Improving procedural fidelity of behavioural interventions for people with intellectual and developmental disabilities: A systematic review

Lucy Brady[1]

Ciara Padden[1]

Peter McGill[1]

[1]Tizard Centre, University of Kent, United Kingdom

Corresponding Author: Lucy Brady, Tizard Centre, University of Kent, United Kingdom. Email: lmb47@kent.ac.uk

Note: Accepted for publication in the *Journal of Applied Research in Intellectual Disabilities* on 25.02.19.

Abstract

Background. Despite its importance within behavioural intervention, it remains unclear how best to achieve high procedural fidelity. This paper reviewed studies on improving procedural fidelity of behavioural interventions for individuals with intellectual and developmental disabilities (IDD).

Method. A systematic literature search was conducted, which identified 20 studies meeting inclusion criteria. Data were extracted on study design, participant characteristics, intervention, target behaviours, effect sizes, maintenance, generalisation, and social validity. A quality rating was also applied.

Results. A total of 100 participants took part in the included studies. Most participants were teachers working with children in school settings. There was a significant positive correlation between level of procedural fidelity and client outcomes. Feedback was the most commonly employed intervention to improve procedural fidelity.

Conclusions. More research should be conducted in environments with high levels of variability such as community homes to determine how to reach and maintain high levels of procedural fidelity.

Keywords: procedural fidelity; intellectual disability; behavioural intervention; treatment integrity; human services; staff

Introduction

Intellectual and developmental disabilities (IDD) are lifelong conditions with individuals requiring some level of ongoing support throughout their lives. Such supports may include educational, medical, social and residential components. With a prevalence of approximately 1% of the world's population (Maulik, Mascarenhas, Mathers, Dua, & Saxena, 2011), providing support to people with IDD is a significant financial consideration. It has been estimated that the cost is $1m per person over their lifetime in the United States (US; Moeschler, 2013). In the US, approximately 756,000 individuals with IDD need residential support (Larson et al., 2015), with 43% of these people requiring behaviour support (National Core Indicator Data, 2015). Braddock, Hemp, Rizzolo, Tanis, Haffer and Wu (2015) suggest that this amounts to a staggering $61 billion dollar cost annually for the US government in long term supports and residential services for adults alone. This situation is similar in other countries. In the United Kingdom (UK), approximately 189,000 adults with IDD using community residential services at a cost of approximately £5.3 billion per year (Emerson, Vick, Rechel, Muñoz-Baell, Sørensen, & Färm, 2012). Currently 2,600 adults with IDD and challenging behaviour are residing full time in English mental health facilities at a cost of £557 million per year to the National Health Service (National Audit Office; NAO, 2015). These costs to governments across the globe highlight how vital it is to address the effectiveness of the interventions and services these clients receive. Not only is this important to ensure that people with IDD are able to progress, moving away from living with behavioural issues and towards an

improved quality of life, but also to ensure governments are able to provide sustainable services for future generations.

In the UK, scandals such as Winterbourne View (Kenyon & Chapman, 2011), where individuals with IDD living in residential services were subjected to ongoing abuse and ineffective services, have created a sense of urgency to improve the quality of lives of people with IDD. This has led to increasing demand for positive behaviour support (PBS), an evidence-based approach to supporting individuals who display challenging behaviour. PBS is a person-centred approach which seeks to improve the quality of life of individuals with challenging behaviour by drawing on behavioural interventions, valued outcomes and promoting choice, inclusion and equality in the community (Gore et. al, 2013). A review conducted by La Vigna and Willis (2012) found PBS to be an effective approach for individuals with severe challenging behaviour and also challenging behaviour that occurred at high rates. They also found that it was a cost effect method which could be applied in community and institutional settings. PBS emphasises the use of proactive interventions and strategies to support individuals who display challenging behaviour with the goal of improving quality of life. However, when considering the effectiveness of such supports, it is important to consider whether they are being implemented as intended. Otherwise, the outcomes of reducing challenging behaviour and achieving improved quality of life may be less likely to be achieved.

Procedural fidelity refers to the extent to which an intervention is carried out as intended; in the case of PBS or other behavioural interventions, this is likely to refer to the extent to which a behavioural intervention is carried out according to a behavioural intervention plan (Gresham, Gansel, Nowell, Cohen & Rosenblum,

1993). There is growing evidence suggesting that stronger procedural fidelity is associated with more successful interventions and better outcomes for the individual (DiGennaro, Martens & Kleinmann, 2007; Vollmer, Roane, Ringdahl & Marcus, 1999). Similarly, Wilder, Atwell and Wine (2006) concluded that new skills were mastered more quickly when the intervention was carried out with high fidelity, leading to lower levels of challenging behaviour. Thus, it is important to have systems in place to measure procedural fidelity in human services. Research has been conducted on effective ways to record and monitor procedural fidelity levels across schools and human services (Horner, Todd, Lewis-Palmer, Irvin, Sugai & Boland., 2004; LaVigna, Willis, Shaull, Abedi & Sweitzet., 1994). However, it is often not measured or reported. For instance, reviews of the school-wide positive behaviour support literature report that procedural fidelity was recorded in less than half of studies (Bruhn, Hirsch & Lloyd, 2015) and that the actual level of procedural fidelity reached by staff dropped significantly in environments with higher variability such as high schools (Horner, Sugai & Anderson., 2010). This is an area that is also often overlooked in practice, meaning the huge financial costs of providing support to these individuals may build.

Several factors may interfere with implementation of an intervention, contributing to low procedural fidelity and poor outcomes for the client (DiGennaro et al., 2007). These factors may include inadequate staff training, incomplete training on the delivery of specific interventions or complex protocols (Vollmer, Sloman & St-Peter-Pipkin, 2008). This raises an ethical issue for any clinician involved, as they may inadvertently be allowing their clients to receive ineffective treatment. It may also lead to suggestions of negligence in cases of challenging behaviour if the

intervention prescribed was not delivered properly. Life changing decisions may be made based on the outcome of interventions that are ineffective, which may include residential placements, use of restrictive procedures or introduction or withdrawal of medications. Vollmer et al. (2008) highlight that few would make these decisions without being certain of procedural fidelity if the problem was medical rather than behavioural. Thus, when low procedural fidelity has been identified, this raises the question of which strategies are most effective in improving procedural fidelity, an outcome that could have far-reaching benefits.

To date, there is no comprehensive review of approaches to improving procedural fidelity of behavioural interventions. This review aims to summarise the findings of previous studies regarding improvement of procedural fidelity in human services for individuals with IDD. It will examine the participants used, the settings of the interventions and the different interventions used. The review will also examine the included studies for quality indicators including social validity, maintenance, and generalisation.

Method

Search Methods

Comprehensive database searches were carried out by entering keyword combinations (see Table 1) into the PsycInfo, SCOPUS, Web of Science and ERIC databases. The reference lists of articles that met inclusion criteria were also hand searched for possible citations of papers not found electronically. Publication year was not restricted, but only peer-reviewed papers published in the English language were considered for inclusion.

Inclusion Criteria. Studies were included if they had:

(1) Included a baseline and post-intervention measure of procedural fidelity of behaviour interventions delivered by staff, family or other carers (whether this was the primary focus of the study or was included as a secondary outcome), and

(2) Included family or staff who worked in any setting with individuals with an intellectual or developmental disability (e.g., homes, hospitals, group homes, day services, schools, outpatient clinics, etc.). There was no restriction on the type of staff or setting. One study (Vince Garland, Holden & Garland, 2016), used a simulation avatar and this study was retained in the review as the simulation was designed to replicate real-life interactions with individuals with intellectual disabilities.

Selection of studies

Initial screening of titles and abstracts by the researcher eliminated all those citations irrelevant to the topic and studies in a language other than English. Thereafter, full-texts of the remaining papers were screened against the inclusion and exclusion criteria to identify the final sample of papers included (See Figure 1). A second reviewer assessed the final studies included to ensure they met criteria and there was 100% agreement between the researchers on included studies.

Data extraction

The researcher extracted data from the identified studies and recorded it in a specially designed data extraction form. The following data were extracted.

Study design. The type of design used in each study was recorded here.

Participant characteristics and setting. Details of number of participants, participant's occupation, gender and the setting were recorded for each study where available.

Primary participant intervention. The main focus of the review was any intervention aimed at improving procedural fidelity of the staff or family member. Interventions were pre-defined by the reviewer into 11 categories and the intervention that was used within each study was determined by reviewing the elements that were reported within the study. This ensured consistency, since studies often used different terms to refer to the same interventions. See Table 2 for definitions of each intervention.

Primary participant target behaviour. The target behaviour for the primary participant (e.g., staff or family member) was identified, such as implementing discrete trial teaching.

Secondary participant characteristics. Three age categories were used to classify the samples: (a) Child (1-11years); (b) Adolescent (12-17 years); and (c) Adult (18+ years). Details of the client's diagnosis were also gathered where possible.

Secondary participant target behaviour. Where available, details of the behaviour being targeted for the secondary participant were also included for each study.

Effect size. Effect size was calculated for both the primary and secondary participants (where possible) using non-overlap of all pairs (NAP; Parker & Vannest, 2009). NAP was used to measure effect sizes of the procedural fidelity intervention and the client's intervention where applicable. NAP is used to display the percentage

of data which improve across phases (Parker & Vannest, 2009). NAP was chosen as it is appropriate for single case designs. Additionally, it is not affected by ceiling effects and is appropriate to use where there are a small number of data points, which was important in the present review because several of the studies had small data sets and included interventions that produced ceiling effects. It is also relatively simple to calculate NAP by hand and it has strong statistical power relative to other approaches (Parker & Vannest, 2009).

To calculate NAP, pairs were identified by comparing each phase A data point with each phase B data point. NAP was calculated as the number of improving or positive pairs (POS) plus half of the tied pairs (TIES), divided by the total number of pairs (PAIRS): NAP = [POS + 0.5TIES]/PAIRS(Parker & Vannest, 2009). NAP effect sizes were coded according to Parker and Vannest's (2009) guidelines using the following ranges: weak effects: 0 - .65; medium effects: .66 -.92; strong effects: .93 - 1.0.

Quality assessment. A quality assessment was conducted on each study using an adapted version of the evaluation of research report strength from Reichow's evaluative method (Reichow, Volkmar & Cicchetti, 2008). This approach can be used to assess the quality of both single subject and group design studies. Ratings were awarded to each quality indicator as detailed in Reichow et al. (2008). However, given that Reichow et al.'s (2008) approach was designed for use with studies relating to children with autism, the primary quality indicator for participants was adapted to make it appropriate for the present review (since this included studies focusing on staff and parents rather than children with autism). The criteria for this quality

indicator originally required inclusion of age, gender and diagnosis for all participants in a study. This was adapted in the present review such that it was deemed acceptable for participant ages not to be reported, since this information (i.e., that they were adults) could be deduced from the information provided (e.g., occupation). Furthermore, there was no diagnosis to be reported, so this was removed from the requirements. As a result, the ratings for the primary quality indicator of participant characteristics could be coded either as high (if items i-iii were included) or unacceptable (if any of items i-iii were not included), with the acceptable category removed. Furthermore, Reichow et al's (2008) original strength ratings had three categories, which was expanded to five ratings in the present review to produce a more sensitive quality assessment tool with a wider range of ratings. (Tomlinson, Gore & McGill, in press). The adapted requirements for each strength rating are shown in Table 3.

Maintenance, generalisation and social validity. Maintenance was considered to have been included if all aspects of the intervention were removed and the dependent variable assessed. Maintenance was not considered to be included when there was a maintenance/follow up period where the intervention or parts of the intervention were still in place. A study was considered to have assessed for generalisation if the skills were later assessed in untrained settings, with untrained people or with untrained materials. Data from studies that assessed for social validity was extracted regarding the method of assessment, if the method was standardised or non-standardised and the social validity score recorded.

Interobserver Agreement (IOA)

A second researcher conducted IOA on 100% of the included studies. The second researcher independently conducted data extraction on all studies. Agreements were defined as both observers identifying the same characteristics for extraction or arriving at the same NAP figure or quality rating. Disagreements were defined as observers recording different characteristics for extraction or producing a different NAP figure or quality rating. Mean IOA was calculated using the following formula: [Number of agreements/Number of agreements plus number of disagreements] * 100. Mean inter-observer agreement was found to be 84.3% (range 63.6% - 100%). Any disagreements were discussed between the researchers and resolved.

Results

A total of 20 papers published between 2004 and 2016 met the inclusion criteria. Table 4 summarises experimental design, participant characteristics and setting, intervention, participant target behaviour, participant effect size, client characteristics, target behaviour and effect size (where applicable), quality assessment, and information on generalisation, maintenance and social validity.

Study Design

The majority of studies (n = 19) used a single case research design, with only one study (Minjarez, Williams, Mercier, & Hardan, 2011) reporting use of a group (pretest-posttest) experimental design. Of the studies that had used a single subject design, multiple baseline across participants design was most common, with 95% (n

= 18) of the studies reporting use of this design and only 5% (n = 1) reporting use of a within subjects changing criterion design (Weinkauf, Zeug, Anderson & Ala'i-Rosales, 2011). Given that only one study (Minjarez et al. 2011) had reported use of a group experimental design, the results of this study have been presented separately first, followed by the findings for the studies that used a single case research design.

The one study (Minjarez et al., 2011) that reported using a group experimental design used a pretest-posttest design conducted over an 18-month period. Seventeen parents of children with ASD (all male) took part in the study which consisted of behavioural skills training, assessment, goal setting and observations being conducted on a weekly basis either in person or via video in a clinical setting. The parents were trained to implement Pivotal Response Training with their children with high fidelity targeting specific language goals identified by the parents and the researchers in a clinical setting. Results found significant increases both in the parents' treatment fidelity and the level of children's utterances. The quality assessment of the study carried out according to the Reichow et al. (2008) method found the study to be borderline adequate. The study did not assess for generalisation, maintenance or social validity.

Studies that used a single case research design have been presented in Table 4. The rest of the Results section refers to these papers.

Participants

Primary participant characteristics. Primary participants refer to the staff who were the focus of the procedural fidelity intervention. Across the 19 studies, a total of 83 participants took part with between three to nine participants in each study. Sixty

three participants were female and 13 were male. Two studies (Courtemanche, Sheldon, Sherman, Schroeder, Bell & House., 2014 and McKenny & Bristol, 2015) did not report details of gender. As shown in Table 5, the majority (41%) of primary participants were teachers. Front-line support staff made up only 11% of primary participants.

Secondary participant characteristics. Secondary participants referred to the individuals with IDD who were in receipt of intervention from the primary participants. The ages of the secondary participants were divided into three categories: children (age 0-11), adolescents (age 12-17) and adults (aged 18 years or older). As seen in Table 5, children made up the greatest number of secondary participants (76%) with adults (18+) as the least frequently included in studies (4%). With respect to diagnoses, two studies (DiGennaro-Reed et al., 2010; Miller et al., 2014) did not provide any information on the diagnoses of secondary participants. For the remaining studies, the majority of participants were reported to have a single diagnosis (71%) rather than multiple diagnoses (29%). The most common diagnosis reported was ASD (75%). . Finally, more than half of the reported studies (88%) were carried out in a classroom setting and only 12% took place in more uncontrolled settings such as the community or residential homes.

Target behaviour of primary participants. The target behaviour in all studies was increasing implementation accuracy of interventions. There was an even balance between behaviour reduction and skills teaching, with 47% of the interventions being implemented to reduce client problem behaviours using function-based behaviour support plans (Codding et al., 2005, Codding et al., 2008; Courtemanche et al., 2014;

DiGennaro et al., 2005; DiGennaro et al., 2007; DiGennaro-Reed et al., 2010; Flynn & Lo, 2015; Pelletier et al., 2010) and the Good Behaviour Game (Maggin et al., 2012), while 53% of the interventions were implemented to increase positive behaviours or teach skills with discrete trial training (Belfiore et al., 2008; McKenny & Bristol, 2015; Miller et al., 2014; Pollard et al., 2014), pivotal response training (Coolican et al., 2010), system of least prompts (Vince Garland et al., 2016), a token economy (Plavnick et al., 2010) and Project ImPACT (Ingersoll & Wainer, 2013) which is a social communication intervention for children which uses modelling to increase spontaneous language.

Intervention. Interventions were coded into 11 types of intervention. Most studies used a combination of interventions and the results reported highlight the inclusion of a particular intervention in the study. As shown in Table 6, the most commonly reported intervention was feedback (68%), followed by observation (37%), role-play (37%), modelling (37%), behavioural skills training (32%), self-monitoring (26%), teaching (26%), quizzes (16%) and negative reinforcement, financial incentives and goal setting the least frequently reported in only 11% of studies.

Effect sizes (NAP)

Primary participants. NAP effect sizes were calculated to determine the effect the intervention had on procedural fidelity for the primary participant. A total of 32 effect sizes were calculated across the 19 studies as some studies had several phases to their interventions. These effect sizes were then coded into weak, medium and strong effects according to Parker and Vannest's (2009) guidelines. Table 7 shows that the majority of primary participants' intervention phases were weak and only 3%

of phases were found to have a strong effect size. Most secondary participants intervention effect sizes were found to be medium with only 11% of phases were found to be strong.

Correlation between effect sizes. Pearson's *r* correlation was conducted to determine if there was a relationship between the primary participant effect size (i.e., for procedural fidelity) and the client effect size (i.e., for the client's behaviour that was being targeted by the intervention). There was a statistically significant positive correlation between the primary and secondary participants' effect sizes, $r(14) = 0.3365$, $p = 0.002$, such that improvements in procedural fidelity were associated with improvements in client behaviour. However, this needs to be interpreted cautiously given that it represents a relatively weak effect size and given the lack of variation in the participant effect sizes.

Effect sizes of interventions. Effect sizes were further examined across different interventions. This was complicated by the fact that interventions were often used in combination with other interventions rather than being used in isolation. Across the 19 studies that used a single-case design, a single intervention was used in 11 out of a possible 32 intervention phases. The combination of interventions used and their effect sizes are displayed in Table 6.

Quality Assessment. The Reichow et al. (2008) quality assessment was carried out on all 19 studies. Two (11%) of the studies were found to be strong (Flynn & Lo, 2015; Vince Garland et al., 2016), five (26%) were adequate (DiGennaro et al., 2005; DiGennaro-Reed et al., 2010; Miller et al., 2014; Pelletier et al., 2010; Pollard et al., 2014), seven (37%) were found to be borderline adequate (Belfiore et al., 2008;

Coolican et al., 2010; Courtemanche et al., 2014; Ingersoll & Wainer, 2013; Maggin et al., 2014; Mouzakitis et al., 2015; Weinkauf et al., 2011), while five (26%) were found to be weak (Codding et al., 2005; Codding et al., 2008; DiGennaro et al., 2007; McKenny & Bristol, 2015; Plavnick et al., 2010).

Correlation between NAP and quality rating. Pearson's *r* correlation was conducted to determine if there was a relationship between the primary participant effect size and the quality of the study. There was a statistically significant but weak positive correlation between the primary participant effect size and the quality of the study: $r(14) = 0.226$, $p = .002$. such that larger improvements in procedural fidelity were associated with higher quality of the study and vice versa.

Maintenance, generalisation and social validity

Maintenance. Only seven (37%) studies assessed for maintenance (Belfiore et al., 2008; Codding et al., 2005; Coolican et al., 2010; DiGennaro-Reed et al., 2010; Ingersoll & Wainer, 2013; Maggin et al., 2012; Vince Garland et al., 2016). The maintenance probe times ranged from one week follow up to four months. Of the seven studies assessed, all reported evidence of maintenance in all primary participants.

Generalisation. The assessment of generalisation was described in two (11%) studies (Flynn & Lo. 2015; Mouzakitis et al., 2012). In both studies, generalisation was assessed across other people and generalisation was achieved.

Social validity. Thirteen (68%) of the studies assessed for social validity (Codding et al., 2005; Codding et al., 2008; Coolican et al., 2010; Courtemanche et al., 2014; DiGennaro et al., 2005; DiGennaro et al., 2007; DiGennaro-Reed et al.,

2010; Flynn & Lo 2015; Maggin et al., 2012; McKenny & Bristol, 2015; Miller et al., 2014; Pollard et al., 2014; Vince Garland et al., 2016). Only four (30%) of these studies used standardised questionnaires, (DiGennaro et al., 2005; DiGennaro et al., 2007; DiGennaro-Reed et al., 2010; Maggin et al., 2012) while the other studies used surveys or questionnaires developed by the authors. All studies included an element of a questionnaire or study and all reported positive outcomes for participants.

Discussion

The systematic review examined interventions used to improve the level of procedural fidelity with which professionals and carers implement behaviour support plans for individuals with IDD. Twenty studies published between 2004 and 2016 were included in the review. Findings suggest that there are several ways to improve procedural fidelity, with all studies showing some increase in procedural fidelity following intervention. Notably, however, the level of improvement was not consistent across studies, suggesting that some interventions may be more effective than others. Maintenance data was only measured in 37% of the studies (Belfiore et al., 2008; Codding et al., 2005; Coolican et al., 2010; DiGennaro-Reed et al., 2010; Ingersoll & Wainer, 2013; Maggin et al., 2012; Vince Garland et al., 2016) so it is not possible to determine if the interventions were generally effective for long term implementation.

Children made up the largest age group for secondary participants, with only two adults included across all studies. Surprisingly, none of the studies in the review were conducted in a residential setting for adults with IDD. The main setting for the

studies was in schools and educational settings, with a focus on academic skills and behaviours. Furthermore, the primary settings included (i.e., clinics and separate classrooms within schools) were mainly controlled environments, while more naturalistic settings (e.g., homes or adult residential services) were not utilised. This is surprising considering the huge amounts of money paid out across the world for adults with IDD who live in residential settings (e.g., £5.3billion per year in the UK; NAO, 2015), highlighting the need for research within this area. Furthermore, different factors can be relevant within more naturalistic settings compared to controlled research environments. Saunders and Spradin (1991) highlighted the unpredictable nature of residential settings caused by variables such as the presence of other service users, members of the public and staff turnover. Thus, further research is needed to identify how to promote procedural fidelity in settings such as adult community based services, and to understand how these issues of unpredictability in certain settings might impact on intervention outcomes.

The client behaviours targeted within the studies were a balanced mix of reducing problem behaviours (47% studies) and increasing or teaching new skills (53% studies). While it was positive to see an emphasis on skill development as well as behaviour reduction, this may be reflective of the fact that most studies were conducted in schools with children. The most commonly applied intervention for improving procedural fidelity was feedback, used in 21.7% of studies as an individual intervention and in 13.3% of studies as part of a behavioural skills training package. However, there were inconsistencies across studies about the definition of feedback. Types of feedback included verbal, written feedback, immediate or in-vivo feedback and delayed feedback. Most feedback was delivered by a supervisor or the researcher

but peer feedback was also included in one study. The intervention with the highest effect size was teaching and instruction with 100% effect size across five studies. However, in each study, teaching and instruction was used as part of a package so it is not possible to say that the effect size of each study was only as a result of this intervention. An interesting point to note is the strength of the effect sizes for each group of participants. The effect size of the primary participants' intervention was mainly strong with 78.1% of effect sizes recorded as such. Only 2.7% of the primary participant effect sizes were recorded as weak which would suggest that the interventions implemented were effective in increasing the accuracy with which staff were implementing behavioural interventions. In contrast to this, most secondary participant effect sizes (63.2%) were recorded as being of medium strength. The lack of weak effect sizes in the interventions may suggest however, that there is an element of publication bias in determining how studies are chosen. Further research into the efficacy of different interventions is needed in order to develop an efficient method of improving and maintaining procedural fidelity. Currently, the combination of interventions used in intervention packages makes it difficult to conclude which elements of the package contributed to the outcome.

It is also clear from the current review that the interventions applied are effective and successful in improving procedural fidelity, but the combination of interventions and how they should be selected is less clear. Feedback and observation were by far the most widely used interventions, but this is possibly due to the fact they are convenient, cheap and easy to use. They both had strong effect sizes on procedural fidelity, but so did other interventions which would be more time consuming and costly and so may not have been included in as many studies. The

strongest effect size noted was teaching and instruction, which had a consistent score of 100% effect size in each programme it was included in. This is, however, a particularly time consuming intervention which requires expertise and resources. Moving forward, it would be important to explore the efficacy of different interventions to investigate how they can be used in a larger context to yield positive results for staff and clients.

The quality of studies reviewed is also notable, with 63.2% of single case design studies falling into weak or borderline adequate ratings and only 10.5% of studies being rated as strong. This highlights significant quality issues within much of the procedural fidelity literature. This is reflected in the fact that only 31.6% of studies assessed for maintenance and only 10.5% assessed for generalisation of behaviour change. As these key quality indicators have been omitted in so many cases, it is difficult to predict if the interventions and techniques applied would be successful in different settings or with a different population. However, it is important to note that there was a statistically significant positive relation between quality of the study and the primary participant effect size. This suggests that the higher quality studies, which assessed for maintenance, social validity, generalisation and displayed experimental control, were more likely to produce larger effects, meaning the primary participant was more likely to increase procedural fidelity. However, the effect size was weak and overlap between experimental control and quality of study may be confounding this finding, so caution is warranted. More methodologically rigorous research should be conducted, particularly in the areas of maintenance and generalisation to determine how these impact on procedural fidelity and outcomes for clients.

There was a positive correlation between the primary and secondary participant effect sizes suggesting that high level procedural fidelity for staff is associated with higher intervention effects for clients. This finding is consistent with previous research (DiGennaro et al., 2007; Vollmer et al., 1999; Wilder et al., 2006) that suggested that an intervention will be more successful if it is carried out with high fidelity. This finding suggests that it would be valuable to establish a technique to deliver interventions with high fidelity, on a consistent basis, across staff and carers to ensure clients receive the best possible outcomes. However, the effect size recorded was weak, perhaps due to the lack of variation in participant effect sizes and so should be interpreted with some caution. Other variables may need to be taken into account such as age of client, client diagnosis, number of sessions during which data was recorded, staff experience, level of training etc. While this is a small sample and a small effect size, it does suggest the need to work to improve procedural fidelity. Further exploration into factors that contribute to low levels of procedural fidelity would be valuable to identify areas for improvement or change within human service settings.

The current review was limited in some ways. By only including individuals with IDD, the review may have missed studies conducted with other populations that require behavioural input and may have broadened the age range slightly. These could include young offenders, dementia patients, individuals with mental health issues and typically developing children (particularly for interventions for skills acquisition). It may be possible to learn more about procedural fidelity by conducting future reviews of intervention in this area for other populations. By requiring pre and post intervention data of procedural fidelity levels, it is possible that the review may have

excluded broader literature on procedural fidelity. However, this criterion was included to be able to examine the effect size of different interventions.

In conclusion, there are several studies that highlight how procedural fidelity can be improved in various settings. Considering the huge costs of providing support and the implications for individuals with IDD of receiving ineffective, poorly implemented interventions, it will be important for future research to explore the efficacy of these individual interventions to develop quality supports across human services. Most of the research is currently with children in school settings and it will be important to expand to other populations to explore if different interventions are effective at improving procedural fidelity in services that have more variability such as residential services or in community-based settings. Studies are relatively weak from a quality perspective with many not including vital assessments for generalisation and maintenance. Therefore, further high-quality research is warranted to determine the most effective approaches for achieving and maintaining high procedural fidelity across a variety of settings to ensure that adults with IDD are receiving high quality support.

References

Belfiore, P.J., Fritts, K.M., & Herman, B.C. (2008). The role of procedural integrity. Using self-monitoring to enhance discrete trial instruction (DTI). *Focus on Autism and Other Developmental Disabilities. 23* (2), 95-102.

Braddock, D., Hemp, R., Rizzolo, M. C., Tanis, E. S., Haffer, L., & Wu, J. (2015). *The State of the States in Intellectual and Developmental Disabilities.* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Bruhn, A.L., Hirsch, S.E., & Lloyd, J.W. (2015). SWPBS in school-wide programs: a review of the literature (1993 -2012). *Journal of Primary Prevention, 36* (5), 335-349.

Codding, R.S., Feinberg, A.B., Dunn, E.K., & Pace, G.M. (2005). Effects of immediate performance feedback on implementation of behaviour support plans. *Journal of Applied Behaviour Analysis*, 38, 205-219.

Codding, R.S., Livanis, A., Pace, G. M., & Vaca, I. (2008). Using performance feedback to improve treatment integrity of classwide behaviour plans: An investigation of observer reactivity. *Journal of Applied Behavior Analysis*, *41*, 417-422.

Coolican, J., Smith, I.M., & Bryson, S.E. (2010). Brief parent training in pivotal response treatment for pre-schoolers with autism. *Journal of Child Psychology and Psychiatry, 51:12,* 1321-1330.

Courtemanche, A., Sheldon, J., Sherman, J., Schroeder, S., Bell, A., & House, R. (2014). Assessing the effects of a staff training package on the treatment integrity of an intervention for self-injurious behaviour. *Journal of Developmental and physical disabilities.* 26: 371-389

DiGennaro, F.D., Martens, B.K., & McIntyre, L.L. (2005). Increasing treatment integrity through negative reinforcement: Effects on teacher and student behaviour. *School Psychology Review. 34*, 220-231.

DiGennaro, F. D, Martens, B. K. & Kleinmann, A. E. (2007). A comparison of performance feedback procedures on teachers' treatment implementation integrity and students' inappropriate behaviour in special education classrooms. *Journal of Applied Behaviour Analysis, 40,* 447-461.

DiGennaro-Reed, F.D., Codding, R., Catania, C.N., & Maguire, H. (2010). Effect of video modelling on treatment of integrity of behavioural interventions. *Journal of Applied Behavior Analysis.* 43, 291-295.

Emerson, E, Vick, B, Rechel, B, Muñoz-Baell, I, Sørensen, J & Färm, I 2012, Health inequalities and people with disabilities in Europe, social exclusion, vulnerability and disadvantage task group background paper 5, European Regional Office of the World Health Organization, Copenhagen.

Flynn, S.D. & Lo, Y. (2016). Teacher implementation of trial based functional analysis and differential reinforcement of alternative behaviour for students with challenging behaviour. *Journal of Behavioral Education,* 25, 1-31.

Gresham F. M, Gansle K. A,Nowell G. H, Cohen S & Rosenblum S.(1993) Treatment integrity of school based behavioural intervention studies: *School Psychology Review. 22*, 254-272

Gore, N.J., McGill, P., Toogood, S., Allen, D., Carl Hughes, J., Baker, P., Hastings, R.P., Noone, S.J., & Denne, L.D. (2013). Definition and scope for positive behavioural support: *International Journal of Positive Behavioural Support,* 3, 2.

Hagermoser-Sanetti, L. M., Gritter, K. L., & Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review, 40*(1), 72–84

Horner, R.H., Todd,A.W., Lewis-Palmer, T., Irvin, L.K., Sugai,G., & Boland, J.B. (2004). The school-wide evaluation tool (SET): A research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions,*7(4), 224-236.

Horner, R.H., Sugai, G., & Anderson, C.M. (2010). Examining the evidence base for School-Wide Positive Behavior Support. *Focus on Exceptional Children*, 42(8)

Ingersoll, B., & Wainer, A. (2013). Initial efficacy of project ImPACT: A parent mediated social communication intervention for young children with ASD. J*ournal of Autism and Developmental Disorders. 43*, 2943-2952.

Kenyon, P. (Writer) & Chapman, M. (Director). (2011). Undercover care: The abuse exposed [Television episode] in Chapman, M (Producer) *Panorama*. British Broadcasting Corporation.

Larson, S.A., Faykin, F., Muchow-Hallis, L., Pettingel, S., Taylor, B., Hewitt, A., Sowers, M., Fay, M.L. (2015). In-home and residential long-term supports and services for people with intellectual or developmental disabilities: Status and trends through 2013. Minneapolis: University of Minnesota, Institute on Community Integration, Residential Information Systems Project.

La Vigna, G.W., Willis T.J., Shaull J., Abedi M., & Sweitzet M. (1994). *The Periodic Service Review: A total quality assurance system for human services and education*. Baltimore: Paul H. Brookes.

LaVigna, G W.; & Willis, T, J. (2012). The efficacy of Positive Behavioural Support with the most challenging behaviour: The evidence and its implications *Journal of Intellectual & Developmental Disability, (7,3)* p185-195

Maggin, D.M., Fallon, L.M., Hagermoser Sanetti, L.M. & Ruberto, L.M. (2012). Training paraeducators to implement a group contingency protocol: Direct and collateral effects. *Journal of behavioural disorders, 38* (1), 28-37.

Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., & Saxena, S. (2011). Prevalence of Intellectual disability: a met-analysis of population based studies. *Research in Developmental Disabilities,* 32(2), 419-436.

McKenny, E.L.W., & Bristol, R.M. (2015). Supporting intensive interventions for students with autism spectrum disorder: Performance feedback and discrete trial training. *School Psychology Quarterly. 30* (1). 8-22.

Miller, M.V., Carlson, J. & Sigurdsson, S.O. (2014). Improving treatment integrity in a human service setting using lottery based incentives. *Journal of organisational behaviour management. 34*(1), 29-38.

Minjarez, B.M., Williams, S.E., Mercier, E.M., & Hardan, A,Y. (2011). Pivotal response group treatment program for parents of children with austism. *Journal of Autism and developmental disorders. 41*: 92-101.

Moeschler, J.B. (2013). Neurodevelopmental Disabilities: Global developmental delay, Intellectual disability and autism. In D.L. Rimoin, R.E. Pyrotiz & B. Korf (Eds) *Emery and Rimoin's Principles and practice of medical Genetics (Sixth Edition)* pp 1-15: Elsevier Science Ltd.

Mouzakitis, A., Codding, R.S., & Tryon, G. (2015). The effects of self-monitoring and performance feedback on the treatment integrity of behaviour intervention plan implementation and generalization. *Journal of positive behaviour interventions. 17*(4) 223-234.

National Audit Office (2015) Care services for People with Learning Disabilities and Challenging Behaviour.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367.

Pelletier, K., McNamara., Braga-Kenyon., & Ahern, W.H. (2010). Effect of video self-monitoring on procedural integrity. *Behavioural Interventions, 25*, 261-274.

Plavnick, J.B., Ferreri, S.J., & Maupin, A.N. (2010). The effects of self-monitoring on the procedural integrity of behavioural intervention for young children with developmental disabilities. *Journal of Applied Behavior Analysis*, 43, 315-320.

Pollard, J.S., Higbee, T.S., Akers, J.S., Brodhead, M.T. (2014). An evaluation of interactive computer training to teach instructors to implement discrete trials with children with Autism. *Journal of applied behaviour analysis. 47,* 765-776.

Reichow, B., Volkmar, F.R., & Cicchetti, D.V. (2008). Development of the evaluative method for determining the strength of research evidence in autism. *Journal of Autism and developmental disorders, 38,* 1311-1319.

Saunders, R.R., & Spradin, J.E. (1991). A supported routines approach to active treatment for enhancing independence, competence and self-worth. *Behavioral Residential Treatment,* 6(1), 11-37.

The British Psychological Society. (2010). Code of ethics.

Tomlinson, S., Gore, N. & McGill, P. (in press) Training individuals to implement applied behaviour analytic procedures via telehealth: a systematic review of the literature. *Journal of Behavioral Education.*

Vince Garland, K.M., Holden, K., Garland, D. P. (2016). Individualized Coaching in the TLE TeachLivE Lab: Enhancing fidelity of implementation of system of least prompts among novice teachers of students with autism. *Teacher Education and special education. 39* (1), 47-59.

Vollmer, T.R., Roand, H.S., Ringdahl, J. E., & Marcus, B.A. (1999). Evaluating treatment challenges with differential reinforcement of alternative behaviour. *Journal of Applied Behavior Analysis*, 32, 9-23.

Vollmer T.R., Sloman, K.N., & St-Peter Pipkin C (2008) Practical implications of Data Reliability and Treatment Integrity Monitoring. *Behavior Analysis in Practice* 1, 4-11.

Weinkauf, S.M., Zeug, N.M., Anderson, C.T., & Ala'i-Rosales, S. (2011). Evaluating the effectiveness of a comprehensive staff training package for behavioral interventions for children with autism. *Research in Autism Spectrum Disorders. 5*, 864-871.

Wilder, D.A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behaviour Analysis. 39* (3), 369-373.

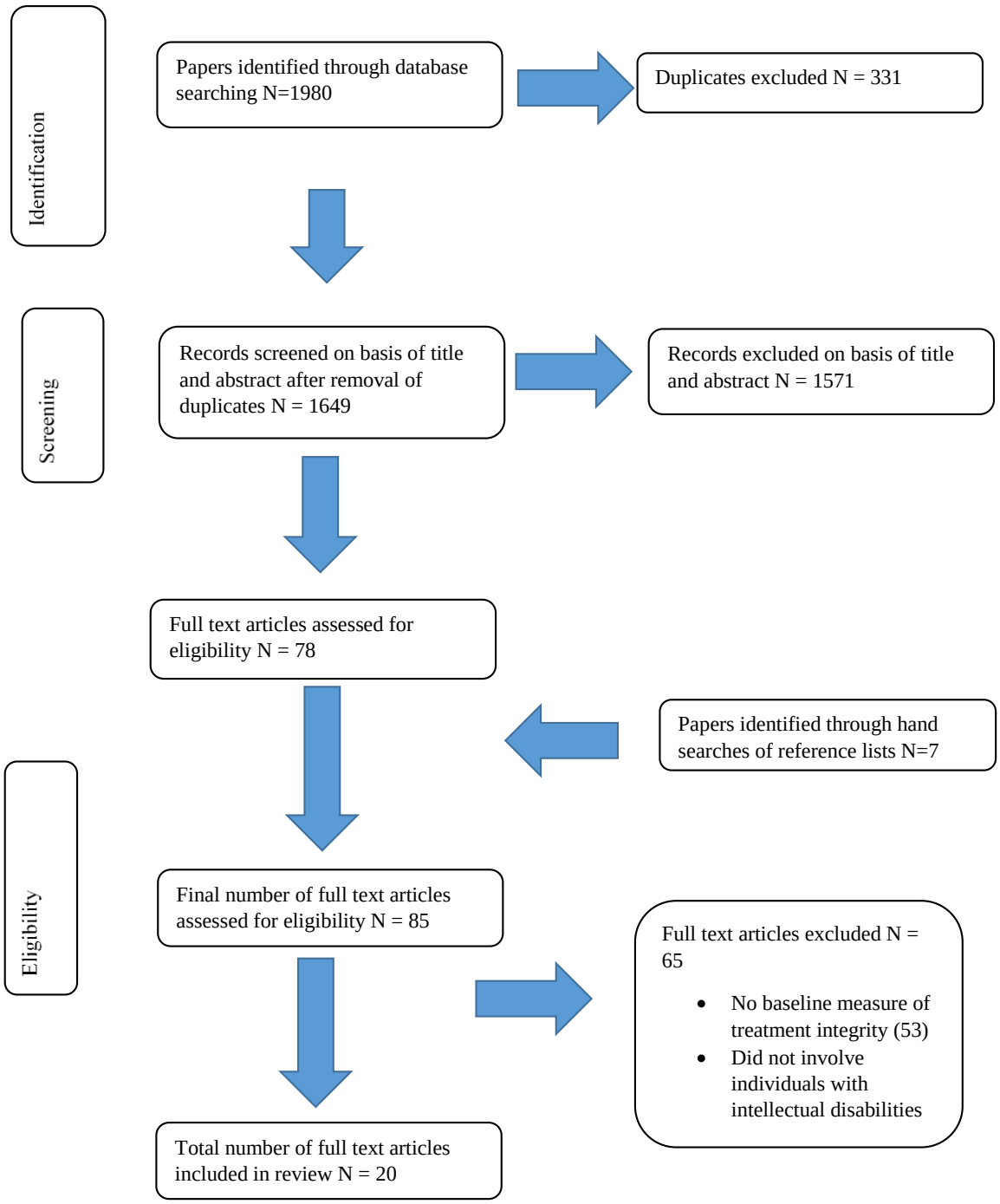

Figure 1 Flow diagram showing inclusion/exclusion of studies identified during database search process

Table 1 Search terms

Participants		Intervention		Outcome		Terms related to disability
Frontline staff OR Frontline employees OR Frontline workers Staff OR Residential Staff OR Employee OR Worker OR Nurse OR Manager OR Care Assistant OR Support worker OR Parent* OR Care* OR Teacher OR Tutor OR Direct care OR Direct worker Social care worker OR Direct support OR Attendant OR Social worker	AND	Organisational behav* management OR Organizational behav* management OBM OR Periodic service review OR PSR OR PSR Model OR Behav* OR Intervention System-wide positive behaviour support SWPBS	AND	Treatment integrity OR Fidelity OR Procedural	AND	Disability OR Autism* OR ASD OR Asperger* OR Down Syndrome OR Fragile X syndrome OR Disorder OR Learning disability OR Intellectual disability OR Developmental disability OR Special educational needs OR Mental retardation OR Special needs

Table 2 Intervention codes and definitions

Intervention code	Intervention definition
1. Feedback	To include praise for steps followed and corrective feedback for incorrect implementation. Feedback might be provided by a supervisor, trainer or researcher in vivo, immediately following implementation or in days following intervention. Feedback may be delivered in person, via skype, via email or memo.
2. Role play	Participants act out steps of intervention in a contrived situation, supervised by researcher, manager or peers.
3. Modelling	Trainer or researcher carries out intervention while being observed by participant. This may be done in person or recorded for "video modelling" which can be viewed at the participants' leisure. Modelling may be carried out with a client or with a substitute actor playing the role of client.
4. Self-monitoring	Researcher develops data sheet/ task analysis of steps in intervention. To include when participant scores themselves on sheet as they carry out intervention. This may be done during the intervention, after as reflective practice or by reviewing video tape of themselves carrying out the intervention and scoring the tape.
5. Quiz/Assessment	To include when the participant is provided with written quiz about intervention or theory to complete within 24 hours. Can also include when researcher or trainer assesses participant while observing them carry out intervention. Participant is required to reach certain criteria to pass quiz/assessment. If the participant does not reach criteria, they must repeat the assessment until they do
6. Teaching/Instruction	Sessions dedicated to giving participants background knowledge of theory for basis of intervention. May be provided 1:1 or in group situations. May be one off session or provided regularly over a number of weeks. May be provided in person or using computer training programmes.
7. Financial incentive	Participants receive monetary reward on achieving certain pre-agreed criteria.
8. Goal setting	To include when participants set goals for client behaviour and monitor client's progress towards that goal. Also to include when participants set goals for their own progress and targets to be achieved. Goals are set with support from supervisor or researcher.
9. Observation	Participant is watched by a supervisor, trainer or peer when implementing intervention. Participant may or may not be informed why the observer is present. Also to include when observation takes place via video camera or one way mirror.
10. Negative reinforcement	If a participant does not achieve criteria for the implementation of an intervention, they must attend a meeting with consultant/supervisor. If the participant does achieve criteria for the implementation of an intervention they do not have to attend a meeting.
11. Behavioural skills training	Training package that includes feedback, role-play, modelling and instruction.

Table 3 Guidelines for determination of research report strength ratings (adapted from Reichow et al., 2008)

Strength rating	Criteria
Weak	Received fewer than three high quality grades on primary quality indicators or showed evidence of less than two secondary quality indicators
Borderline adequate	Received high quality grades on three primary quality indicators with no unacceptable quality grades on any primary indicators and showed evidence of at least two secondary quality indicators
Adequate	Received high quality grades on four primary quality indicators with only one unacceptable quality grades on any primary indicators and showed evidence of at least two secondary quality indicators
Borderline strong	Received high quality grades on five primary quality indicators with no unacceptable quality grades on any primary indicators and showed evidence of three or more secondary quality indicators
Strong	Received high quality grades on all primary quality indicators and showed evidence of three or more secondary quality indicators

Table 4 Data Extraction table

Included studies	Design	Participant characteristics & setting	Participant intervention	Participant target behaviour	Participant Effect size NAP	Secondary Participants characteristics	Secondary Participants Target behaviour	Secondary participants effect size NAP	Quality Assessment of primary participants	Maintenance (M), generalisation (G), social validity (SV)
Belfiore et al. 2008	Multiple baseline design across participants	**N=** 3 staff (all female) **Setting**: private school for children with ASD	4	Administer discrete trial instruction (DTI)	M: 99% (Range 98-100%) Strong	N = Unknown Gender = Unknown Children ASD	Increasing nonverbal imitation and receptive body parts	Could not calculate NAP (data not provided)	Borderline adequate	M: *Follow up observations after 4 weeks *1/3 participants maintained intervention level of implementation G: None SV: None
Codding et al. 2005	Multiple baseline design across teacher – student dyads	**N** = 5 teachers (3 male 2 female) **Setting**: Private school for students with acquired brain injury & behaviour problems	A: 1 + 9 (with antecedent strategies) B: 1 + 9 (with consequence strategies)	Administer behaviour support plan (Antecedent & consequence strategies)	A: M = 78.21% (range 53.85-91.43%) Medium B: M = 92.74% (range 69.23-100%) Strong	N= 5 5 male Adolescent Acquired brain injury	Tantrums, preservative speech, inappropriate speech, inappropriate social behaviour, teasing, noncompliance, inappropriate social behaviour, major aggression, minor aggression, inappropriate touching, public exposure, instigation, destruction, peer instigation, invasion of space, property destruction, mimicking, wandering	Could not calculate NAP (data not provided)	Weak	M: *Follow up after 5, 10 & 15 weeks *5/5 ppts maintained intervention level implementation G:NONE SV: 10 item questionnaire (item value range 1-5) Average 4.5-5.0 for each item
Codding et al. 2008	Multiple baseline design across staff members with alternating treatments	**N =** 3 teachers (1 male 2 female) **Setting**: Mainstream school.	A: 1 & 9 (present) B: 1 & 9 (absent)	Administer behaviour support plan	A: M = 99% (range: 98 – 100%) Strong B: M= 99% (range: 98-100%) Strong	N = 6 5 female 1 male Adolescent 5 ADHD, 3 Bi-polar, 3 conduct disorder, 1 anxiety disorder	Prosocial behaviours and noncompliance	Could not calculate NAP (data not provided)	Weak	M: None G: None S.V: 10 item questionnaire (item value range 1-5) M=4.8

Coolican et al. 2010	Non – concurrent Multiple baseline design across participants	**N** = 8 parents of children with ASD (5 female 3 male) **Setting:** Clinical laboratory and family home	11	Pivotal Response Training	M = 80.21% (range = 55-100%) Medium	N = 8 7 Male, 1 Female Children ASD	Increase children's utterances	M: 89% (Range: 71 – 100%) Medium	Borderline adequate	M:Follow up at 2 to 4 months, gains maintained G: None S.V. Parent satisfaction questionnaire (item value range 1-10) Rated the whole training experience as very helpful (M =- 9/10)
Courteman che et al. (2014)	Multiple baseline design across participants (dyads)	**N**= 3 (1 para professional 2 direct care staff) **Setting**: 1 in SEN classroom, 1 in community home, 1 in group home van	A: 5,7,11 B: 1, 2, 7, 9, 10 (observer present) C: 1, 2, 7, 9, 10 (observer absent)	Administer behaviour support plan	A:M= 62% (Range:8-100) Weak B:M = 80% Range:79-100 Medium C:M= 100% Strong	N= 3 1 Adolescent, 2 adults 1 ASD, 1 ASD + profound ID + ADHD, 1 profound ID + anxiety disorder	Reduction of SIB	A:M= 74% Range: 71 – 76 Medium B:M=81 Range: 71-100 Medium C:M= 61.09% Range: – 47.92 – 69.39 Weak	Borderline adequate	M: None G: None SV: Participants completed a 10 item questionnaire & agreed they liked the teaching procedures.
Di Gennaro et al. 2005	Multiple baseline design across dyads	**N =** 4 Teachers (all female) **Setting:** 3 mainstream school, 1 special education classroom	A: 11 BASELINE C: 1, 2, 3, 10	Administer behaviour support plan	A:M = 100% Strong Return to baseline C: 100% Strong	N = 4 3 male, 1 female Children 3 ADHD 1 intellectual disability	Off task behaviours	A:M = 80.55% (Range: 0 – 9450 - 100%) Medium Return to baseline C: M: 81.82% Range: 71.67 – 88.24 Medium	Adequate	M: None G: None S.V. 15-item standardised questionnaire (Intervention Rating Profile-15 (item value range 1-6) M = 4.8/6
Di Gennaro et al. 2007	Multiple baseline design across dyads	**N =** 4 teachers (2 male, 2 female) **Setting**: Residential and educational facility for students with brain injuries	A: 11 B: 1 C: 1 & 8 D: 1 & 10	Administer behaviour support plan	A:M = 98.33% Range 93.33 - 100% Strong B:M = 70.83% Range50 - 100% Medium C:M= 85.71% Range 50 - 100% Medium	N = 4 3 Male, 1 Female 3 Children, 1 adult 3 Seizure disorder, 1 anxiety, 1 developmental disorders, 1 intellectual disability, 3	Off task behaviour	A:M = 80.59% Range 61.11-97.92% Medium B:M = 75.35% Range –55.56-87.5% Medium C:M = 80.88% Range 56.25-100% Medium D: M= 85.33%	Weak	M: None G: None S.V: 15-item standardised questionnaire (Intervention Rating Profile-15 (item value range 1-6) M = 5.2/6

					D: 99.72% Range: 98.86 – 100% Strong .	brain injury, 2 ADHD, 1 blindness		Range: 74.6-99.16% Medium		
Di-Gennaro Reed et al. (2010)	Multiple baseline design across participants	**N =** 3 teachers (All female) **Setting**: Residential and educational facility for students with brain injuries	A: 3 B: 1 & 3	Administer behaviour support plan	A:M = 98% Range 93 – 100% Strong B:M = 100% Strong	N= 3 Children	Problem behaviours	Could not calculate NAP (data not provided)	Adequate	M: 1 Week follow up probe 3/3 ppts maintained intervention levels of implementation G: None SV: 15-item standardised questionnaire (Intervention Rating Profile-15) Item value 1-6. A: *M* = 4.06/6. B: *M* = 4.93/6.
Flynn & Lo. 2015	Multiple probe across participants	**N =** 3 teachers (all female) **Setting**: Special education	A: 11 (with trial-based functional analysis) B: 11, 2, 3, 6 (with DRA)	Implementation of trial-based functional analysis (TBFA) and DRA procedures	**A (with TBFA):** M = 100% Strong **B (with DRA):** M = 100% Strong	N= 6 5 male, 1 female Children 5 ASD 1 EBD	Vocal outbursts, elopement, giggling, self stimulation	DRA Replacement bx M: 100% Strong Challenging bx reduction M = 100% Strong	Strong	M: None G:Included extra students for generalization TBFA: M=94.5% DRA Teacher 1: 98%, Teacher 2: 92% Teacher 3: Did not achieve criterion SV: Adapted version of Teacher Post-Intervention Acceptability and Importance of Effects Survey. 11 items ranked 1-5. All 4 & 5
Ingersoll & Wainer. 2013	Multiple baseline design across participants	**N=** 8 Parents of children with ASD (All female) **Setting:** Research laboratory and family home	11, 8	Administer behaviour intervention (Project ImPACT)	M = 95% (Range: 85 – 99.5%) Strong	N= 8 7 Male, 1 Female Children ASD	Increase spontaneous speech.	M: 73.53% (range: 59.52-92.42%) Medium	Borderline Adequate	M: 1 month follow up * G: None SV: None
Maggin et al. 2012	Multiple baseline design across participants	**N=** 3 Para educators (All female) **Setting**: SEN classroom for students with EBD	A: 2, 3, 6 B: 11	Administer good behaviour game	A:M = 100% Strong B:M = 100% Strong	N = 4 4 male children ADHD, ED, ID	Aggressive behaviours	A:M = 100% Strong B:M = 99.5% Range 99 – 100 Strong	Borderline Adequate	M: 1 day a week for 5 weeks *3/3 ppts maintained high levels of implementation G: None SV: Usage Rating profile intervention 35 items ranked

										1 -6 Highly acceptable M= 5.47
Mc Kenney & Bristol 2015	Multiple baseline design across participants	**N** = 9 3 Special education teachers 1 SLT 5 Teaching assistants (Gender not provided) **Setting**: SEN Classroom	1, 2, 3, 9	Administer Discrete Trial training	M = 95.69% (range: 85.71-100%) Strong	N= 3 Children ASD and intellectual disability	Specific targets not provided	Could not calculate NAP (data not provided)	Weak	M: -None G: None SV: 10-item acceptability survey developed by the authors. Values from 1-7. Overall mean = 5,27
Miller et al. (2014)	Multiple baseline design across participants	**N** = 3 1 Educational aide 1 Teacher's assistant 1 teacher (All female) **Setting**: SEN School	1 & 7	Implement DTT	M = 96% (Range: 87 – 100%) Strong	Data not provided	Specific targets not provided	Could not calculate NAP (data not provided)	Adequate	M: None G: None SV: 3-item social validity survey developed by the authors. Value range 1-5. M = 4.3/5.
Mouzakitis et al. 2015	Multiple baseline design across participants	**N** = 4 Special education teachers (All female) **Setting:** Inclusion programme in a mainstream school	A: 4 B: 1 & 4	Administer behaviour support plan	A: M = 93.43% (range: 80.08-100%) Strong B: M = 90.2% (range: 72.12-100%) Medium	N = 8 8 Male Children ASD	On-task behaviour	A: M = 83.92 (range: 48.82-100) Medium B: M = 51.68 (range: 40-83.71) Weak	Borderline adequate	M: None G: Included extra students for generalization SV:None
Pelletier et al. 2010	Multiple baseline design across participants (dyads)	**N =** 3 Teachers (2 female, 1 male) **Setting**: Residential and day school for children with autism	1, 2, 4, 9	Administer behaviour support plan	M: 100% Strong	N= 1 Adolescent Female ASD	Self-injurious behaviour, physical aggression.	Could not calculate NAP (data not provided)	Adequate	M: None G: None SV: None
Plavnick et al. 2010	Multiple baseline design across participants	**N** = 3 1 Teacher 2 Paraprofessionals (All female) **Setting**: Special education classroom	A: 11 B: 4	Administer token economy	A: M:100% Strong B: 97.2% (Range 92.86 – 100) Strong	N = 2 1 male, 1 female Children 1 ASD 1 Williams syndrome & language impairment	Appropriate vocalising and appropriate sitting	A: 89.17% Range (91.67 – 100%) Medium B: M= 99.34% Range = 98.68 – 100 Strong	Weak	M: None G: None SV: None

Pollard et al. 2014	Multiple baseline design across participants	**N** = 4 Undergraduate students from SEN course (All female) **Setting**: Office setting	1, 2, 6, 9	Implement DTT	M: 99.5% (Range: 98 – 100%) Strong	N = 2 Children ASD	Skill acquisition (targets included nonsense shapes and unknown colours	Could not calculate NAP (data not provided)	Adequate	M: None G: None SV: 8-item questionnaire developed by the authors. Value range: Strongly agree to strongly disagree (5 options)..Participants agreed or strongly agreed with all statements (one negative – that the videos did not always work properly).
Vince Garland et al. 2016	Multiple baseline across participants	**N=** 6 Masters level special education students (3 male 3 female) **Setting:** Research laboratory	1, 3, 5, 6	Implement system of least prompts	M: 100% Strong	N = 1 Simulation Avatar ASD & ID	Skill acquisition (specific targets not provided)	Could not calculate NAP (data not provided and study used an avatar)	Strong	M: At least 2 weekly maintenance phases per participant *6/6 ppts maintained intervention levels of implementation G: None SV: Focus group & 6 item social validity survey developed by the authors. Value range 1-5. All participants highly agreed with the usefulness of the teaching procedures.
Weinkauf et al. 2011	Changing criterion design	**N**= 4 Trainees in autism treatment programme (All Females) **Setting:** Therapy rooms at autism treatment centre (wonder if this could perhaps be classified as a clinical setting?)	1, 3, 4, 5, 6, 9	Increase skills identified as necessary for high procedural fidelity implementation (based on checklist of 125 behavioural skills)	M: 100% Strong	N = 4 Children ASD	Increase engagement, skill acquisition and appropriate transitioning	Could not calculate NAP (data not provided)	Borderline adequate	M: None G: None SV: None

Table 5 Participants characteristics and settings

Primary participants (*N* = 83)	**Percentage of participants**
Relation to secondary participant	
Teacher	41
Parent	19
Paraprofessional	16
Students	12
Front-line staff	11
Clinician	1
Gender	
Female	83
Male	17
Secondary participants (*N* = 72)	**Represented in percentage of studies**
Age group present in studies	
Children (0-11)	76
Adolescents (12-17)	18
Adults (18+)	4
Adolescent avatar character	2
Number of Diagnoses	
Single	74
Multiple	26
Type of Diagnosis	
ASD	75
ABI	10
Emotional disturbance	6
ADHD	6
Intellectual disability	2
Setting	
Special educational needs	42
Clinic/laboratory	21
Mainstream classroom	13
Residential facility for individuals with I.D.	4
Family home	4
Community	4

Table 6 Intervention Combination and effect sizes

Intervention	Belfiore et al. 2008	Codding et al. 2005	Codding et al. 2008	Coolican et al. 2010	Courtemanche et al. 2014	DiGennaro et al. 2005	DiGennaro et al. 2007	DiGennaro-Reed [illegible]	Flynn & Lo. 2015	Ingersoll & Wainer. 2013	Maggin et al. 2012	McKenney &Bristol 2015	Miller et al. 2014	Mouzakitis et al. 2015	Pelletier et al. 2010	Plavnick et al. 2010	Pollard et al. 2014	Vince Garland et al. 2016	Weinkauf et al. 2011	Intervention Effect Size, *M*		
																				% of studies intervention included in	Used in isolation	Used as part of a package
Feedback		X	X		X B C	X C	X B C D	X B				X	X	X B	X		X	X	X	22	Medium 70.83%	Strong 94.85%
Role-play					X B C	X C			X B		X A	X			X		X			12	N/A	Strong 97.07%
Modelling						X C	X	X A B	X B		X A	X						X	X	12	Strong 98%	Strong 99.10%
Self-monitoring	X													X A B	X	X B			X	9	Strong 96.54%	Strong 96.73%
Quiz/assessment					X A													X	X	5	N/A	Medium 97.33%
Teaching & instruction									X B		X A						X	X	X	9	N/A	Strong 100%
Financial incentive					X A B C								X							3	N/A	Medium 85.5%
Goal-setting							X C			X										3	N/A	Medium 90.36%
Observation		X	X		X B C							X			X		X		X	12	N/A	Strong 94.85%
Negative reinforcement					X B C	X C	X D													3	N/A	Strong 94.93%
Behaviour Skills training				X	X A	X A	X A		X A B	X	X B					X A				10	Strong 97.1%	Medium 78.5%

Note: The X in the table marks the interventions which are included in the study and the lettering included below indicates which phase of the study the intervention was included in. This detail matches the data in Table 4.

Table 7 Strength of effect size on participant target behaviour

Primary participants	Weak	78
	Medium	19
	Strong	3
Secondary participants	Weak	26
	Medium	63
	Strong	11