Howard, R., & Petrie, K. (1995). Reducing inappropriate questioning behaviour in an adolescent with autism: a case study. *Behaviour Change*, 12, 163–166.

- *Dorminy, K. P., Luscre, D., & Gast, D. L. (2009). Teaching organizational skills to children with high functioning autism and Asperger's syndrome. *Education and Training in Developmental Disabilities*, 44, 538–550.
- *Ganz, J. B., & Sigafoos, J. (2005). Self-monitoring: are young adults with MR and autism able to utilize cognitive strategies independently? *Education and Training in Developmental Disabilities*, 40(1), 24–33.
- *Holifield, C., Goodman, J., Hazelkorn, M., & Heflin, L. J. (2010). Using self-monitoring to increase attending to task and academic accuracy in children with autism. *Focus on Autism and Other Developmental Disabilities*, 25, 230–238. doi:10.1177/1088357610380137.
- Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: some brief reflections. *Journal of Behavioral Education*, 21, 266–272. doi:10.1007/s10864-012-9152-2.
- Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the Systematic Analysis and Use of Single-Case Research. *Education and Treatment of Children*, 35(2), 269–290. doi:10.1353/etc.2012.0011.
- *Kern, L., Marder, T. J., Boyajian, A. E., Elliot, C. M., & McElhattan, D. (1997). Augmenting the independence of self-management procedures by teaching self-initiation across settings and activities. *School Psychology Quarterly*, 12(1), 23–32. doi:10.1037/h0088944.
- *Koegel, L. K., Koegel, R. L., Hurley, C., & Frea, W. D. (1992). Improving social skills and disruptive behavior in children with autism through self-management. *Journal of Applied Behavior Analysis*, 25, 341–353. doi:10.1901/jaba.1992.25-341.
- *Koegel, R. L., & Frea, W. D. (1993). Treatment of social behavior in autism pivotal social skills. *Journal of Applied Behavior Analysis*, 26, 369–377. doi:10.1901/jaba.1993.26-369.
- *Koegel, R. L., & Koegel, L. (1990). Extended reductions in stereotypic behavior of students with autism through a self-management package. *Journal of Applied Behavior Analysis*, 23, 119–127. doi:10.1901/jaba.1990.23-119.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). WWC Single-case designs technical documentation Version 1.0.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013a). What Works Clearinghouse procedures and standards handbook (Version 2.1).
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013b). *What Works Clearinghouse procedures and standards handbook* (Version 3.0).
- Leland Wilkinson and the Taskforce on Statistical Inference. (1999). The taskforce on statistical inference statistical methods in psychology journals; guidelines and explanations. *American Psychologist*, 54, 594–604. doi:10.1037/0003-066X.54.8.594.
- *Lofstin, R. L., Odom, S. L., & Lantz, J. F. (2008). Social interaction and repetitive motor behaviors. *Journal of Autism and Developmental Disorders*, 38, 1124–1135. doi:10.1007/s10803-007-0499-5.
- Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011a). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality*, 19(2), 109–135. doi:10.1080/09362835.2011.565725.
- Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011b). A generalized least squares regression approach for computing effect sizes in single-case research: application examples. *Journal of School Psychology*, 49(3), 301–321. doi:10.1016/j.jsp.2011.03.004.
- *Mancina, C., Tankersley, M., Kamps, D., Kravits, T., & Parrett, J. (2000). Brief report: reduction of inappropriate vocalizations for a child with autism using a self-management treatment program. *Journal of Autism and Developmental Disorders*, 30(6), 599–606. doi:10.1023/a:1005695512163.
- *Mithaug, D. K., & Mithaug, D. E. (2003). Effects of teacher-directed versus student-directed instruction on self-management of young children with disabilities. *Journal of Applied Behavior Analysis*, 36(1), 133–136. doi:10.1901/jaba.2003.36-133.
- *Morrison, L., Kamps, D., Garcia, J., & Parker, D. (2001). Peer mediation and monitoring strategies to improve initiations and social skills for students with autism. *Journal of Positive Behavior Interventions*, 3, 237–250. doi:10.1177/109830070100300405.
- National Standards Report. (2009) *National Autism Center*; Randolph MA.
- *Newman, B., Buffington, D., & Hemmes, N. (1996). Self-reinforcement used to increase the appropriate conversation of autistic teenagers. *Education and Training in Mental Retardation and Developmental Disabilities*, 31, 304–309.
- *Newman, B., Buffington, D., O'Grady, M., McDonald, M., Poulson, C., & Hemmes, N. (1995). Self-management of schedule following in three teenagers with autism. *Behavioral Disorders*, 20, 190–196.
- *Newman, B., Reinecke, D. R., & Meinberg, D. L. (2000). Self-management of varied responding in three students with autism. *Behavioral Interventions*, 15, 145–151. doi:10.1002/(SICI)1099-078X(200004/06)15:2<145::AID-BIN50>3.0.CO;2-3.
- *Newman, B., & Ten Eyck, P. (2005). Self-management of initiations by students diagnosed with autism. *The Analysis of Verbal Behavior*, 21, 117–122.
- *Palmen, A., Didden, R., & Arts, M. (2008). Improving question asking in high-functioning adolescents with autism spectrum disorders: effectiveness of small-group training. *Autism: The International Journal of Research and Practice*, 12(1), 83–98. doi:10.1177/1362361307085265.
- *Parker, D., & Kamps, D. (2010). Effects of task analysis and self-monitoring for children with autism in multiple social settings. *Focus on Autism and Other Developmental Disabilities*, 26, 131–142. doi:10.1177/1088357610376945.
- Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., De-Alba, R. G., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: how large is large? *School Psychology Quarterly*, 34, 116–132.
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): an alternative to PND. *The Journal of Special Education*, 40, 194–204. doi:10.1177/00224669070400040101.
- Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behavior Therapy*, 40, 357–367. doi:10.1016/j.beth.2008.10.006.
- Parker, R. I., Vannest, K., & Davis, J. (2011). Effect size in single-case research: a review of nine nonoverlap techniques. *Behavior Modification*, 34, 303–322. doi:10.1177/0145445511399147.
- *Pierce, K. L., & Schreibman, L. (1994). Teaching daily living skills to children with autism in unsupervised settings through pictorial self-management. *Journal of Applied Behavior Analysis*, 27, 471–481. doi:10.1901/jaba.1994.27-471.
- *Reinecke, D. R., Newman, B., & Meinberg, D. L. (1999). Self-management of sharing in three preschoolers with autism. *Education and Training in Mental Retardation and Developmental Disabilities*, 34, 312–317.
- *Sainato, D. M., Strain, P. S., Lefebvre, D., & Rapp, N. (1990). Effects of self-evaluation on the independent work skills of preschool children with disabilities. *Exceptional Children*, 56, 540–549.
- Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: issues and applications. *Behavior Modification*, 22, 221–242. doi:10.1177/01454455980223001.
- Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: past, present, and future trends in summarizing single-subject research. *Remedial and Special Education*, 34, 9–19. doi:10.1177/0741932512440730.

- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: methodology and validation. *Remedial and Special Education, 8*(2), 24–33. doi:[10.1177/074193258700800206](https://doi.org/10.1177/074193258700800206).
- *Shabani, D. B., Wilder, D. A., & Flood, W. A. (2001). Reducing stereotypic behavior through discrimination training, differential reinforcement of other behavior, and self-monitoring. *Behavioral Interventions, 28*(6), 279–286. doi:[10.1002/bin.096](https://doi.org/10.1002/bin.096).
- Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: an introduction. *Journal of School Psychology, 52*(1), 1–12. doi:[10.1016/j.jsp.2013.11.009](https://doi.org/10.1016/j.jsp.2013.11.009).
- *Shearer, D. D., Kohler, F. W., Buchan, K. A., & McCullough, K. M. (1996). Promoting independent interactions between preschoolers with autism and their nondisabled peers: an analysis of self-monitoring. *Early Education and Development, 7*, 205–220. doi:[10.1027/s15566935eed0703_1](https://doi.org/10.1027/s15566935eed0703_1).
- *Shogren, K., Lang, R., Machalicek, W., Rispoli, M., & O'Reilly, M. (2011). Self- versus teacher management of behavior for elementary school students with Asperger syndrome: impact on classroom behavior. *Journal of Positive Behavior Interventions, 13*, 87–96. doi:[10.1177/1098300710384508](https://doi.org/10.1177/1098300710384508).
- Skiba, R. J., Casey, & Center, B. (1985). Nonaversive procedures in the treatment of classroom behavior problems. *The Journal of Special Education, 19*(4), 459–481. doi:[10.1177/002246698501900409](https://doi.org/10.1177/002246698501900409).
- *Soares, D., Vannest, K., & Harrison, J. (2009). Computer aided self-monitoring to increase academic production and reduce self-injurious behavior in a child with autism. *Behavioral Interventions, 24*, 171–183. doi:[10.1002/bin.283](https://doi.org/10.1002/bin.283).
- *Stahmer, A. C., & Schreibman, L. (1992). Teaching children with autism appropriate play in unsupervised environments using a self-management treatment package. *Journal of Applied Behavior Analysis, 25*, 447–459. doi:[10.1901/jaba.1992.25-447](https://doi.org/10.1901/jaba.1992.25-447).
- *Strain, P. S., & Danko, C. D. (1995). Caregivers' encouragement of positive interaction between preschoolers with autism and their siblings. *Journal of Emotional and Behavioral Disorders, 3*(1), 2–12. doi:[10.1177/106342669500300101](https://doi.org/10.1177/106342669500300101).
- *Strain, P. S., Kohler, F. W., Storey, K., & Danko, C. D. (1994). Teaching preschoolers with autism to self-monitor their social interactions: an analysis of results in home and school settings. *Journal of Emotional and Behavioral Disorders, 2*, 78–88. doi:[10.1177/106342669400200202](https://doi.org/10.1177/106342669400200202).
- *Todd, T., & Reid, G. (2006). Increasing physical activity in individuals with autism. *Focus on Autism and Other Developmental Disabilities, 21*, 167–176. doi:[10.1177/10883576060210030501](https://doi.org/10.1177/10883576060210030501).
- *Todd, T., Reid, G., & Butler-Kisber, L. (2010). Cycling for students with ASD: self-regulation promotes sustained physical activity. *Adapted Physical Activity Quarterly, 27*, 226–241.
- *Wehmeyer, M. L., Yeager, D., Bolding, N., Agran, M., & Hughes, C. (2003). The effects of self-regulation strategies on goal attainment for students with developmental disabilities in general education classrooms. *Journal of Developmental and Physical Disabilities, 15*(1), 79–91. doi:[10.1023/a:1021408405270](https://doi.org/10.1023/a:1021408405270).