Special Education Outcomes and Young Australian School Students: A Propensity Score Analysis Replication*

Ian Dempsey and Megan Valentine
University of Newcastle, Australia

Using a second cohort of Australian school students, this study repeated the propensity score analysis reported by Dempsey, Valentine, and Colyvas (2016) that found that 2 years after receiving special education support, a group of infant grade students performed significantly less well in academic and social skills in comparison to matched groups of students who did not receive support. Using Longitudinal Study of Australian Children data, the present study found that the second cohort of students with additional needs also performed less well than matched groups of peers and that these results also held true for the specific subgroup of these children with learning disability/learning problems. The ramifications of these results to the delivery of special education in Australia are discussed.

Keywords: research methods, measurement, statistics, academic assessment, content area assessment, instruction

Over the last 40 years, researchers in special education have made substantial contributions in developing and testing a variety of instructional approaches and organisational structures to help children with additional needs to better achieve at school (Hanley-Maxwell & Bottge, 2006). Notwithstanding these contributions and changes in both legislation and increased awareness of inclusion of students with special needs (Dempsey, 2014), substantial numbers of these students encounter poor postschool outcomes. The research consistently shows that children who are unable to show proficiency in basic academic and social skills are at considerable risk of ongoing limitations in their future opportunities (National Early Childhood Technical Assistance Center, 2011; Schaeffer, Petras, Ialongo, Poduska, & Kellam, 2003). However, as Kauffman and Lloyd (2011) correctly note, statistical and mathematical realities mean that there will always be a group of school students who perform substantially less well than their peers.

Challenges in Assessment of Special Education Outcomes

Regardless of these statistical certainties, interest in the efficacy of expensive public education programs, including special education programs, has increased in recent times. Identification and use of evidence-based practice (Slavin, 2002) and the development of

Correspondence: Ian Dempsey, Special Education Centre, University of Newcastle, Callaghan, NSW 2308, Australia. Email: Ian.Dempsey@newcastle.edu.au
*This manuscript was accepted under the Editorship of Umesh Sharma.

 © The Author(s) 2017 |
doi: 10.1017/jse.2017.1

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

implementation fidelity (Fixsen, Blasé, Metz, & Vandyk, 2013) have emerged as potential strategies to maximise student outcomes. Allied with this, measurement of national student academic outcomes has continued apace in most developed countries, including Australia. In this country, the National Assessment Plan – Literacy and Numeracy (NAPLAN; Australian Curriculum, Assessment and Reporting Authority, 2015) allows broad longitudinal conclusions to be reached at school, state, and national levels with regard to students' academic skills. However, there are several reasons why NAPLAN does not permit conclusions to be reached about the academic skills of students with additional needs. NAPLAN results cannot be disaggregated into groups of students with and without additional needs, and a substantial proportion of students with additional needs do not take NAPLAN tests (Dempsey & Davies, 2013).

These logistical problems with national assessment are compounded by the difficulties associated with evaluating the effectiveness of special education via large-scale experimental studies. Randomised control trials (RCTs), regarded as the gold standard in efficacy research, are generally impossible to run in special education contexts because of the diversity of characteristics of students with additional needs and, more importantly, because the withholding of access to special education support for a control group of students will be unethical and likely illegal. Often the best that can be achieved in special education experimental research is to draw conclusions about the effectiveness of an approach for a group of participants in a particular situation (Carter & Wheldall, 2008).

These limitations have not prevented some researchers from attempting to determine the effectiveness of special education. However, virtually all existing studies have substantial methodological flaws that include the lack of adequately matched treatment and control groups, groups matched on a limited range of covariates, and a reliance on cross-sectional designs. A further limitation in this area is that replication research is relatively rare in special education (Travers, Cook, Therrien, & Coyne, 2016). Although written three decades ago, Tindal's (1985) remark still holds: 'The only conclusion that can be made at this time is that no conclusion is yet available about special education efficacy. ... Without sound and valid methodology, the question of effectiveness is simply not worth asking' (p. 109).

Propensity Score Analysis

The difficulties of conducting RCTs are not limited to education and special education. Other human sciences also experience such problems, and so in the last 15 years many researchers have turned to propensity score analysis (PSA) to reduce the imbalance between important covariates (selection bias) and to allow contrasts between naturally occurring experimental and control groups. The control group is a subset of the untreated groups who display very similar likelihoods of experiencing the intervention because of their observed characteristics (Austin, 2011). First proposed by Rosenbaum and Rubin (1984), PSA is a procedure intended to provide an unbiased estimate of treatment outcomes by reducing the confounding effects of covariates and consequentially increasing confidence that differences in dependent variables across groups are due to the treatment. Fundamental to PSA is the calculation of the propensity score for all participants. The propensity score is '... the conditional probability of receiving the treatment given the observed covariates' (Rosenbaum, 2002, p. 296). A wide range of covariates with known relationships with the treatment should be used in this calculation. When participants are grouped into treatment and non-treatment groups (e.g., children receiving and not receiving special education support), then a logistic regression analysis (with treatment as the dependent

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

variable and covariates as independent variables) allows the probability of treatment to be saved from the analysis. This probability serves as the propensity score for each participant.

The next step in PSA involves matching participants who did and did not receive treatment on their propensity scores. A number of different matching methods are available to be used with the goals of matching participants with adequately similar propensity scores and either eliminating or substantially reducing significant imbalance in covariates across matched groups. The final step in PSA uses standard bivariate and multivariate analyses to assess the magnitude of differences in effect across treatment and control groups (Caliendo & Kopeinig, 2008).

There has been conjecture that PSA offers no substantial advantages over traditional multivariate regression methods (Stürmer, Schneeweiss, Avorn, & Glynn, 2003). However, reviews demonstrate that, when the conditions warrant, PSA should be the preferred method (Glynn, Schneeweiss, & Stürmer, 2006). Winklemayer and Kurth (2004) noted that '... if the outcome is rare relative to the number of confounders and the number of study subjects in the smaller exposure group is sufficiently large to warrant multivariable PS estimation, then this statistical technique has a ... role to potentially reduce bias' (p. 1673).

PSA Studies in Special Education

Several PSA studies related to special education have been published in the last decade and each is now briefly reviewed. In the first of these, Morgan, Frisco, Farkas, and Hibel (2010) used PSA to develop two matched groups of students from the Early Childhood Longitudinal Study who were receiving ($n = 363$) and not receiving ($n = 5{,}995$) special education services in schools in the United States (US). Propensity scores were derived from 35 covariates and students' school placements ranged from regular classrooms with assistance, to brief class withdrawal from the regular class, and to special school placement. The results of this study were not consistent across study outcomes. Special education services made either a negative or a statistically nonsignificant improvement on children's learning and behaviour. However, special education services did provide a small, positive effect on children's learning-related behaviours (i.e., remaining attentive, persistence at tasks, being organised).

The second study also made use of the same early childhood US longitudinal database with a younger cohort ($N = 8{,}000$), the Early Childhood Longitudinal Study – Birth Cohort (Sullivan & Field, 2013). Over 30 covariates were used with propensity score weighting methods to generate two matched groups of young children who either did or did not receive special education services. The results demonstrated that receiving special education support had significant moderate negative effects on children's reading and mathematics skills.

The final PSA study in special education is that reported by Dempsey, Valentine, and Colyvas (2016), which used the Longitudinal Study of Australian Children database. Eight different PSA matching methods were used with young school children receiving ($n = 291$) and not receiving ($n = 1{,}926$) special education services. Again, students' school placements included regular classrooms with assistance, brief class withdrawal from the regular class, and special school placement. Across all eight matching methods, the group of children receiving special education assistance performed significantly less well than their matched peers not receiving such support in literacy, numeracy, and their behavioural and social skills. The effect size of this difference ranged from large to small across outcome measures.

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

Taken together, these three studies suggest that special education services may not be bringing expected benefits to children over and above the benefits they might experience in regular classrooms without special education support. However, the pool of PSA studies in special education is very small and the studies cover educational jurisdictions with quite different legislative and school delivery systems, which makes generalisation of these results imprudent at this time. A further limitation of the existing work in this area is that the studies report outcomes for all children receiving special education support. Although conclusions about the efficacy of special education for the total group of children receiving those services may be helpful, it does not permit conclusions to be reached about the effectiveness of special education for some groups of students with additional needs or for students receiving special education support across different settings.

The present study sought to make a contribution to our limited knowledge base on PSA and special education by replicating the study by Dempsey et al. (2016) with a second cohort of school students from the same database. The second goal of the study was to determine if the average treatment effects (ATEs) of special education support for the specific subgroup of students with learning disability/learning problems (students with literacy and/or numeracy problems but without a diagnosis of developmental disability) were consistent with the ATE of all students with special needs.

Method

Participants

The Longitudinal Study of Australian Children (LSAC) began recruitment in 2004 of over 10,000 children and their families and teachers in a stratified random sample from the Medicare (national healthcare system) database. The first wave of data collection involved approximately equal numbers of children in two cohorts of 0–1 (birth cohort) and 4–5 years of age (kindergarten cohort). LSAC has collected data from participants every 2 years and later data collection waves are planned. The purpose of LSAC is to permit examination of the interaction between a variety of social and environmental variables and childhood development (Australian Institute of Family Studies, 2015a). Information on children's physical and mental health, their education, and social, cognitive, and emotional development is being collected from parents, carers, and teachers, and from the children themselves. Specific detail on overall response rates and response rates from subpopulations are available in several LSAC technical papers (Australian Institute of Family Studies, 2015b).

In this paper we reported on data collected during the period mid-2004 (Wave 1) to mid-2012 (Wave 5). In particular, this paper relates to the birth cohort of study children (SC) who were 8 or 9 years of age in 2012. The SC included in this research were those reported as receiving or not receiving special education support in 2010 and for whom data were available for all included covariates and for all 2012 measures of children's learning and behaviour (n ranged from 1,835 to 1,857 depending on the outcome measure).

Study Outcome Measures

LSAC data is collected by parent interview, parent questionnaire, SC interview, and teacher questionnaires (Australian Institute of Family Studies, 2015c). The four outcome measures used in the present research were two measures of child learning (literacy and numeracy), and two measures of the child's social/emotional development (behaviour problems and prosocial skills). These four measures were completed by the teacher of the SC.

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

The literacy and numeracy measures were an adapted version of the Academic Rating Scale (ARS) that was developed for the US Early Childhood Longitudinal Study, Kindergarten Cohort (National Center for Education Statistics, 2008). There were 10 Wave 5 literacy items of increasing complexity that included 'contributes relevant information to classroom discussion' and 'able to write sentences with more than one clause'. The eight numeracy items ranged from 'can continue a pattern with three items' to 'uses a variety of strategies to solve maths problems'. Teachers rated each SC for each skill on a 5-point scale from *not yet displayed* to *proficient*. Rasch-modelled literacy and numeracy scores, which are standardised measures taken at Wave 5 (2012), were used in this paper. Higher scores indicated higher academic skills.

The language and literacy section of the ARS has a moderate correlation (.34) with the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007) and the correlation between the numeracy and the literacy sections was high (.82) in LSAC Wave 3. Internal reliability (Cronbach's α) of both components of the ARS ranged from .95 to .97 in Wave 3 (Australian Institute of Family Studies, 2005).

The study measures of social and emotional rating of behaviour were derived from the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). The SDQ is a widely used 25-item scale with good psychometric properties (Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004; Hawes & Dadds, 2004). The instrument subscales measure the level of conduct behaviour problems, difficulties with peer relationships, hyperactivity, and emotional difficulties, as perceived by the teacher. These are typically combined into a total SDQ score that is a measure of the extent of behavioural difficulties, with higher scores indicating a higher level of behaviour problems. A further subscale measures a range of prosocial or appropriate behaviour with higher scores showing a higher level of prosocial skills. The prosocial skills scale and the total SDQ score measured at Wave 5 (2012) were the behaviour measures outcomes used in this study.

Receiving Special Education Support

The Wave 4 (2010) teacher questionnaire included an item, 'Does this child receive any specialised services provided within the school because of a diagnosed disability or additional need?' A SC was regarded as receiving special education support in 2010 if their teacher responded 'yes' to this question. Given that some of the children in this cohort were in their second year of school, this special education support may have been provided to some children for over 12 months.

A subsequent question asked, 'What is the main reason that this child requires additional assistance or specialised services to enable them to succeed in the regular school program?' The 11 response options for this item were intellectual and physical disability; hearing, sight, or speech/language impairment; learning disability/learning problems in reading and maths; emotional/behavioural problems; poor understanding of Australian English/ESL; autism spectrum disorder; and giftedness. In the Australian context, children with poor understanding of Australian English/ESL and gifted students are not regarded as having special needs (Foreman, 2014). Consequently, these students ($n = 28$) were excluded from the overall treatment group (ALL group $n = 257$) but were included in the contrast group of children.

As the second goal of the present study was to compare results for different groups of SC with special needs, a separate treatment group was identified comprising only SC with learning disability/learning problems in reading or maths (LD group $n = 148$). The second set of PSA analyses compared the LD group with a matched group not receiving

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

additional support. The 109 SC from the ALL group who received additional support for intellectual and physical disability, hearing, sight, or speech/language impairment, emotional/behavioural problems, or autism spectrum disorder were excluded from this analysis. The rationale for this approach relied on several facts about students with learning disability/learning problems. In comparison to other special education needs groups, there is a relatively high incidence of students with learning disability/learning problems (Banks & McCoy, 2011; Dempsey & Davies, 2013), and their support needs are quantitatively and qualitatively different to other students with additional needs, such as students with intellectual, physical and sensory disability, students with autism spectrum disorder, and students with emotional and behavioural problems (Lane, Carter, Pierson, & Glaeser, 2006).

The nature of data collected by the teacher and parent questionnaires in Waves 4 and 5 did not permit further differentiation of services. For example, the LSAC database does not allow meaningful conclusions to be made about the location of delivery of special education services (i.e., regular classroom, segregated classroom within a regular school, or a special school), or the duration or intensity of that support (e.g., teacher aide assistance, short-term withdrawal from class).

Predictors of Special Education Services

Twenty-two covariates were considered for inclusion in the present study to model a child's likelihood (propensity) to receive special education support in 2010. A combination of theoretical studies (Kavale, 1988; van Kraayenoord & Elkins, 2004) and empirical research (Delgado & Scott, 2006; Donovan & Cross, 2002; Louden et al., 2000) was used to select variables associated with the use of special education services. Child and family demographic variables were SC gender and age, birth weight, whether birth was premature or was a multiple birth, whether the SC had repeated a school year, a physical health index, whether the child had a medical condition or disability for at least 6 months in 2006, the remoteness of the family home, and socioeconomic status. Parental variables were whether the primary parent was living with a partner, the extent of the parent's school involvement and frequency of homework checking, how far they thought their child would go with their education, and the parent's level of alcohol consumption. Also included in this category was the parent's age, their English-language proficiency, whether they were of Aboriginal or Torres Strait Islander origin, and two measures of their parenting skills: angry and consistent parenting scales. Finally, the 2010 teacher variables of teacher qualifications and years of teaching experience were added as covariates. See Table 1 for a full list of these covariates. With the exception of teacher experience, teacher qualifications, whether the SC had repeated a year of school, and the extent of parent–school involvement, all covariate data were collected from the primary parent (typically the mother) of the SC.

PSA Procedure

The PSA procedures used in this study are those recommended by Guo and Fraser (2010) and are consistent with those reported in Dempsey et al. (2016). In brief, PSA was used to examine the effect on each of four outcome measures of receiving additional services for the two groups of SC (ALL and LD) receiving special education (treated) in comparison to matching groups of SC who did not receive those services (not treated).

Raw LSAC data were screened for missing data and working datasets created using SAS/STAT Version 9.3. Bivariate relationships between receiving special services and each of the potential covariates were tested using SPSS Version 21 (see Table 1). This approach

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

TABLE 1

Comparison of ALL Treated and Non-Treated Students at Age 6 ($N = 1,856$)

	ALL SC receiving special services ($n = 254$)	SC not receiving special services ($n = 1,602$)	p
SC demographic and health covariates			
SC gender (male %)	62.6	47.6	< .001*
SC age (*M* years)	5.9	5.9	.357
Premature birth (yes %)	5.8	6.4	.749
Multiple birth (yes %)	3.5	3.8	.814
SC has disability (yes %)	5.4	4.2	.381
SC repeated a year (yes %)	5.8	1.4	< .001*
Socioeconomic (*z* score)	− .10	.18	< .001*
Remoteness %			.829
Accessible	77.8	79.5	
Moderately accessible	19.1	17.6	
Remote	3.1	2.9	
Physical Outcome Index (*z* score)	98.2	101.3	.002*
Birth weight (*z* score)	− .08	.03	.111
Parental variables			
Parent age (*M*)	37.4	37.5	.822
Parent has partner (yes %)	89.9	92.0	.264
Parent school involvement (none %)	4.7	2.4	.041*
How far SC will go with education %			< .001*
Under 12 years school	3.1	0.7	
12 years school	25.3	14.7	
Trade qualification	23.7	12.2	
University qualification	47.9	72.5	
Frequency of homework check/help %			.062
Monthly or less	0.4	1.2	
Weekly	27.2	33.5	
Daily	72.4	65.2	
Parent heavy alcohol consumption (yes %)	5.1	5.0	.961
Consistent parenting (*M*)	4.28	4.35	.163
Angry parenting (*M*)	1.98	1.95	.823
Parent ATSI status (%)	2.3	1.1	.106
English language proficiency			
Teacher variables			
SC teacher qualifications %			.433
Master's degree or higher	5.1	5.1	
Graduate diploma or bachelor's	81.7	78.8	
degree	13.2	16.1	
Diploma or certificate			
SC teacher experience (*M* years)	15.7	16.6	.134

Note. SC = study children; ATSI = Aboriginal and Torres Strait Islander.
*Indicates differences are statistically significant.

identified which covariates were statistically associated with receiving services (are imbalanced and are therefore potential sources of selection bias).

Before beginning PSA analyses, *t* tests and OLS regression were conducted for each of the outcome measures to provide comparisons with the different methods of PSA that

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

followed. All PSA methods calculated the ATE, which is a measure of the differences in outcomes for those SC who received the special services compared to a 'corresponding' set of SC who did not receive the services (on an 'intention to treat' basis; Guo & Fraser, 2010, p. 47). The method of calculating the propensity score and its appropriate matching process determine this corresponding set of students.

PSA methods assume that the distribution of the propensity scores overlap each other and therefore share sufficient common scores or a common support region (overlap assumption) from which to draw matching SC. Each of the PSA methods utilised used all or most of the treated SC and a selection of the untreated SC according to the rules and assumptions of the individual methods and the options to trim a percentage of the SC with the 'weakest' matching (Guo & Fraser, 2010). Unless stated otherwise, PSA procedures were conducted using Stata (Stata Corporation, 2011).

The first step in the PSA analysis, estimation of the conditional probability of receiving special services, was conducted by logistic regression in order to specify the functional form of the covariate for the propensity score model. The propensity score was then calculated using the logit of the probability. Matching (resampling) was conducted using four greedy matching procedures: nearest neighbour with callipers of 0.25 standard deviations of the propensity score and 0.1, and Mahalanobis distances with and without propensity score (PSA methods 1–4). No higher order or interaction terms were considered for these first four PSA methods. Postmatching analysis utilised the *t* test on these matched SC to calculate the ATE of the special services intervention.

A fifth PSA method used Generalised Boosted Modelling (GBM) in Stata to create the propensity score followed by various optimal matching techniques and postmatching analysis via the Hodges–Lehmann aligned rank test. The key advantage of this GBM regression tree method is that the functional form of the covariates or interactions do not need to be specified, but are tested within the modelling process up to order 4 interaction (Guo & Fraser, 2010, p. 143). Several boosted regression models were created using different proportions of training data, and Test *R* squared was used to determine the most appropriate model. GBM does not estimate regression coefficients, but calculates the relative predictive influence of each covariate. Highly influential covariates would indicate imbalance between the treated and untreated SC in this multivariate regression. The optimal matching procedures were conducted in *R* (R Core Team, 2013).

Optimal matching reduces the chance of poor matching where the propensity score difference between matched subjects is large, increases the chance of desirable matching where the difference is minimised (Rosenbaum, 1989), and so may be more robust against violations of overlap (Guo & Fraser, 2010, p. 213). The Stata *imbalance* command was used to evaluate whether the optimal matching balanced an observed covariate between those receiving (treated) or not receiving special services (untreated) and to calculate the ATE and Cohen's *d* effect size. The Hodges–Lehmann aligned rank test used the *hodgesl* Stata command (Guo & Fraser, 2010, p. 18) to gauge statistical significance. Calculation of confidence intervals is not included in this procedure.

All of the above five methods are three-step propensity score analyses (i.e., calculation of propensity score, followed by appropriate matching techniques, and postmatching analyses). Common support regions (overlap) may or may not cover the whole range of study participants, but the key objective is to make the two groups of participants (those receiving special services and those who are not) as much alike as possible in terms of their estimated propensity score.

Depending on the extent to which covariate bias was reduced or eliminated with these five PSA methods, consideration was then given to using three additional methods. In the

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

first of these methods, the propensity scores are used as sampling weights to improve the representativeness of treated and non-treated SC (McCaffrey, Ridgeway, & Morral, 2004). The seventh and eighth potential methods of PSA used kernel-based matching estimators to conduct a latent matching, using nonparametric local linear (Heckman, Ichimura, & Todd, 1998, p. 131) and Epanechnikov kernel regression (Guo & Fraser, 2010, p. 255).

Results

Following data screening, the parent variable of proficiency in spoken English was dropped from further analyses because of a large proportion of missing cases. Table 1 shows the relationships between the remaining 21 covariates and SC receiving and not receiving additional support for the ALL group. For this special education needs group there were six covariates with significant associations with receiving support. These covariates were unbalanced and would likely lead to selection bias. For the ALL group, SC who received special education services at age 6/7 were more likely to be male, have repeated a year, come from a lower socioeconomic status, and have a lower physical health index, had less parent school involvement, and lower parent expectations on how far they would progress their education. Although not shown in Table 1 for reasons of conciseness, for the LD group, SC receiving additional support were more likely to be male, have repeated a school year, have a lower physical health index, and have parents with lower expectations about their child's education. The presence of these covariates with statistically significant associations with treatment showed that there was substantial imbalance of covariates in the dataset and that without adequate matching procedures any attempts to determine the effectiveness of special education services could be biased.

There were also significant differences between the LD group and the group of SC receiving support who did not have LD across all the outcome measures, except numeracy, at the start of the intervention in 2010. The LD group ($M = 2.63$, $SD = 0.45$) had lower literacy skills than the non-LD group, $M = 2.90$, $SD = 0.71$, $t(249) = -3.61$, $p < .001$, $d = 0.45$, although there was no significant difference in their maths skills. The non-LD group had higher levels of behaviour problems ($M = 12.11$, $SD = 6.98$) than the LD group, $M = 8.48$, $SD = 5.25$, $t(249) = 4.71$, $p < .001$, $d = 0.59$. Finally, SC with LD had higher scores on the measure of prosocial behaviour ($M = 7.40$, $SD = 2.05$) than the non-LD group of SC, $M = 6.10$, $SD = 2.71$, $t(249) = 4.31$, $p < .001$, $d = 0.54$.

The initial step in the first four PSA analyses reported here was the calculation of the propensity score by logistic regression. Figure 1 shows the box plots of the propensity score distributions demonstrating considerable overlap in propensity scores for SC receiving (treated) and not receiving (non-treated) special education support, for the ALL ($n = 254$) and the LD ($n = 147$) groups, and for whom all literacy outcome data and covariate data were available. As there were only small differences in propensity score distribution for the four outcomes considered, just the literacy distribution is presented here.

The next step in PSA analyses was the completion of the four greedy matching and boosted regression methods detailed earlier in the paper. As at least two of the nearest neighbour greedy matching methods consistently removed all bias from the dataset for all four outcome variables and for the ALL and LD groups, additional PSA analyses (i.e., the methods using propensity scores as weights and kernel regression) were not conducted. However the PSA that utilised the boosted regression to calculate the propensity score, followed by the optimal matching procedures was conducted. For each of the outcomes of literacy, numeracy, behaviour problems, and prosocial skills measured at age 8, Tables 2 to 5 report the number of ALL and LD children for whom data were available in the groups

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

TABLE 2

Differences in Literacy Skills for SC Participants Across ALL and LD Groups at Age 8 for PSA and Non-PSA Methods

	ALL group				LD group			
Method	Received support $n = 255$	No support $n = 1,602$	Magnitude and type of effect $N = 1,857$	Degree of covariate imbalance	Received support $n = 147$	No support $n = 1,602$	Magnitude and type of effect $N = 1,749$	Degree of covariate imbalance
Non-PSA methods								
t test	$M = 3.00$, $SD = 0.76$	$M = 3.80$, $SD = 0.74$	ATE = –0.80, $p < .001$, CI [–0.89, –0.70], $d = 1.07$	6 covariates	$M = 2.91$, $SD = 0.70$	$M = 3.80$, $SD = 0.74$	ATE = –0.88, $p < .001$, CI [–1.01, –0.76], $d = 1.23$	4 covariates
Linear regression			ATE = –0.64, $p < .001$, CI [–0.74, –0.54]	5 covariates			ATE = –0.74, $p < .001$, CI [–0.86, –0.62]	7 covariates
PSA methods								
1. Nearest neighbour calliper .25 *SD*	$n = 254$	$n = 254$	ATE = –0.70, $p < .001$, CI [–0.84, –0.57], $d = 0.92$	None	$M = 2.91$, $SD = 0.70$, $n = 147$	$M = 3.66$, $SD = 0.83$, $n = 147$	ATE = –0.75, $p < .001$, CI [–0.93, –0.58], $d = 0.98$	None
2. Nearest neighbour calliper .1	$n = 250$	$n = 250$	ATE = –0.69, $p < .001$, CI [–0.83, –0.56], $d = 0.90$	None	$M = 2.91$, $SD = 0.70$, $n = 147$	$M = 3.66$, $SD = 0.83$, $n = 147$	ATE = –0.75, $p < .001$, CI [–0.93, –0.58], $d = 0.98$	None
3. Mahalanobis covars as logistic regression	$n = 221$	$n = 221$	ATE = –0.63, $p < .001$, CI [–0.77, –0.49], $d = 0.83$	How far SC will progress education	$M = 2.91$, $SD = 0.72$, $n = 132$	$M = 3.66$, $SD = 0.77$, $n = 132$	ATE = –0.75, $p < .001$, CI [–0.93, –0.57], $d = 1.01$	None
4. Mahalanobis covars as logistic regression + propensity score	$n = 221$	$n = 221$	ATE = –0.63, $p < .001$, CI [–0.78, –0.49], $d = 0.83$	How far SC will progress, SC physical health	$M = 2.91$, $SD = 0.71$, $n = 132$	$M = 3.66$, $SD = 0.77$, $n = 132$	ATE = –0.88, $p < .001$, CI [–1.01, –0.76], $d = 1.01$	None
5. Boosting and optmatch (5SVM3)	$n = 255$	$n = 825$	ATE = –0.55, $p < .001$, $d = 0.84$	None	$n = 147$	$n = 544$	ATE = –0.69, $p < .001$, $d = 1.04$	Teacher experience

Note. SC = study children; LD = learning disability/learning problems; PSA = propensity score analysis; ATE = average treatment effect; CI = 95% confidence interval.

TABLE 3

Differences in Maths Skills for SC Participants Across ALL and LD Groups at Age 8 for PSA and Non-PSA Methods

	ALL group				LD group			
Method	Received support $n = 253$	No support $n = 1,582$	Magnitude and type of effect $N = 1,835$	Degree of covariate imbalance	Received support $n = 145$	No support $n = 1,582$	Magnitude and type of effect $N = 1,727$	Degree of covariate imbalance
Non-PSA methods								
t test	$M = 3.07$, $SD = 0.80$	$M = 3.74$, $SD = 0.76$	ATE = –0.67, $p < .001$, CI [–0.78, –0.57], $d = 0.86$	6 covariates	$M = 2.99$, $SD = 0.71$	$M = 3.74$, $SD = 0.74$	ATE = –0.75, $p < .001$, CI [–0.88, –0.62], $d = 1.00$	4 covariates
Linear regression			ATE = –0.55, $p < .001$, CI [–0.65, –0.45]	10 covariates			ATE = –0.64, $p < .001$, CI [–0.76, –0.51],	8 covariates
PSA methods								
1. Nearest neighbour calliper .25 *SD*	$n = 252$	$n = 252$	ATE = –0.57, $p < .001$, CI [–0.71, –0.44], $d = 0.70$	None	$n = 141$	$n = 141$	ATE = –0.68, $p < .001$, CI [–0.86, –0.50], $d = 0.92$	None
2. Nearest neighbour calliper .1	$n = 249$	$n = 249$	ATE = –0.58, $p < .001$, CI [–0.71, –0.44], $d = 0.73$	None	$n = 141$	$n = 141$	ATE = –0.68, $p < .001$, CI [–0.86, –0.50], $d = 0.92$	None
3. Mahalanobis covars as logistic regression	$n = 221$	$n = 221$	ATE = –0.59, $p < .001$, CI [–0.74, –0.44], $d = 0.75$	How far SC will progress	$n = 132$	$n = 132$	ATE = –0.67, $p < .001$, CI [–0.85, –0.48], $d = 0.88$	None
4. Mahalanobis covars as logistic regression + propensity score	$n = 221$	$n = 221$	ATE = –0.61, $p < .001$, CI [–0.76, –0.46], $d = 0.76$	How far SC will progress	$n = 132$	$n = 132$	ATE = –0.65, $p < .001$, CI [–0.83, –0.46], $d = 0.85$	None
5. Boosting and optmatch	$n = 253$	$n = 781$	ATE = –0.51, $p < .001$, $d = 0.66$	Teacher experience	$n = 145$	$n = 398$	ATE = –0.66, $p < .001$, $d = 0.90$	None

Note. SC = study children; LD = learning disability/learning problems; PSA = propensity score analysis; ATE = average treatment effect; CI = 95% confidence interval.

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

TABLE 4

Differences in Behaviour for SC Participants Across ALL and LD Groups at Age 8 for PSA and Non-PSA Methods

	ALL group				LD group			
Method	Received support $n = 257$	No support $n = 1,604$	Magnitude and type of effect $N = 1,861$	Degree of covariate imbalance	Received support $n = 148$	No support $n = 1,604$	Magnitude and type of effect $N = 1,752$	Degree of covariate imbalance
Non-PSA methods								
t test	$M = 9.62$, $SD = 6.91$	$M = 5.36$, $SD = 5.42$	ATE = 4.26, $p < .001$, CI [3.52, 5.00], $d = 0.69$	6 covariates	$M = 7.95$, $SD = 5.56$	$M = 5.36$, $SD = 5.42$	ATE = 2.58, $p < .001$, CI [1.67, 3.50], $d = 0.47$	4 covariates
Linear regression			ATE = 3.47, $p < .001$, CI [2.73, 4.20]	6 covariates			ATE = 2.01, $p < .001$, CI [1.12, 2.91]	6 covariates
PSA methods								
1. Nearest neighbour calliper .25 *SD*	$n = 256$	$n = 256$	ATE = 3.38, $p < .001$, CI [2.26, 4.50], $d = 0.53$	None	$n = 147$	$n = 147$	ATE = 2.27, $p < .001$, CI [1.03, 3.50], $d = 0.42$	None
2. Nearest neighbour calliper .1	$n = 254$	$n = 254$	ATE = 3.48, $p < .001$, CI [2.36, 4.61], $d = 0.54$	None	$n = 147$	$n = 147$	ATE = 2.27, $p < .001$, CI [1.03, 3.50], $d = 0.42$	None
3. Mahalanobis covars as logistic regression	$n = 223$	$n = 223$	ATE = 3.14, $p < .001$, CI [1.94, 4.34], $d = 0.49$	How far SC will progress	$n = 133$	$n = 133$	ATE = 1.24, $p < .001$, CI [0.81, 2.56], $d = 0.23$	None
4. Mahalanobis covars as logistic regression + propensity score	$n = 223$	$n = 223$	ATE = 4.26, $p < .001$, CI [1.96, 4.35], $d = 0.49$	How far SC will progress	$n = 133$	$n = 133$	ATE = 1.56, $p < .001$, CI [0.21, 2.90], $d = 0.28$	None
5. Boosting and optmatch (5VM3)	$n = 257$	$n = 795$	ATE = –3.01, $p < .001$, $d = 0.56$	None	$n = 148$	$n = 540$	ATE = –1.32, $p < .001$, $d = 0.29$	Consistent parenting

Note. SC = study children; LD = learning disability/learning problems; PSA = propensity score analysis; ATE = average treatment effect; CI = 95% confidence interval.

TABLE 5

Differences in Prosocial Skills for SC Participants Across ALL and LD Groups at Age 8 for PSA and Non-PSA Methods

	ALL group				LD group			
Method	Received support $n = 257$	No support $n = 1,604$	Magnitude and type of effect $N = 1,861$	Degree of covariate imbalance	Received support $n = 148$	No support $n = 1,604$	Magnitude and type of effect $N = 1,752$	Degree of covariate imbalance
Non-PSA methods								
t test	$M = 7.16$, $SD = 2.42$	$M = 7.98$, $SD = 2.10$	ATE = –0.81, $p < .001$, CI [–1.09, –0.53], $d = 0.36$	6 covariates	$M = 7.66$, $SD = 2.13$	$M = 7.98$, $SD = 2.10$	ATE = –0.31, $p = .042$, CI [–0.04, 0.67] $d = 0.15$	4 covariates
Linear regression			ATE = –0.51, $p < .001$, CI [–0.79, –0.24]	8 covariates			ATE = –0.11, $p < .001$, CI [–0.45, –0.23]	5 covariates
PSA methods								
1. Nearest neighbour calliper .25 *SD*	$n = 256$	$n = 256$	ATE = –0.41, $p = .0542$, CI [–0.01, 0.83], $d = 0.18$	None	$n = 147$	$n = 147$	ATE = –0.07, $p = .3945$, CI [–0.43, 0.57] $d = 0.03$	none
2. Nearest neighbour calliper .1	$n = 254$	$n = 254$	ATE = –0.43, $p = .0447$, CI [–0.01, 0.85], $d = 0.20$	None	$n = 147$	$n = 147$	ATE = –0.07, $p = .395$, CI [–0.43, 0.57] $d = 0.03$	none
3. Mahalanobis covars as logistic regression	$n = 223$	$n = 223$	ATE = –0.48, $p = .0295$, CI [0.05, 0.91], $d = 0.20$	How far SC will progress	$n = 133$	$n = 133$	ATE = –0.02, $p = .480$, CI [–0.52, 0.55] $d = 0.05$	none
4. Mahalanobis covars as logistic regression + propensity score	$n = 223$	$n = 223$	ATE = –0.46, $p = .0369$, CI [0.03, 0.90], d=0.19	How far SC will progress	$n = 133$	$n = 133$	ATE = –0.31, $p = .804$, CI [–0.47, 0.60] $d = 0.01$	none
5. Boosting and optmatch (5VM3)	$n = 257$	$n = 795$	ATE = –0.46, $p = .003$, $d = 0.36$	None	$n = 148$	$n = 435$	ATE = –0.15, $p = .32$, $d = 0.04$	none

Note. SC = study children; LD = learning disability/learning problems; PSA = propensity score analysis; ATE = average treatment effect; CI = 95% confidence interval.

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

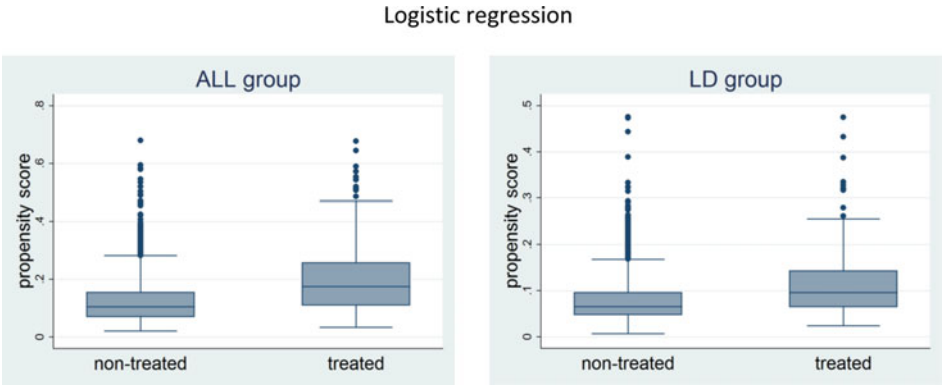

FIGURE 1

(Colour online) Box plots of PS Distributions From Logistic Regression for ALL Treated ($n = 255$) and LD Treated Groups ($n = 147$) Versus Non-Treated ($n = 1{,}602$) for the Literacy Outcome.

receiving or not receiving special education services. For the five different PSA techniques used, the tables also report estimates of the effect of the special services intervention (ATE), effect size, and confidence intervals (with the exception of boosted regression methods).

Literacy

For the ALL group, there were 1,857 observations with 255 of these SC receiving special education support. The *t* test with six unbalanced covariates likely overestimated the difference between the two groups' literacy skills. Both nearest neighbour PSA methods eliminated covariate bias, and the remaining greedy matching methods (models three and four) had one and two covariates showing imbalance. The boosted regression and optimal matching method showed that physical health, birth weight, and how far parents thought their child would progress with their education were the most influential covariates. However, the variable method using Hansen's equation (5VM3) removed bias from all influential covariates. Across all PSA methods, the students receiving treatment scored around 0.6 points lower on literacy skills than their matched peers not receiving assistance (moderate effect size).

Results for the LD group ($n = 147$) were similar with the *t* test, again likely biased. All covariate imbalance was eliminated with both the nearest neighbour and the Mahalanobis methods and, for the 5VM3 optimal matching procedure, one covariate remained imbalanced (teacher experience). The LD group was about 0.7 points lower on literacy skill scores than their peers; a large effect size.

Numeracy

With regard to ALL maths skills ($n = 253$), the ATE from the *t* test had six imbalanced covariates and is therefore likely to be overestimated. Two of the greedy matching methods removed covariate bias, and overall children receiving treatment scored 0.6 points lower on numeracy skills than their matched peers, a result reflected by the boosted method. This was a moderate effect size.

Similarly for the LD group, all covariate bias was eliminated using the greedy methods and the boosted methods, and the maths ATE was about 0.7 points less for the LD group than their matched group. Again, this was a moderate treatment effect.

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

Behaviour

There were 1,861 ALL SC included in the analysis and 257 children received support. Again, the ATE from the *t* test had six imbalanced covariates and is therefore likely to be overestimated. The two nearest neighbour methods eliminated all covariate bias, and the remaining greedy methods retained one unbalanced covariate (how far parents thought their child would academically progress). Overall, children in the ALL group had behaviour problems about 3 points higher than the matched group, and this was a moderate effect size.

Every greedy matching PSA method removed all or most covariate bias. On average, the LD group's behaviour was about 2 points higher (worse behaviour) than SC not receiving special education services — a small effect size.

Prosociality

There were 257 ALL SC receiving assistance and 1,604 children not receiving support. The *t* test, with six imbalanced covariates, showed the ALL group with significantly lower prosocial scores but with a small effect size. Both nearest neighbour greedy PSA methods removed covariate bias and showed SC receiving special education assistance scored, on average, 0.4 points less than matched SC not receiving special education assistance.

For the LD group ($n = 148$), all five matching methods eliminated covariate bias. However, the ATE estimates for the SC with LD were not significantly different from their matched peers.

Discussion

The research reported here had two goals. The first objective was to replicate the PSA analysis completed by Dempsey et al. (2016) with a different cohort of children. However, as covariate bias was eliminated by using the nearest neighbour and boosted regression PSA methods, the additional methods of using propensity scores as weights and kernel regression were not required in the current study. The second goal was to check if the ATEs for four outcome variables for the group of SC with LD were broadly consistent with the ATEs of the total group of SC receiving special education support.

With regard to the first goal, the results of the present study were consistent with those reported earlier by Dempsey and colleagues (2016). The total group of children receiving additional support (ALL group) performed less well in literacy, numeracy, behaviour, and prosocial skills in comparison to a matched group of SC not receiving support. Logistic regression and *t* tests estimates of ATE used unbalanced covariates and are therefore likely to have overinflated differences between the groups. All propensity score analyses gave statistically significant differences between treated and untreated groups. At least two of the greedy matching methods eliminated covariate bias and the effect size of the unbiased ATEs ranged from small to large depending on the outcome measure under consideration.

There were several findings in relation to the second goal. The LD group of children receiving additional support performed significantly less well than their matched peers in literacy, numeracy, and behaviour outcomes across the biased *t* test and regression analyses, as well as the propensity score analyses. Covariate bias was eliminated in at least four of the five propensity score analyses. For prosocial skills, there was no significant difference between the two groups. All these results mean that, for both the ALL and the LD groups, the provision of special education support appeared to provide no benefits in

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

terms of improvements in their academic skills and behaviour in comparison to matched peers who did not receive additional support.

In conjunction with the three other studies using PSA methods to examine the effectiveness of special education that were reviewed in the introduction, this study has assisted in building a consistent evidence base that special education may not be providing the outcomes expected of it in Australia and the US. The word 'may' is used judiciously here because there are some limitations in the research design used by Dempsey et al. (2016) and in the present study that need to be acknowledged. The mid-year timing of LSAC data collection and that some of the SC were in their second year of school in 2010 means that children in the treatment groups had been receiving special education services for varying lengths of time. Furthermore, the LSAC database does not provide detailed information on the intensity or type of special education support provided. For example, it is not possible to draw conclusions about the extent to which the duration, location, and intensity of special education services may be associated with the research findings. A final limitation is that low cell counts for some groups of students with additional needs (e.g., students with hearing impairment, visual impairment, or with physical disability) did not permit differential analysis for these groups of SC. Given the specialised equipment and technologies used with these children in special education settings, it may be that additional supports do indeed provide demonstrable benefits for these children over and above what they may receive in the regular classroom.

Regardless of these limitations, the present findings must be of concern for special education professionals and for educational administrators. Beyond evidence that special education teaching strategies and technologies are effective in highly controlled environments, in cross-sectional studies, or in longitudinal studies with biased comparison groups (Kavale & Dobbins, 1993), the discipline of special education lacks confirmation that it is effective for the majority of students with additional needs (Carter & Wheldall, 2008).

In a helpful discussion of stages of programs of educational research, Sam Odom and colleagues (2005) note that such research logically progresses through four steps. First, preliminary ideas, hypotheses, and pilot studies; second, controlled laboratory and classroom-based experiments; third, randomised trials; and finally, informed classroom practice. The difficulty of conducting RCTs in special education no doubt explains why special education research has largely bypassed the third step in its goal to improve outcomes for students with additional needs. Nevertheless, without evidence from control trials (or from quasi-experimental methods, such as PSA, that control for bias), then claims about the effectiveness of special education for the majority of students with additional needs are unjustified.

At the moment, special education in developed countries is maintained by philosophical arguments and legislative requirements (Foreman, 2014) rather than by an evidence base that maintains human professions such as medicine, nursing, and the sciences. Special education is not alone in lacking solid research evidence that it is effective, over and above what may be provided in regular settings. For example, a variety of social programs in the criminal justice system lack a sound research base (Australian Institute of Health and Welfare, 2013). However, without evidence that the special education system is effective then the profession leaves itself open to accusations that special education acts as little more than a form of respite for regular education.

The statistical realities identified by Kauffman and Lloyd (2011) mean that there will always be a group of students in the education system that perform substantially less well than their peers. However, the research reviewed and reported in this paper suggests that the current special education services provided to these Australian students delivers poorer

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

outcomes than regular classroom teaching. There is a range of potential explanations for this situation. It may be that the relatively poorer outcomes in special education settings is related to the skill base of special educators and the ineffectiveness of preservice and inservice special education training. The inconsistent fidelity of special education support within and across schools may also contribute to the apparent ineffectiveness of special education. It could also be the case that the quality of teaching in special and in regular education settings is little different to each other. A final possible explanation may be that the administrative and support structures in special education offer no advantages over those that exist in regular schools.

Acknowledgements

This report makes use of data from Growing Up in Australia: the Longitudinal Study of Australian Children (LSAC). LSAC is conducted in partnership with the Department of Social Services (DSS), the Australian Institute of Family Studies (AIFS), and the Australian Bureau of Statistics (ABS), with advice provided by a consortium of leading researchers. Findings and views expressed in this publication are those of the individual authors and may not reflect those of the AIFS, DSS, or the ABS.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424. doi:10.1080/00273171.2011.568786

Australian Curriculum, Assessment and Reporting Authority. (2015). *NAPLAN*. Retrieved from http://www.nap.edu.au/naplan/naplan.html

Australian Institute of Family Studies. (2005). *Summarising children's wellbeing: The LSAC Outcome Index* (LSAC Technical Paper No. 2). Retrieved from http://www.growingupinaustralia.gov.au/pubs/technical/tp2.pdf

Australian Institute of Family Studies. (2015a). *Growing Up in Australia: The Longitudinal Study of Australian Children*. Retrieved from http://www.growingupinaustralia.gov.au/

Australian Institute of Family Studies. (2015b). *Technical papers*. Retrieved from http://www.growingupinaustralia.gov.au/pubs/technical/index.html

Australian Institute of Family Studies. (2015c). *Study questionnaires*. Retrieved from http://www.growingupinaustralia.gov.au/studyqns/index.html

Australian Institute of Health and Welfare. (2013). *Diverting Indigenous offenders from the criminal justice system* (Resource Sheet No. 24). Retrieved from http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129545614

Banks, J., & McCoy, S. (2011). *A study on the prevalence of special educational needs* (National Council for Special Education Research Report No. 9). Retrieved from http://ncse.ie/wp-content/uploads/2014/10/Prevalence_of_SEN_10_09_12.pdf

Becker, A., Woerner, W., Hasselhorn, M., Banaschewski, T., & Rothenberger, A. (2004). Validation of the parent and teacher SDQ in a clinical sample. *European Child & Adolescent Psychiatry*, *13*(Suppl. 2), ii11–ii16. doi:10.1007/s00787-004-2003-5

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Carter, M., & Wheldall, K. (2008). Why can't a teacher be more like a scientist? Science, pseudoscience and the art of teaching. *Australasian Journal of Special Education*, *32*, 5–21. doi:10.1080/10300110701845920

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

Delgado, C. E. F., & Scott, K. G. (2006). Comparison of referral rates for preschool children at risk for disabilities using information obtained from birth certificate records. *The Journal of Special Education*, *40*, 28–35. doi:10.1177/00224669060400010301

Dempsey, I. (2014). Legislation, policies and inclusive practices. In P. Foreman & M. Arthur-Kelly (Eds.), *Inclusion in action* (4th ed., pp. 47–72). South Melbourne, Australia: Cengage.

Dempsey, I., & Davies, M. (2013). National test performance of young Australian children with additional educational needs. *Australian Journal of Education*, *57*, 5–18. doi:10.1177/0004944112468700

Dempsey, I., Valentine, M., & Colyvas, K. (2016). The effects of special education support on young Australian school students. *International Journal of Disability, Development and Education*, *63*, 271–292. doi:10.1080/1034912X.2015.1091066

Donovan, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education*. Washington, DC: National Academy Press.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test – Fourth Edition*. San Antonio, TX: Pearson.

Fixsen, D., Blasé, K., Metz, A., & Vandyk, M. (2013). Statewide implementation of evidence-based programs. *Exceptional Children*, *79*, 213–230.

Foreman, P. (2014). Introducing inclusion in education. In P. Foreman & M. Arthur-Kelly (Eds.), *Inclusion in action* (4th ed., pp. 2–46). South Melbourne, Australia: Cengage.

Glynn, R. J., Schneeweiss, S., & Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, *98*, 253–259. doi:10.1111/j.1742-7843.2006.pto_293.x

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *38*, 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Hanley-Maxwell, C., & Bottge, B. A. (2006). Reconceptualizing and recentering research in special education. In C. F. Conrad & R. C. Serlin (Eds.), *The Sage handbook for research in education: Engaging ideas and enriching inquiry* (pp. 175–195). Thousand Oaks, CA: Sage. doi:10.4135/9781412976039.n10

Hawes, D. J., & Dadds, M. R. (2004). Australian data and psychometric properties of the Strengths and Difficulties Questionnaire. *Australian and New Zealand Journal of Psychiatry*, *38*, 644–651. doi:10.1080/j.1440-1614.2004.01427.x

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, *65*, 261–294. doi:10.1111/1467-937X.00044

Kauffman, J. M., & Lloyd, J. W. (2011). Statistics, data, and special education decisions: Basic links to realities. In J. M. Kauffman & D. P. Hallahan (Eds.), *Handbook of special education* (pp. 27–36). New York, NY: Taylor & Francis.

Kavale, K. A. (1988). Learning disability and cultural-economic disadvantage: The case for a relationship. *Learning Disability Quarterly*, *11*, 195–210. doi:10.2307/1510764

Kavale, K. A., & Dobbins, D. A. (1993). The equivocal nature of special education interventions. *Early Child Development and Care*, *86*, 23–37. doi:10.1080/0300443930860103

Lane, K. L., Carter, E. W., Pierson, M. R., & Glaeser, B. C. (2006). Academic, social, and behavioral characteristics of high school students with emotional disturbances or learning disabilities. *Journal of Emotional and Behavioral Disorders*, *14*, 108–117. doi:10.1177/10634266060140020101

Louden, W., Chan, L. K. S., Elkins, J., Greaves, D., House, H., Milton, M., … van Kraayenoord, C. E. (2000). *Mapping the territory — Primary students with learning difficulties: Literacy and numeracy. Volume 2: Analysis*. Canberra, Australia: Department of Education, Training and Youth Affairs.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425. doi:10.1037/1082-989X.9.4.403

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*, 236–254. doi:10.1177/0022466908323007

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press

National Center for Education Statistics. (2008). *Academic Rating Scale*. Retrieved from http://nces.ed.gov/ECLS/kinderinstruments.asp

National Early Childhood Technical Assistance Center. (2011). *The outcomes of early intervention for infants and toddlers with disabilities and their families*. Retrieved from http://www.nectac.org/~pdfs/pubs/outcomesofearlyintervention.pdf

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*, 137–148. doi:10.1177/001440290507100201

R Core Team. (2013). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer. doi:10.1007/978-1-4757-3692-2

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032. doi:10.1080/01621459.1989.10478868

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524. doi:10.1080/01621459.1984.10478078

Schaeffer, C. M., Petras, H., Ialongo, N., Poduska, J., & Kellam, S. (2003). Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology*, *39*, 1020–1035. doi:10.1037/0012-1649.39.6.1020

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. doi:10.3102/0013189X031007015

Stata Corporation. (2011). Stata release 11 [Computer software]. College Station, TX: Author.

Stürmer, T., Schneeweiss, S., Avorn, J., & Glynn, R. J. (2003). Determinants of use and application of propensity score (PS) methods in pharmacoepidemiology. *Pharmacoepidemiology & Drug Safety*, *12*, S121–S122.

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology*, *51*, 243–260. doi:10.1016/j.jsp.2012.12.004

Tindal, G. (1985). Investigating the effectiveness of special education: An analysis of methodology. *Journal of Learning Disabilities*, *18*, 101–112. doi:10.1177/002221948501800209

Travers, J. C., Cook, B. G., Therrien, W. J., & Coyne, M. D. (2016). Replication research and special education. *Remedial and Special Education*, *37*, 195–204. doi:10.1177/0741932516648462

van Kraayenoord, C. E., & Elkins, J. (2004). Learning difficulties in numeracy in Australia. *Journal of Learning Disabilities*, *37*, 32–41. doi:10.1177/00222194040370010401

Winklemayer, W. C., & Kurth, T. (2004). Propensity scores: Help or hype? *Nephrology, Dialysis, Transplantation*, *19*, 1671–1673. doi:10.1093/ndt/gfh104

https://doi.org/10.1017/jse.2017.1 Published online by Cambridge University Press