Taylor & Francis
Taylor & Francis Group

Journal of Intellectual and Developmental Disability

ISSN: 1366-8250 (Print) 1469-9532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjid20

Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice

Shane T. Harvey, Diana Boer, Luanna H. Meyer & Ian M. Evans

Journal of Intellectual & Developmental Disability, March 2009; 34(1): 67–80

informa
healthcare

LITERATURE REVIEW

Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice*

SHANE T. HARVEY[1], DIANA BOER[2], LUANNA H. MEYER[2] & IAN M. EVANS[1]

[1]*Massey University, Palmerston North and Wellington, New Zealand, and* [2]*Victoria University of Wellington, New Zealand*

Abstract

Background This meta-analysis of interventions with challenging behaviour in children with disabilities updates a comprehensive meta-analysis that previously addressed reported standards of practice and effectiveness of different strategies.

Method Four effect-size algorithms were calculated for published intervention cases, and results analysed and compared to previous findings by behaviour target, intervention type, and other factors.

Results The evidence largely supports intervention effectiveness, with some inconsistency reflecting the fact that the four metrics assess different aspects of change. Skills replacement, consequence combined with systems change, and antecedent interventions generated selective positive results, large enough to be clinically meaningful.

Conclusions Behavioural interventions effectively reduce challenging behaviour, particularly when preceded by a functional analysis. Teaching replacement skills was most effective, especially if used in combination with systems change and/or traditional antecedent and consequence manipulation. Positive changes as well as enduring limitations to both research design and standards of clinical practice in comparison to 18 years ago are discussed.

Keywords: *challenging behaviour, behavioural intervention, standards of practice, children and youth, meta-analysis*

Introduction

There is strong evidence that challenging behaviour in children and youth with developmental disabilities both interferes with quality of life and predicts future negative outcomes (Emerson, 2003; Murphy et al., 2005). Disruptive, dangerous, life-threatening, inappropriate, and socially undesirable behaviours present major difficulties for family, peer, and other community relationships. These behaviours can also represent significant challenges to professionals responsible for providing rehabilitative and other educational services. Intervention with difficult behaviour as early as possible is an agreed priority for those supporting the best possible quality of life for individuals with intellectual disabilities (Carr, Horner, et al., 1999).

Eighteen years ago, the first comprehensive meta-analysis of intervention research focused on problem behaviour in persons with developmental disabilities appeared (Scotti, Evans, Meyer, & Walker, 1991). In contrast to traditional literature reviews relying on expert judgment, meta-analyses of large samples of published studies provide an objective aggregation of outcomes as a function of intervention variables. Of all meta-analyses conducted on studies involving some aspect of developmental disability, the Scotti et al. study has been ranked among the highest in terms of including the domains of information necessary for ensuring face valid results concerning outcomes (Mostert, 2001).

The 1991 report also investigated the evidence with regard to issues and controversies prominent at the time. These included standards of research practice, such as reporting relevant independent variables and client information, use of aversive interventions, availability of alternative positive interventions, degree to which treatments were associated with unintended side-effects, and clinical

*This manuscript was accepted under the Editorship of Roger J. Stancliffe.

Correspondence: Professor Ian M. Evans, School of Psychology, Massey University, Private Box 756, Wellington, 6140, New Zealand.
E-mail: i.m.evans@massey.ac.nz

ISSN 1366-8250 print/ISSN 1469-9532 online © 2009 Australasian Society for the Study of Intellectual Disability Inc.
DOI: 10.1080/13668250802690922

significance or meaningfulness of reported benefits (Helmstetter & Durand, 1991; Meyer & Evans, 1993; Voeltz & Evans, 1982). Discrepancies were highlighted between accepted best-practices and what was actually being published in the literature, purportedly representing the highest standard of validated intervention research reports disseminated by key international journals. These inconsistencies included the finding that only a small proportion of reports reported a functional analysis as part of the assessment process (e.g., Carr & Durand, 1985). Scotti et al. (1991) argued that significant improvements were needed in research and publication standards, leading to seven recommendations toward ensuring that future published interventions could provide the solid evidence-base needed for evaluating treatment strategies and verifying particular interventions or treatment protocols. Of special importance, they emphasised, was the need to report collateral change (both positive and negative behaviours), and the need for longer baseline and intervention phase data. Despite these limitations, however, the overarching conclusion was that the available empirical evidence supported the effectiveness of treatments based on principles of behaviour modification for reducing serious challenging behaviour in individuals with an intellectual disability (Chambless et al., 1996).

At the time of that review, major developments were already underway in educational programs and supports for children and youth with disabilities, including increased emphasis on individualised education, school inclusion, family support, and early intervention. Better lifestyles and services for young people might be expected to contribute to reductions in those severe behaviour problems previously associated with institutionalisation. Also emerging in 1991 were a few examples of non-aversive, educative approaches with good reported outcomes (cf. Evans & Meyer, 1985; Meyer & Evans, 1989), and the implementation and evaluation of more holistic packages such as Positive Behaviour Support (Carr, Dunlap, et al., 2002) and positive parenting programs (Stepping Stones Triple P; Sanders, Mazzucchelli, & Studman, 2004) has risen dramatically in the intervening years. Progress in the validation of effective interventions with the full range of social and emotional needs should now be providing authoritative guidance to the field regarding the best practices and the degree of benefit in the treatment of challenging behaviour.

Given these developments, regular updates on evidence of effectiveness would seem to be important, and more recent meta-analyses have been conducted. Carr, Horner, and their colleagues (1999) carried out a quantitative analysis of intervention outcomes, but their review did not assess effect sizes. Subsequently, the same group of investigators (Marquis et al., 2000) used the identical literature base of 109 articles published between 1985 and 1996, calculated effect sizes, and conducted a formal meta-analysis. They concluded that positive behaviour support was effective, however, they coded articles only in terms of whether they used "stimulus based" interventions or "positive reinforcement".

Didden, Duker, and Korzilius (1997) published a detailed meta-analytic study on interventions with problem behaviours, one which these authors described as designed to rectify certain limitations of the Scotti et al. (1991) review. This study drew on a wider range of journals and sorted both problem behaviours and treatment procedures into a larger number of discrete categories. However, the authors used only one effect-size metric (PND, percentage of non-overlapping data, adjusted for the occurrence of data points of zero). Despite this limitation, the study suggested that response contingent behavioural interventions are more effective than other types of treatment including medication, but that externally destructive behaviours are less successfully treated than internally maladaptive or socially disruptive behaviours. Other findings which replicated those of the Scotti et al. study were that conducting a prior functional analysis of the challenging behaviour resulted in better outcomes and that only a small percentage of published studies (20%) produced outcomes that could be rated as highly effective.

More recently, Didden, Korzilius, van Oorsouw, and Sturmey (2006) focused a new meta-analysis on studies involving persons with mild mental retardation (mostly children and youth), for whom a wider variety of psychotherapeutic interventions were used. They calculated the two effect-size metrics used by Scotti et al. (1991): PND and percentage of zero data (PZD). Again, treatments based on a functional analysis produced significantly better outcomes, and behavioural interventions (including antecedent control, differential reinforcement, and functional communication training) were more effective than cognitive or self-management approaches. Similar findings were reported by Campbell (2003), reviewing outcome studies for persons with autism. Mathur, Kavale, Quinn, Forness, and Rutherford (1998) reported a meta-analysis of the effects of social skills interventions on emotional and behavioural problems. Based on the PND metric these authors found that social skills instruction alone was only mildly or questionably effective and even less effective for students with autism.

While there is some convergence of findings from these various meta-analyses, there is a degree of ambiguity caused by different methods, metrics, and strategies for coding variables. Given recent widespread changes in intervention approaches and improvements in service settings and attitudes, we felt there was a need to replicate as systematically as feasible the 1991 comprehensive meta-analysis. Furthermore, with the ever-growing emphasis on evidence-based practice in health and education, government agencies have supported new reviews to provide guidance to public policy and spending for children and youth in particular (e.g., Meyer & Evans, 2006). Young people are more likely to have experienced recently improved educational services, and more likely to have been the recipients of new holistic and naturalistic intervention packages relevant to families and schools. We also investigated results associated with different statistical effect size algorithms (metrics) that have been developed and discussed in related literature. In addition to evaluating the effectiveness (validity) of interventions featured in the recent research literature, we were interested in whether progress had been made regarding standards of practice, as these are reflected in the design and reporting of research studies. The present review, therefore, focuses on the relative effectiveness of different intervention approaches for changing what sorts of challenging behaviours, in which contexts, and for which children and youth having developmental disabilities, and how these results are being obtained and reported.

Method

Inclusion criteria

English-language research reports published between January 1988 and mid-2006 were identified according to the following criteria: (a) participants were diagnosed with a developmental disability and exhibited a challenging behaviour; (b) participants were aged birth to 21 years; (c) the focus of the report was on psychological intervention (i.e., behavioural, educational, psychotherapeutic) rather than medication as the sole treatment; (d) data were reported through formal observation assessments; and (e) the data were suitable for calculation of effect size (see below). Studies had to include independent data sets as recommended by Rosenthal (1995) whenever they reported data across several participants.

Literature search

Relevant articles were located by searching recognised databases as well as 22 specific journals known to publish this type of research (see listing in Table 1).

Consistent with the recommendations of Hunter and Schmidt (2004) and Lipsey and Wilson (2001), the search included review articles, references within studies, bibliographic databases, and making contact with 20 leading experts. A librarian specialising in database administration searched the Educational Resources Information Clearinghouse (ERIC) and the Psychological Literature (PsycLit) databases.

Key terms used to search both specific journals and databases encompassed: (a) symptom terms (such as behaviour disorder, aggression, self-injury, self-stimulation); (b) diagnostic terms and disability labels (such as autistic disorder, pervasive developmental disorders, brain damage, mental retardation); and (c) intervention and treatment terms (such as cognitive behaviour therapy, applied behaviour analysis, family therapy, social skills training). The search initially located 1,086 journal articles, for which the full citation and abstract were printed. Abstracts were then checked against the review criteria by two of the authors working independently. This reduced the original pool to 680 potential articles, and we were able to source 635 of the articles in time for the project (45 could not be accessed through any available inter-library loan

Table 1. Specific journals searched

American Journal on Mental Retardation
Behavioral Disorders
Behavior Modification
Behaviour Research and Therapy
Behavior Therapy
Child & Family Behavior Therapy
Disability and Rehabilitation
Education & Training in Developmental Disabilities
Exceptional Children
Intellectual and Developmental Disabilities (formerly *Mental Retardation*)
International Journal of Disability, Development and Education
Journal of Abnormal Child Psychology
Journal of Applied Behavior Analysis
Journal of Autism and Developmental Disorders
Journal of Behavior Therapy and Experimental Psychiatry
Journal of Consulting and Clinical Psychology
Journal of Experimental Child Psychology
Journal of Intellectual & Developmental Disability
Journal of Positive Behavior Interventions
Journal of Special Education
Research & Practice for Persons with Severe Disabilities (formerly *Journal of the Association for Persons with Severe Handicaps*)
*Research in Developmental Disabilities*1

procedure). These articles were read in full by at least one of the authors, resulting in a further 436 being eliminated because they did not meet all of the inclusion criteria. Of the 199 articles remaining at this stage, a further 57 had to be excluded because their data proved unsuitable for the calculation of effect sizes; for example, 13 of these were excluded because they contained less than the minimum number of 3 data points across each of the baseline and treatment phases. The final database included 142 articles reporting studies with 316 individuals (299 single case studies and one group study including 17 participants). The list of articles included in the meta-analysis is available from the authors.

To assess agreement on the selection of articles according to the criteria, a person not involved in the research at any stage chose a random sample of approximately 25% ($N = 36$) of the articles retained for the meta-analysis and an equal number of the articles excluded due to the absence of data needed for the analyses. This person then combined the set of articles in random order, and one of the co-authors who had not been involved in the original process sorted the 72 articles according to the inclusionary and exclusionary criteria. The rater agreed with the original coder's decision that all 72 articles met the initial selection criteria and that the 36 articles subsequently excluded did not in fact include data suitable for use in the meta-analysis sample (100% agreement for each decision).

In the 1991 meta-analysis, the study was the basic unit of analysis. For the present review we used the individual participant as the primary element of analysis where individualised data were available. As treatments are designed to suit the needs of the individual, it seems reasonable that the individual be the focus of analysis, not the article. From the articles analysed, 305 individual participants were available for the meta-analysis.

Coding

Reports (individuals) were coded according to the study characteristics and moderating variables utilised by Scotti et al. (1991). However, changes in the context and priorities now being reported in the literature required additional codes. Coding was carried out independently by two of the authors, with any discrepancies resolved by team discussion including the other two authors. Studies were coded and grouped according to the following:

(A) *Participant variables*: age, gender, primary diagnosis, secondary diagnosis, target behaviour, behaviour severity, intellectual (functioning) level, sensory impairment, motor impairment, communication ability, and previous intervention. The wide variety of individual target behaviours could be conveniently classified under the following categories: self-injurious, aggressive, destructive, stereotypic, inappropriate social, and disruptive behaviours. These categories are very similar to those being used in recent epidemiological studies (e.g., Lowe et al., 2007). Three levels of severity of any of these behaviours were identified, adapted from Scotti et al (1991; based on Meyer & Evans, 1989), from the least to the most severe. Level 1 was chronic behaviour little changed over time but likely to interfere with community acceptance. Level 2 was more serious behaviour that was likely to increase in severity if left untreated, was a priority concern for caregivers, and/or interfered with learning. Level 3 was behaviour that was health or life threatening and/or dangerous to others, requiring immediate, urgent attention.

(B) *Setting and context*: primary treatment setting (home, school, community, hospital, clinical lab), secondary treatment setting, mainstream (inclusion) context, intervention agent (staff/teacher, mental health professional, parent, sibling, peer), family involvement, and peer involvement. Duration of treatment was divided into 6 bands, the shortest band being 1–3 weeks and the longest band 20 weeks or more.

(C) *Treatment:* treatment strategy, level of intervention intrusiveness, use of formal (not necessarily experimental) functional analysis, medication, and use of restraints. Treatment strategies were further analysed by (i) modification of antecedents/stimulus triggers; (ii) teaching or promoting alternative replacement skills; (iii) contingency management (reward, punishment, extinction); (iv) systems change, in which the whole service context for the individual is modified, including for example changes in placement and the introduction of new activity contexts presumed to be age-appropriate and normalised – see, for example, McClean and Grey (2007); (v) two or more treatment conditions implemented (with and without systems change); (vi) use of aversives; and (vii) treatment design (number and type of phases and reversals). Level of intervention intrusiveness was categorised from 1 through 6, including: Level 1, the least intrusive, involving ecological (environmental) changes, modifying task difficulty, positive social and material rewards, redirecting; Level 2, including extinction, brief restraint, social disapproval, within-room time-out, removal of desired objects; Level 3 involved overcorrection, contingent exercise, time-out involving removal from the room; Level 4 included visual screening, mandatory relaxation, time-out in a restraining room; Level 5 was mechanical restraint, application of noxious stimuli; and Level 6 slapping, pinching, electric shock, food deprivation, and noxious stimuli (Scotti et al., 1991).

(D) *Practicality*: this category included variables such as cost and duration of treatment.

Data analysis

Effect-size algorithms. The choice of effect-size metric in single subject research is a complex issue debated extensively in the literature (Parker & Brossart, 2003; Scruggs & Mastropieri, 1994; Swanson & Sachse-Lee, 2000; White, Rusch, Kazdin, & Hartmann, 1989). Each time a new metric is proposed, its advocate typically emphasises its virtue by pointing out flaws in other metrics, and it is not possible to revisit all of the arguments here (see Busk & Serlin, 1992). Simulations favour the Hierarchical Linear Modeling approach, but this requires very long baselines (Jenson, Clark, Kircher, & Kristjansson, 2007). In practice, much depends on which components of the series of data points are used in the calculations (Hartmann et al., 1980), but the most frequently expressed statistical concern is that data points in a time-series may be autocorrelated (Crosbie, 1995). In fact, most data points in treatment studies are not serially dependent, as challenging behaviour is often recorded days, or even weeks apart (Huitema, 1985). The frequency of an undesirable behaviour on one day is largely independent of its frequency on another day. Although interrupted time-series computer programs such as ITSACORR are still being reported (e.g., Karasu, 2006), Huitema (2004) has made a compelling argument against such methods.

Somewhat more problematic is the issue of trend in the baseline (Van den Noortgate & Onghena, 2003). If an increasing baseline trend in a challenging behaviour is reversed to a decreasing trend during intervention, it can be argued that this is as clinically significant as a larger decrease in absolute level. But the major concern for calculating a meaningful effect size arises if the data points in the baseline show a decreasing linear trend. The intervention could appear to be successful in reducing the behaviour when in fact the behaviour was decreasing anyway. Rather than being strictly a statistical problem, this issue instead reflects a weakness in experimental design: if the baseline period was not long enough to establish a "steady state" (a fixed level of the behaviour), then it is difficult to attribute change unambiguously to the intervention – especially if the study design is basically a simple AB (baseline versus intervention comparison). Since most published clinical studies use these simple designs and since most report only a few baseline data points, the meta-analytic reviewer has only two choices: eliminate studies that have weak designs and short baselines (which would be the majority), or include them and trust that, in most clinical studies worthy of publication, the problematic behaviour was not declining and would not have declined without the intervention. This in turn requires effect-size metrics that can be sensibly calculated with just a few data points.

However, given sufficient recorded data points, it is possible to calculate an effect-size metric that uses regression logic to decrease the possible influence of common trends in the baseline that would otherwise mask the importance of the changes observed during intervention. Allison and Gorman (1993; Gorman & Allison, 1996) suggested a linear regression technique in which baseline data are used to predict values, these are subtracted from the observed data, and the detrended data can then be regressed on treatment and on treatment by time interaction, with the R^2 value then converted to *d*. While Campbell (2004) in a direct comparison study found that regression-based effect-size metric did not improve understanding of treatment outcomes, we were persuaded that where sufficient numbers of data points were reported, one such metric should be calculated. We used the procedure recommended by Allison and Gorman (1993) for all cases in which there were 5 or more data points in both baseline and treatment phases.

Different methods of representing effect size are also strongly argued on statistical grounds, but the reality is that each one represents a slightly different facet of clinical outcome. For example, if the challenging behaviour is very harmful, an outcome that reduces its frequency to zero levels (percentage zero data points) is clinically more meaningful than one that merely reduces observed frequency relative to the lowest frequency seen during baseline (percentage of treatment data points not overlapping with intervention data points). We therefore used different metrics, each one of which has some advantages as well as having some well-publicised disadvantages. Since the present study is an update of the Scotti et al. (1991) analysis, it was important to use the two metrics that were employed in that earlier study. When there are a reasonable number of data points, some of their limitations can be addressed by use of the Standard Mean Difference metric introduced by Busk and Serlin (1992), in which the difference between the mean for the baseline and the mean for the intervention is determined and then divided by the standard deviation of the baseline data. This approach performed very well in a comparative study of sample data by Olive and Smith (2005). The metrics finally adopted and the pros and cons of each are summarised in Table 2.

Effect-size calculations and meaning. Beyond the selection of metrics, important methodological decisions are also needed for how to aggregate these and how to report their meaning in terms of the size of an effect (Parker et al., 2005). We averaged effect-size estimates only within each of the four algorithms selected. Since the four metrics understandably do not correlate highly, findings for each are reported separately, emphasising those for which there was convergence. In this study, findings refer to either the descriptive mean effect sizes for different procedures or, where there were sufficient cell sizes to allow a comparison, comparing interventions and other independent variables using analysis of variance or *t*-test comparisons of means. Towards providing a verbal descriptor of how large any given effect size can be considered to be (i.e., degree of treatment effectiveness), we relied on quartile splits for three of the metrics and an absolute level for PND based on the cumulative frequencies reported by Scotti et al. (1991) requiring a modification of those originally proposed by Scruggs, Mastropieri, and Casto (1987) (see Table 2).

Results

Descriptive statistics: Participants, type of study, and setting

The sample of children and youth had a mean age of 9.7 years ($SD = 4.6$), with 27% between 1 and 5, 32% between 6 and 10, 27% between 11 and 15, and 14% between 16 and 20. Two-thirds of the participants were male. The majority of studies failed to report ethnicity or cultural group, the presence or nature of sensory and motor difficulties, or any secondary diagnosis. Regarding overall functioning level, only communication ability was likely to be reported; 15% were described as having no communication skills, 52% as having limited ability, and less than 2% as having age-appropriate communication skills. Other information on cognitive level and/or academic skills was generally absent. Forty-four percent had a primary diagnosis of mental retardation, followed by autistic disorder (33%), and multiple disabilities (17%). Only 9% of participants were reported to be taking psychotropic medication during the study; other reports were generally silent on this issue. Eighty-five percent of the case studies did not report whether or not there had been previous treatment attempts. Most target behaviours were at Level 2 severity (72%) and included self-injury (33%), destructive (18%), stereotypic (16%), and aggressive (12%) categories. Disruptive and inappropriate social behaviours were the targets in 9% and 10% of the participants respectively.

Treatment designs typically were simple treatment vs. baseline comparisons (AB designs, 76%). In 23% of the studies, there were additional treatment phases added (ABC, 13%; ABCD, 10%). In 20% of the studies, the duration of the treatment was less than 20 weeks; for 75% of the studies, the duration was not reported and could not be calculated from the data provided. It was reported that 37% of interventions occurred in schools, with 72% of secondary contexts occurring in the community. Overall, 71% of the studies occurred in mainstream settings, mostly delivered by professionals (64%), care staff (23%), or parents (11%). Twenty percent of studies reported family involvement in the intervention, while 33% reported peer involvement. Service providers' existing resources were sufficient in 85% of the treatment studies, with treatment costs being supplemented by the researchers in 14% of the cases. Major additional resources were utilised in only 1% of the studies.

Descriptive statistics: Type of intervention

Sixty two percent of all interventions were at Level 1 in terms of intrusiveness, and 29% were at Level 2. Sixteen interventions (about 5%) were at Level 5 and only 3 (about 1%) were at Level 6, the most intrusive level recorded. Only 4 of over 300 cases involved aversives and 6 involved physical restraints. Given few behaviours at the most serious level (16%), the relationship of treatment intrusiveness to behaviour severity was evaluated by collapsing all degrees of intrusiveness into three levels. A cross-tabulation analysis showed that the more severe the behaviour the more likely it was to be treated with more intrusive interventions (χ^2 [4] = 19.2, $p < .01$); for example, 10% of mild behaviours, 6% of moderate behaviours, and 23% of severe behaviours were

Table 2. Features of the effect-size statistics used in the meta–analysis

			Data points used in each phase	Interpretation of effective ranges[e]			
Effect-size algorithm	Definition	Strengths and limitations		Ineffective	Questionable	Fair	Highly effective
Percent Non-overlapping Data (PND)[a]	Change as a result of treatment without indicating absolute magnitude of change.	Easy to calculate Widely used Likely to misrepresent effects when outliers occur in baseline, treatment has a detrimental effect, or there are distinctive trends (slope) in the data	≥3	<50%	50–79%	80–99%	>99%
Percent Zero Data (PZD)[b]	Degree to which behaviour is eliminated in treatment.	Easy to calculate Can be distorted if treatment is terminated immediately after zero data point occurs	≥3	<12%	12–42.9%	43–69.9%	≥70%
Standard Mean Difference (SMD)[c]	An overall estimate of change somewhat corrected for chance.	Easy to calculate Yields an effect-size index that is commonly understood Does not account for trend and autocorrelation	≥3	<.30	.30–.49	.50–.79	≥.80
Allison Mean plus Trend (Allison–MT)[d]	Change from baseline to treatment incorporating linear trends across phases.	More difficult to calculate Can be distorted where baseline phase is long Requires five or more data points	≥5	<.04	.04–.18	.19–.46	≥.47

Notes. [a]Introduced by Scruggs, Mastropieri, & Casto (1987; Scruggs & Mastropieri, 1994). [b]Designed by Scotti et al. (1991). [c]Proposed by Busk & Serlin (1992), transformed to R^2 according to Parker et al.'s (2005) recommendation. [d]Allison & Gorman (1993). [e]For PZD, SMD, and Allison-MT, the four effectiveness ranges represent quartile splits of statistical results; ranges for PND as modified by Scotti et al. (1991).

treated with the more intrusive levels of intervention. No relationship was found between diagnostic groups or specific types of behaviour (aggression, self-injury, etc.) and treatment intrusiveness level.

When delivered as the sole intervention strategy, antecedents were used in 16% of all cases, skills replacement in 5%, and consequences in 26%. As components of multi-element interventions, antecedents were used in 50% of the interventions, skills replacement in 31%, consequences in 75%, and systems change in 15%. Systems change was not used as the sole intervention in any of the studies, but always combined with other interventions.

Meta-analysis: Intervention effect sizes

Self-injury, stereotypy, socially inappropriate, and destructive behaviour typically responded best to behaviourally-based interventions. Disruptive and aggressive behaviour generally responded least well to behaviour change efforts. However, when elimination was the criterion (PZD index), treatments for stereotypy would also have to be judged largely ineffective. The higher the level of severity, the less effective any intervention was in changing behaviour.

For intervention strategies used singly (e.g., an antecedent intervention only), none of the effect sizes indicated highly effective outcomes on any of the four statistics. However, all three treatments of antecedents, skills replacement, and consequences (recall that system change was never used alone) produced effect sizes in the fairly effective range (Table 3). Skills replacement was in the fairly effective range on three of the four statistics, and antecedents on two of the four. Consequences were judged as fairly effective only on the metric that adjusts for prior trend – Allison-MT. When combined with other treatments, antecedent based interventions were not related to significantly better outcomes (Table 3). Skills replacement was consistently associated with superior levels of outcome. Single treatments involving consequences only were associated with modest outcomes, overall yielding small effect sizes. When system change was incorporated, outcomes were better on all statistics except PZD, particularly if more than one other treatment approach was being implemented. Consequences in combination with systems change yielded the largest Allison-MT effect size. Skills training combined with antecedent manipulation produced the greatest absolute decreases in target behaviours as measured by PND and PZD. Incorporating systems change into the intervention was related to better outcomes on all statistics except PZD, where outcomes were fair.

Meta-analysis: Treatment type by target behaviour

We also examined whether there were patterns in which treatment types were associated with significant effects for specific target behaviour categories, across the four effect-size metrics. Since proportionately more cases were reported using treatments with limited results in comparison to fewer cases conducted using treatments showing promising results, this type of analysis might also suggest where further research is needed. We will highlight the most important findings which have to be interpreted with caution at times because of small cell sizes (numbers of cases); a table showing all effect sizes for all treatment/behaviour interactions is available on request.

For self-injurious behaviour (SIB), an antecedent-only intervention was highly effective on PZD (79%) and fairly effective on Allison-MT (.22) for 14 and 12 cases, respectively. In contrast, a skills replacement only intervention was fairly effective in treating SIB according to all four statistics (PZD = 62%; PND = 94%; SMD = .64; Allison-MT = .20), but these results were based on only two cases. Based on a slightly larger sample size of four cases, a combined treatment including systems change for SIB was fairly effective on three of the four statistics (PND = 80%; SMD = .67; Allison-MT = .21). Interestingly, combined treatments that did not involve systems change were reported much more often – 29 cases – and associated with fairly effective results only on Allison-MT (.21). The other treatment used in a large number of cases – 23 cases – was consequences-only, which was also associated with fairly effective results only on Allison-MT (.37).

Combination treatments that did not include systems-change were used most often for aggression ($N = 12$ cases), resulting in PZD (61%), SMD (.51), and Allison-MT (.28) effect sizes that were fairly effective. Consequences-only treatments and combined treatments including systems change were both used in 6 cases, with fairly effective results for only PZD (61%) and Allison-MT (.20) respectively. Only one research report used a skills replacement or antecedent-only treatment for aggression, precluding further analysis. For destructive behaviour, antecedent-only interventions carried out for four cases were highly effective on PZD (71%) and fairly effective on Allison-MT (.35), whereas skills replacement only interventions based on three cases were also highly effective on PZD (77%) as well as being fairly effective on SMD (.69) and Allison-MT (.25). A consequences-only treatment for seven cases resulted in effect sizes that were fairly effective for PZD (46%) and Allison-MT (.31). A combined

Table 3. Comparison of treatment outcomes

			95% CI					95% CI		
	ES	*SD*	lower	Upper		ES	*SD*	lower	Upper	
Type of treatment	Single treatment approach				*N*	In combination with other treatments				*N*
SMD										
Antecedents	.44	.26	.36	.52	38	.44	.25	.39	.50	96
Skills replacement	.57*	.28	.39	.74	13	.52*	.27	.46	.59	68
Consequences	.46	.27	.39	.53	59	.46	.26	.42	.50	156
System change	–	–	–	–	0	.49	.26	.39	.58	30
PZD										
Antecedents	43*	40	30	56	38	43*	33	36	50	96
Skills replacement	52*	22	38	65	13	44*	30	37	51	68
Consequences	43*	31	35	51	61	42	30	37	47	156
System change	–	–	–	–	0	32	30	21	43	30
PND										
Antecedents	60	40	47	73	38	57	41	49	66	96
Skills replacement	79	31	60	98	13	70	37	60	78	68
Consequences	49	43	38	60	61	57	42	50	64	156
System change	–	–	–	–	0	59	44	43	76	30
Allison-MT										
Antecedents	.33*	.28	.24	.43	34	.24*	.28	.18	.29	96
Skills replacement	.38*	.26	.22	.54	12	.26*	.28	.19	.32	68
Consequences	.33*	.28	.25	.40	59	.28*	.30	.23	.32	156
System change					0	.35*	.41	.19	.50	30
Combination treatments and system change										
SMD										
No system change	.46	.26	.41	.52	79					
Includes system change	.49	.26	.39	.58	30					
Single treatment	.47	.26	.42	.53	99					
PZD										
No system change	44*	31	37	51	79					
Includes system change	32	30	21	43	30					
Single treatment	44*	32	38	50	99					
PND										
No system change	60	41	50	69	79					
Includes system change	59	44	43	76	30					
Single treatment	58	41	50	67	99					
Allison-MT										
No system change	.21*	.27	.15	.27	79					
Includes system change	.35*	.41	.19	.50	30					
Single treatment	.32*	.26	.26	.37	99					

Note. * = fairly effective.

treatment including systems change used for destructive behaviour in five cases resulted in one highly significant effect size (Allison-MT = .47), whereas combined treatments that did not include systems changed with destructive behaviour in 18 cases resulted in fair PND (81%) and fair Allison-MT (.28) effect sizes.

There were sufficient numbers of cases where different interventions were reported with stereotypic behaviours, showing fairly effective PZD (44%) and Allison-MT for skills replacement only with five and four cases, respectively. Consequences-only treatments reported for 11 cases of stereotypic behaviour resulted in fairly effective effect sizes for SMD (.61) and Allison-MT (.41), and antecedent-only treatments with 15 cases also showed fairly effective effect sizes for SMD (.53) and Allison-MT (.39). A combined treatment including systems change was reported in four cases but associated with only one significant result: being fairly effective on Allison-MT (.32).

Combined treatments were used most for inappropriate social target behaviours; these were fairly effective whether used without systems change (10 cases: SMD = .54; Allison-MT = .23) or with systems change (8 cases: SMD = .50), but results were highly effective on Allison-MT (.53) with systems change. Finally, for disruptive behaviour there were sufficient numbers of cases for both consequences-only and combined treatments to be examined. Consequences-only treatments were fairly effective on PZD (46%) for eight cases and on Allison-MT (.28) for seven cases. Combined treatments without systems change were fairly effective on PZD (57%) for nine cases. All other cell sizes were quite small, making interpretation questionable.

Meta-analysis: Moderator variables' influence on effect size

The impact of selected moderator variables on effect sizes was also tested through planned analyses. Age, gender, and ethnicity appeared not to influence effect sizes achieved (although ethnicity or cultural identity was almost never reported, so that this interpretation is based on very few cases). There was no consistent pattern for good outcomes to be associated with any particular intervention setting, nor were mainstream contexts consistently related to larger effect sizes. Although most reports (75%) failed to record duration of treatment, for those that did three of the four statistics showed significant difference between time groups (PZD index: $F(4, 48) = 5.2$, $p < .01$; PND index: $F(4, 48) = 2.5$, $p < .05$; Allison-MT index: $F(4, 48) = 2.8$, $p < 0.05$. Post-hoc analysis indicated both very short (1–3 weeks) and long (>20 weeks) interventions appeared less effective than those conducted between 3 to 20 weeks. The use of functional analysis in assessing the target behaviour was associated with more effective outcomes in maintaining a zero rate of behaviour (PZD index: $F(1,217) = 12.18$, $p < .01$).

MANOVAs were conducted comparing effect-size statistics to test further for age and for diagnostic differences in responsiveness to particular interventions. First, three age ranges roughly reflective of developmentally-related changes in educational structures were investigated: early childhood (birth to 8 years), middle childhood (ages 8–12 years), and the adolescent years (ages 13–21 years). We found no meaningful differences in treatment responsiveness between these three age groups. We also investigated whether children diagnosed with autism respond differently to various treatments than do children with other diagnoses. We found better outcomes for children with autism to single treatments involving antecedents in comparison to other children on two of the metrics (SMD: 0.54 vs. 0.36, $t(32) = 2.26$, $p < .05$; Allison-MT: 0.42 vs. 0.17, $t(28) = 2.82$, $p < .01$). The magnitude of these effect sizes indicates medium effectiveness. Contrary to the results for SMD and Allison-MT, however, single treatments involving antecedents were not related to elimination of challenging behaviour in autistic children compared to children with other diagnoses as revealed by PZD (27.74 vs. 71.74, $t(32) = -3.80$, $p < .01$). Further, there were no significant differences in responsiveness to other treatment strategies.

Discussion

The movement towards identifying empirically supported clinical treatments has intensified the importance of establishing the effectiveness of behavioural interventions (Chambless & Ollendick, 2001; Task Force on Psychological Intervention Guidelines, 2002). Our meta-analysis affirmed the findings of Scotti et al. (1991) and other more recent reviews that psychological (behavioural) treatments – compared to no treatment or conditions as usual – can clearly reduce even the most severe challenging behaviours. However, there is no one intervention used alone or in combination with others that was associated with highly effective results for all categories of challenging behaviour, nor was any single behavioural strategy significantly more effective than others.

Manipulating antecedents and consequences resulted in questionable to fairly effective outcomes, depending on the metric used. Systems change could not be evaluated as the sole intervention since it was only ever used in combination with another intervention – generally with the two categories that were themselves most effective as the sole treatment with selected behaviours. That no studies used systems change as the sole intervention probably reflects clinical perspectives that systems change is a context for intervention rather than the intervention itself. The one intervention approach that reliably resulted in higher effect sizes was skills replacement. Skills replacement also outperformed other interventions in being consistently higher across algorithms whether used alone or blended with other interventions. Yet here we must express particular caution, as fewer studies included skills replacement in comparison to other interventions (with the smaller sample size resulting in larger variability and overlap in confidence intervals). Based on our results, it would appear that skills replacement is a major area for future research, especially for use in combination with antecedent and systems change approaches.

It might be expected that different types of challenging behaviour respond differentially to different treatment strategies. To investigate this, we evaluated patterns of intervention effectiveness across different challenging behaviours. Changing antecedents was effective (when the metric was the PZD) in eliminating destructive and self-injurious behaviour. Consequences (contingency management) produced medium reductions in inappropriate social behaviour across most effect-size metrics. Skills replacement resulted in medium to high effect sizes across most metrics for stereotypic, destructive, self-injurious, and inappropriate social behaviour. The inclusion of systems change in combination with another intervention type appeared moderately effective with self-injurious and aggressive behaviour, and emerged as highly effective using the Allison-MT metric with inappropriate social and destructive behaviour (but note that Parker et al., 2005, warn that Allison-MT produces large effect sizes, although that would benefit statistical power).

We can put these findings together with one strong moderating procedural variable: as in Scotti et al. (1991), we found that interventions that were preceded by an assessment involving a formal functional analysis produced larger effects. These analyses were not necessarily ones confirmed by experimental manipulation, but could be derived from careful caregiver observation. Didden and his colleagues (2006) also reported that the use of functional analysis was associated with certain more positive effects. Taken together, these results support multi-component intervention approaches that include as an essential component skills replacement designed to accomplish the function identified for the targeted challenging behaviour, as well as incorporating attention to environmental and systems-change variables. Thus, the empirical literature supports theoretical perspectives advanced for some time acknowledging assessment to determine the functions of challenging behaviour and the importance of replacement strategies to address those functions in the design of effective interventions.

Practical issues for the implementation of effective interventions

Programs that include teaching replacement skills appear most promising across the various types of challenging behaviour, reliably associated with outcomes in the fairly effective range. This finding has important implications for the next phase of developing effective interventions, as the involvement of mediators such as teachers, paraprofessionals, parents, and peers is likely to be crucial for any program designed to teach new skills to a child with disabilities. Although there were not many studies with lengthy baselines or the kinds of experimental controls advocated in the single-subject methodology literature, the studies reviewed did require somewhat sophisticated measurement and professional-level interventions and reporting standards. Not surprisingly, therefore, over 60% of the interventions were conducted by educational and clinical professionals, and 32% were conducted in specialised medical-type settings. Motivating paraprofessionals and non-professionals (family members and peers) to implement systematic interventions may entail more than the traditional viewpoint that sees their role as one of maintenance and generalisation of changes initially established by highly skilled professionals. This might explain why the formal evidence did not reveal any benefit for treatment effects through the involvement of families or peers.

The literature has only begun to deal systematically with issues such as how to engage those mediators present in the child's daily life effectively in the ongoing work needed for teaching new skills and maintaining the conditions for their use. For example, there is evidence that once challenging behaviour has responded well to a "high effort" intervention (such as functional communication training), multiple mediators in different contexts of the child's natural environments can spread positive effects through a "low effort" intervention (such as praising communication attempts) that fits more easily into typical interactions (Schindler & Horner, 2005; see also Barnett, Daly, Jones, & Lentz, 2004).

Re-looking at standards of practice

The 1991 meta-analysis critically examined standards of practice and was concerned at the lack of rigour in some of the published literature. Has this situation changed? Firstly, the 1991 report examined whether it was the most serious and severely challenging behaviours that were being treated with the most intrusive interventions, which was typically the justification given for using intrusive, particularly aversive, interventions. Scotti et al. found that interventions at all level of intrusiveness were being implemented for behaviours at all levels of severity: very intrusive interventions were being applied to very minor problems. This is, pleasingly, no longer the case; overall, the most intrusive interventions are barely evident with less than 2% of the studies reporting the use of aversives or restraints while more than 90% were at the two lowest levels of intrusiveness. This marked change in the pattern of intervention types in the 18 intervening years is one of the most striking findings in the present update. The great aversives debate of the 1980s appears to be over, with aversive interventions no longer allowed by many agencies, schools, advocacy groups, and ethics committees and only rarely reported in the research literature.

Procedurally, Scotti et al. (1991) noted a lack of information about contextual or ecological variables. This has improved, perhaps as a result of increasing awareness of the role of systems change. Scotti et al. were surprised at deficits in reporting important participant characteristics, such as level of functioning, adaptive behaviours (strengths), and ethnicity, and this type of missing information was still very noticeable. We would encourage researchers to incorporate crucial demographic information about participants, especially ethnicity and cultural identity, assessment of communication and other skills, previous interventions, duration of the reported intervention, use of medication, and how mediators were involved across the intervention phases. The majority of studies in this meta-analysis (74%) reported assessment by means of functional analysis, but given its relationship to effectiveness this too could be improved.

We found a high proportion of simple AB experimental designs. This is not in itself problematic, but becomes more of a concern if there are not sufficient baseline and intervention phase data points reported to allow adequate effect-size calculations. While some of the metrics used in this analysis are not totally dependent on longer sets of data points, having few such points introduces highly suspect issues for effect-size calculation, such as lack of steady state in the baseline, the possibility that treatment was stopped early (as soon as an improved result was noted), and inability to detect and thus compensate for trend. If journals require that published studies include the data needed for the calculation of effect sizes, the research community will doubtless ensure intervention reports are underpinned by the appropriate evidence. Scotti et al. (1991) also commented on the growing evidence of behavioural interrelationships that made it important to monitor possible collateral change, both negative and positive. Lilienfeld (2007) has recently raised the issue of treatments that are not only ineffective but can cause harm. Leading empirical journals should expect clinical researchers to consider iatrogenic and other unanticipated collateral effects systematically. We found little evidence in the present review that this has become a major feature of intervention design.

The 1991 meta-analysis found nearly 20% of studies showed no objectively discernable treatment benefit, whereas now we found ineffective results across all four effect-size statistics in only 3 studies, or 1.4% of the sample. However, the data set allowing us to report all four metrics (those with the necessary 5 data points) comprised only two-thirds of the studies. Research reports are still being published that may appear visually to demonstrate a treatment effect but which lack the requisite data needed for the calculation of effect-size algorithms with sufficient statistical power to allow valid conclusions. This is particularly true for the growing collection of more sophisticated statistical approaches that do correct for serial dependency but which require at least 15 data points (e.g., Jenson et al., 2007).

Summing up and moving on

In the present meta-analysis we aggregated the results for more than 50 individual cases (studies) in which interventions with self-injurious behaviours used either consequences-only strategies or combination treatments, but without any systems change. Although self-injurious behaviour is one of the most distressing of the problematic behaviours of young people with intellectual disabilities, these relatively recently reported treatments were rated ineffective on all metrics with the exception of a "fairly effective" rating derived from the regression-based metric. In contrast to this number of studies, we located less than one-third as many cases involving antecedent-only treatments, four cases with combination treatments including systems change, and only two cases with skills replacement only treatments for self-injurious behaviour. This was despite the fact that all three treatment types were associated with at least fairly effective outcomes estimated by three or all four effect-size metrics. It seems obvious that further research is not needed where there is already ample evidence that a treatment approach has limited impact – yet such studies are continuing to be carried out. What should be a priority is the systematic replication of interventions showing more significant and clinically meaningful results with seriously harmful behaviours, such as skills replacement and combination treatments including systems change.

With the emphasis placed in recent years on more holistic methods like positive behavioural support, we wondered whether our review might show an increase in published reports using key aspects of that approach, such as skills replacement and systems change. We examined the 13 articles in our sample reporting 30 cases of individuals that used a combination treatment involving systems change. We found that no more than two were published in any given year over the time period covered. None were published during the final three years covered by this meta-analysis. Thus, contrary to the most recent best practice recommendations, the literature continues to be dominated by traditional antecedent-only or consequence-only single interventions, or combination treatments that do not involve systems change.

Anyone familiar with the contemporary literature on managing challenging behaviour in young people with development disabilities will know that the exemplary models considered to be the best clinical/educational practice are multi-faceted. Models such as Positive Behaviour Support and Triple P combine systems change, antecedent/ecological change, social and family support, and teaching new skills (ideally with the same function as the challenging behaviour), with traditional positive intervention practices such as reward contingencies and reward saturation, and mild negative consequences such as planned ignoring. In other words, we have complex packages of interventions suggesting that research reports based on single intervention approaches are now rather passé. While there is still room for demonstrating the specific benefits of individual procedures, all of the conventional methods of applied behaviour modification will take place against a backdrop of broader behavioural support and of educative approaches (Carr, Dunlap, et al., 2002; Evans & Meyer, 1985). What has been seen until recently as a values-based position is supported objectively by findings from this meta-analysis. We also now have sound evidence in favour of interventions that focus on teaching new adaptive skills as a major treatment strategy. The next 18 years of behavioural intervention research must focus on how these more general packages of interventions can be made still more effective, accessible for mediators, made available for the highest need clients, and do-able in children's typical contexts and everyday environments.

Author note

This research was supported in part by a contract awarded to Victoria University by the New Zealand Ministry of Education. No restrictions have been imposed on free access to or publication of the research data by the funding body, and the authors declare that they have no financial or non-financial conflicts of interest either direct or indirect. An earlier version of portions of this report was presented at the World Congress of Behavioral and Cognitive Therapies, Barcelona, Spain, July 2007.

References

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.

Barnett, D. W., Daly, E. J., Jones, K. M., & Lentz, F. E. (2004). Response to intervention: Empirically based special service decisions from single-case designs of increasing and decreasing intensity. *The Journal of Special Education*, *38*, 66–79.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, *24*, 120–138.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification*, *28*, 234–246.

Carr, E. G., & Durand, V. M. (1985). Reducing behavior problems through functional communication training. *Journal of Applied Behavior Analysis*, *18*, 111–126.

Carr, E. G., Dunlap, G., Horner, R. H., Koegel, R. L., Turnbull, A. P., Sailor, W., et al. (2002). Positive behavior support: Evolution of an applied science. *Journal of Positive Behavior Interventions*, *4*, 4–16, 20.

Carr, E. G., Horner, R. H., Turnbull, A. P., Marquis, J. G., Magito-McLaughlin, D., McAtee, M. L., et al. (1999). *Positive behavior support as an approach for dealing with problem behaviour in people with developmental disabilities: A research synthesis* (AAMR Monograph). Washington, DC: American Association on Mental Retardation.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., Crits-Christoph, P., et al. (1996). An update on empirically validated therapies. *The Clinical Psychologist*, *49*(2), 5–18.

Crosbie, J. (1995). Interrupted time-series analysis with short series; why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Lawrence Erlbaum Associates.

Didden, R., Duker, P. C., & Korzilius, H. (1997). Meta-analytic study on treatment effectiveness for problem behaviors with individuals who have mental retardation. *American Journal on Mental Retardation*, *101*, 387–399.

Didden, R., Korzilius, H., van Oorsouw, W., & Sturmey, P. (2006). Behavioral treatment of challenging behaviors in individuals with mild mental retardation: Meta-analysis of single-subject research. *American Journal on Mental Retardation*, *111*, 290–298.

Emerson, E. (2003). Prevalence of psychiatric disorders in children and adolescents with and without intellectual disability. *Journal of Intellectual Disability Research*, *47*, 51–58.

Evans, I. M., & Meyer, L. H. (1985). *An educative approach to behavior problems: A practical decision model for interventions with severely handicapped learners*. Baltimore, MD: Paul H. Brookes.

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single case designs. In R. D. Franklin, D. B. Allison & B.S. Gorman (Eds.), *Design and analysis of single case research* (pp. 159–214). Mahwah, NJ: Lawrence Erlbaum Associates.

Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, *13*, 543–559.

Helmstetter, E., & Durand, V. M. (1991). Nonaversive interventions for severe behavior problems. In L. H. Meyer, C. A. Peck, & L. Brown (Eds.), *Critical issues in the lives of people with severe disabilities* (pp. 559–600). Baltimore, MD: Paul H. Brookes.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*, 107–118.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics*, *3*, 27–46.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analysis, and single subject designs. *Psychology in the Schools*, *44*, 483–493.

Karasu, N. (2006, April). *Evidence-based practices for young children with developmental disabilities: A meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53–70.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lowe, K., Allen, D., Jones, E., Brophy, S., Moore, K., & James, W. (2007). Challenging behaviours: Prevalence and topographies. *Journal of Intellectual Disability Research*, *51*, 625–636.

Marquis, J. G., Horner, R. H., Carr, E. G., Turnbull, A. P., Thompson, M., Behrens, G. A., et al. (2000). A meta-analysis of positive behavior support. In R. Gersten, E. P. Schiller, & S. Vaughn (Eds.), *Contemporary special education research: Synthesis of knowledge base on critical instructional issues* (pp. 137–178). Mahwah, NJ: Lawrence Erlbaum Associates.

Mathur, S. R., Kavale, K. A., Quinn, M. M., Forness, S. R., & Rutherford, R. B. (1998). Social skills interventions with students with emotional and behavioral problems: A quantitative synthesis of single-subject research. *Behavioral Disorders*, *23*, 193–201.

McClean, B., & Grey, I. (2007). Modifying challenging behaviour and planning positive supports. In A. Carr, G. O'Reilly, P. N. Walsh, & J. McEvoy (Eds.), *The handbook of intellectual disability and clinical psychology practice* (pp. 643–684). London: Routledge.

Meyer, L. H., & Evans, I. M. (1989). *Non-aversive intervention for behavior problems: A manual for home and community*. Baltimore: Paul H. Brookes.

Meyer, L. H., & Evans, I. M. (1993). Science and practice in behavioral intervention: Meaningful outcomes, research validity, and usable knowledge. *Journal of the Association for Persons with Severe Handicaps*, *18*, 224–234.

Meyer, L. H., & Evans, I. M. (2006). *Literature review on interventions with challenging behaviour in children and youth with developmental disabilities*. Wellington, NZ: Ministry of Education. Retrieved 12 February 2008 from http://www.educationcounts.govt.nz/publications/special_education/15183

Mostert, M. P. (2001). Characteristics of meta-analyses reported in mental retardation, learning disabilities, and emotional and behavioral disorders. *Exceptionality*, *9*, 199–225.

Murphy, G. H., Beadle-Brown, J., Wing, L., Gould, J., Shah, A., & Holmes, N. (2005). Chronicity of challenging behaviours in people with severe intellectual disabilities and/or autism: A total population sample. *Journal of Autism and Developmental Disorders*, *35*, 267–280.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology*, *25*, 313–324.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy*, *34*, 189–211.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., et al. (2005). Effect sizes in single case research: How large is large? *School Psychology Review*, *34*, 116–132.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183–192.

Sanders, M. R., Mazzucchelli, T. G., & Studman, L. J. (2004). Stepping Stones Triple P: The theoretical basis and development of an evidence-based positive parenting program for families with a child who has a disability. *Journal of Intellectual & Developmental Disability*, *29*, 265–283.

Schindler, H. R., & Horner, R. H. (2005). Generalized reduction of problem behavior of young children with autism: Building trans-situational interventions. *American Journal on Mental Retardation*, *110*, 36–47.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation*, *96*, 233–256.

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy*, *32*, 879–883.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*, 24–33.

Swanson, H. L., & Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, *33*, 114–136.

Task Force on Psychological Intervention Guidelines. (2002). *Template for developing guidelines: Interventions for mental disorders and psychosocial aspects of physical disorders*. Washington, DC: American Psychological Association.

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*, 1–10.

Voeltz, L. M., & Evans, I. M. (1982). The assessment of behavioral interrelationships in child behavior therapy. *Behavioral Assessment*, *4*, 131–165.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment*, *11*, 281–296.