An Assessment of Treatment Integrity in Behavioral Intervention Studies Conducted with Persons with Mental Retardation

John J. Wheeler and Michael R. Mayton
Tennessee Technological University

Stacy L. Carter
Texas Tech University

Morgan Chitiyo
Southern Illinois University

Anthony L. Menendez
Cleveland State University

Ann Huang
Duquesne University

Abstract: The purpose of this study was to assess the degree to which behavioral intervention studies conducted with persons with mental retardation operationally defined the independent variables and evaluated and reported measures of treatment integrity. The study expands the previous work in this area reported by Gresham, Gansle, and Noell (1993) and Wheeler, Baggett, Fox, and Blevins (2006) by providing an evaluation of empirical investigations published in multiple journals in the fields of applied behavior analysis and mental retardation from 1996 –2006. Results of the review indicated that relatively few of the studies fully reported data on treatment integrity.

The field of applied behavior analysis has been steadfast in promoting evidence-based research as one of its central tenets toward ensuring sound scientific practice, yet in spite of this rigorous adherence to principle, the absence of uniformity found among behavioral intervention studies with respect to treatment integrity has been disappointing as documented in the literature. Treatment integrity, also commonly referred to as treatment fidelity, has been defined as the degree to which an independent variable has been implemented as intended (Peterson, Horner, & Wonderlich, 1982).

The lack of emphasis on reporting treatment integrity in intervention studies is alarming. The obvious ethical and methodological dilemmas imposed on researchers, practitioners and most importantly consumers, i.e., persons with disabilities, by these marginal research practices are disconcerting. Research in applied behavior analysis has strongly reinforced the importance of determining the presence of a functional relationship as a means of evaluating empirically-derived treatment outcomes (Peterson et al., 1982). Yet if one examines the trend in the literature from comprehensive reviews that have been conducted related to this question the rate at which treatment integrity has been reported remains abysmally low. Consider the findings from reviews that have been conducted to examine this practice in the literature.

Gresham et al., (1993) conducted a comprehensive review of applied behavior analysis studies published in The Journal of Applied Behavior Analysis between the years 1980–1990. Gresham and colleagues found that approximately 16% of the reviewed studies (n = 158)
assessed the accuracy of independent variable implementation, whereas two-thirds of the studies did not operationally define the independent variable. In a subsequent review, Gresham, MacMillan, Frankenberger, and Bocian (2000) examined this question as it pertained to intervention studies in the area of learning disabilities across a five-year period and published within three major journals in the field of learning disabilities. Gresham et al. also concluded that only 18.5 percent of reviewed studies assessed treatment integrity.

In a subsequent study conducted by Wheeler et al. (2006), behavioral intervention studies were reviewed that had been conducted with children ages 18 or younger who were diagnosed with autism. These studies had been published across nine different journals prominent in the area of autism spectrum disorders and/or developmental disabilities between the years 1993–2003. Of the studies reviewed (n = 60), only 11 studies (or 18%) operationally defined the independent variables and assessed treatment integrity.

In a more recent study, McIntyre, Gresham, DiGennaro, and Reed (2007) reviewed school-based intervention studies conducted with children ages 0–18 years and published in the Journal of Applied Behavior Analysis between the years 1991 and 2005. Of the 142 studies reviewed, only 30% provided data regarding treatment integrity. One could surmise from this investigation and the others preceding it that the absence of treatment integrity in behavioral intervention studies represents a consistent trend across journal sources and disability categories.

From a methodological standpoint, treatment integrity is important to both the internal and external validity of a study. The inability to draw reasonable conclusions concerning effects of the independent variable on treatment outcomes given the absence of such data is obviously hampered (Gresham et al., 1993; Kazdin, 1998). The absence of data concerning the implementation of the independent variable among published studies does not conclusively support a functional relationship and significantly impairs the ability of researchers to identify treatment efficacy (Wheeler et al., 2006). The second most obvious limitation imposed on researchers by the absence of treatment integrity is that it effects replication of the study by other researchers. Replication in a field such as applied behavior analysis is vitally important given that single subject research has often been criticized concerning its presumed lack of external validity, or the inability to generalize findings beyond a single experiment (Kazdin, 1982). Gresham et al. emphasized, “. . . failure to assess the degree to which treatments are implemented as planned compromises the science of building a replicative history” (p. 238). This is also important when considering that most single subject experimental studies have small n sizes by design, thus making replication most important for understanding the efficacy of an intervention across the range of settings, participants and conditions.

The purpose of this review was to assess the presence of treatment integrity as reported in behavioral intervention studies conducted with persons with mental retardation published in multiple journals across the years 1996–2006. Specifically, the study examined whether the reviewed studies operationally defined the independent variables, adequately reported interrater reliability data and evaluated and reported measures of treatment integrity.

Method

Criteria for Inclusion

The studies selected for review had to be published, experimental studies that: (a) utilized a single subject research design, (b) utilized behavioral treatments, and (c) were conducted with children and youth (ages birth to 18) and or adults (ages 18+) with a primary diagnosis of mental retardation (studies with participants who had dual diagnoses were rejected), (d) were found within the following peer-reviewed journals published from 1996 to 2006: American Journal on Mental Retardation (AJMR), Education and Training in Developmental Disabilities (ETDD), Journal of Applied Behavior Analysis (JABA), Mental Retardation (MR), and Research in Developmental Disabilities (RDD). Articles within these journals were located by first searching a series of online databases (PsychLit, ERIC, Info Trac, and the journals’ respective websites) and then following up to find any missing journals by searching collections of print copies.
Procedure

Once selected, studies were coded using a matrix that contained the following fields: (a) journal name, (b) article title, (c) year published, (d) number and ages of participants, (e) presence or absence of an operational definition of the independent variable(s), (f) type of experimental design, (g) interrater reliability by sessions and overall percentage, and (h) an assessment of treatment integrity. If a study assessed treatment integrity, provided information regarding the method of assessment utilized, and provided relevant data, the study was coded as “yes.” If treatment integrity was discussed but either or both the assessment method and/or data were not provided, the study was coded as “monitored.” If there was no mention of treatment integrity within a study, it was coded as “no” (see Table 1).

Interrater Agreement

The first and second authors independently read and scored 50% of the studies to determine interrater reliability. Agreement was calculated by dividing the number of agreements plus disagreements and multiplying by 100. Percentage of agreement across all studies for all variables within the coding matrix was 98%.

Results

Treatment Integrity

Once collected and codified, treatment integrity data were analyzed across the following variables: (a) by journal, (b) year of publication, (c) age of participants, and (d) by age and journal. This descriptive analysis was performed to reveal any trends or variances within and among these variables that could be used in a broader discussion of the utilization of treatment integrity procedures.

By journal. Figure 1 displays the treatment integrity reporting trends for each of the five journals examined in this study. The three journals that had the most studies matching search criteria were JABA, ETDD, and RDD (N = 81, 48, and 22, respectively). Among these journals, ETDD had the highest percentage of studies reporting treatment integrity procedures and results (65%). JABA and RDD had very similar profiles across yes, no, and monitored classifications, though it should be noted that the RDD search revealed significantly fewer behavioral studies than did the search of JABA. Results derived from the remaining journals, AJMR and MR, should be interpreted with caution due to the relatively low number of studies identified.

By year of publication. Figure 2 displays the treatment integrity reporting trends for each calendar year included in the present study. Over the eleven-year examination period, the number of studies that failed to report treatment integrity data (or even mention that such procedures were conducted) appears to drop in frequency (for the first five years, M = 12.6, compared to M = 4.2 over the last six years), but so does the overall number of behavioral studies (for the first five years, M = 46.2, compared to M = 8.8 over the last six years). However, a visual comparison of the yes and no classifications within each year reveals that more recent years (e.g., 2005 and 2006) show a greater majority of studies that
reported treatment integrity procedures and results than that found in previous years.

By age of participants. Figure 3 compares trends in reporting treatment integrity between the children and youth and adult categories. Studies involving children and youth (ages 0–18) as participants reported treatment integrity data a little more often (an 11% difference) than studies involving adults (ages 19+). Percentages of studies that either failed to mention treatment integrity at all or reported only that treatment integrity procedures had been conducted looked very similar in both general age groupings. The 11% difference in the yes condition for studies including adult participants appears to be evenly distributed across no (5% difference) and monitored (6% difference) conditions.

By age and journal. Figure 4 displays the trend in reporting treatment integrity data for children and youth and adults by each journal included in the present study. ETDD and RDD both had larger numbers of studies that reported treatment integrity procedures and results for participants in the children and youth category (17% and 36% more, respectively). In the adult category, however, ETDD had a near even split between yes and no treatment integrity ratings (55% yes, compared to 45% no), while RDD had a much higher number of articles receiving a no rating (a 49% difference between yes and no). MR received mostly monitored ratings for both age categories (across three studies that included five adults and two children, with one study failing to report individual ages of participants), and the AJMR data reported more treatment integrity information for children and youth than for adults, although frequent monitoring of treatment integrity appeared for adults (87%).
Operational Definitions of Independent Variables

Figure 5 displays the trends in reporting operational definitions of independent variables by each of the journals included in the present study. JABA, ETDD, and RDD had acceptably high percentages of operational definitions of reported independent variables (91–96%), though any study that does not adequately describe independent variables makes replication unlikely. The extremes within these results (100% and 67%) were derived from the two journals with the fewest number of identified behavioral studies and should therefore be interpreted with caution.

Interrater Reliability

Interrater reliability (IRR) data were collected from each article and analyzed to determine the following for each journal: (a) the minimum and maximum percentages of sessions examined for IRR, (b) the mean percentage of sessions examined for IRR, (c) the percentage of articles that failed to report the percentage of sessions examined for IRR, (d) minimum and maximum IRR, (e) mean IRR, and (f) the percentage of articles that failed to report IRR. These data were then tabulated for comparative examination (see Table 2).

Kennedy (2005) recommended that a minimum of 20% of observations/sessions be examined for IRR. Results of the current study showed that minimum sessions examined for IRR ranged from 7% (AJMR) to 19% (MR), and the maximum for all journals was 100%. Mean sessions for all journals were in the acceptable range (40.9%–72.8%).

Zirpoli (2008) stated that an IRR of 70–80% is adequate, though results become increasingly desirable as they approach 100%. Results of the current study showed that minimum IRR across all journals ranged from 70% (RDD) to 93% (AJMR), and the maximum for all journals except one was 100%.
Mean IRR for all journals was in the desirable range (90%–98.4%).

Summary and Conclusions

The purpose of this review was to determine to what extent treatment integrity had been addressed in behavioral intervention studies (n = 163) conducted with persons with mental retardation published within multiple journal sources (n = 5) during and between the years of 1996–2006. This study extends previous work in the area, most notably Gresham et al., (1993), however with the main difference being the present study examined behavioral intervention studies conducted with persons with mental retardation across both children/youth and adults and published across multiple journal sources.

Of the 163 studies that were analyzed, only 36% (n = 58) fully reported data on treatment integrity. These results are slightly improved from previous studies, for example, Peterson et al., (1982), who reported 20% across 530 studies, Gresham et al., (1993) who reported 16% of 158 studies, and Wheeler et al., (2006), reporting 18% of 60 studies. Although the current results suggest an improvement, there remains significant need for better performance outcomes in this area.

Considering the trend reported from other reviews, one can only speculate as to the reasons surrounding the lack of treatment integrity data in the treatment literature. Perhaps the problem stems from a lack of understanding as to the importance of such data or is due to the fact that these data are simply not reported due to space considerations imposed during the review process, leaving one to question how reviewers weigh treatment integrity on the scale of importance when reviewing a
manuscript. Results from the current investigation and previous reviews lead one to believe that treatment integrity is not an essential piece when considering whether to publish a paper.

Nonetheless, when implementing a behavioral intervention study, researchers should be advised to assess treatment integrity in a reliable manner by first operationalizing their respective treatments in measurable and observable terms. Gresham et al., (2000) has provided some meaningful recommendations for researchers pertaining to this. First, they stated that treatment integrity should be measured over time, similar to the concept of time generalizability. This method will result in repeated measures that can assist in the analysis of treatment integrity over time (Gresham et al.). They also recommended that two measures of treatment integrity be taken, the first being a measure of each treatment component across days of treatment, and a second measure taken to yield session integrity. Greshman and colleagues asserted that a failure to realize significant treatment effects can be potentially explained by either less than optimal component integrity or less than optimal session integrity measures. Secondly, treatment integrity involves the direct observation of treatment implementation by trained observers, or through the use of videotape analysis (Gresham, et al.; Wheeler, Carter, Mayton, & Thomas, 2002).

Methods used in ensuring treatment integrity include the training of those personnel responsible for the implementation of the treatment using a pre-identified protocol that outlines the various component steps that comprise the treatment. One example of this can be found in a study conducted by Wheeler.
et al. (2002). As a means of ensuring treatment integrity, the teacher who was responsible for implementing the treatment, during a structural analysis in this case, was provided with a structured presentation of procedures. This training consisted of: (a) rationale, (b) directions on how to present the task, (c) directions on how to respond to the child during the condition, (d) modeling of the procedures and (e) rehearsal of the procedures with performance feedback provided. This training was conducted one day prior to the actual assessment, with a brief rehearsal being conducted 30-minutes prior to the session. For observers monitoring treatment integrity, a checklist or behavior rating scale can

![Figure 5. Percentage of articles reporting operational definitions of the independent variables.](image)

**TABLE 2**

Reported Interrater Reliability by Journal

<table>
<thead>
<tr>
<th>Journal</th>
<th>Min/Max Sessions</th>
<th>Mean Sessions</th>
<th>Not Reported</th>
<th>Min/Max IRR</th>
<th>Mean IRR</th>
<th>Not Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJMR</td>
<td>7–100%</td>
<td>72.8%</td>
<td>11%</td>
<td>93–100%</td>
<td>98.4%</td>
<td>11%</td>
</tr>
<tr>
<td>ETDD</td>
<td>16–100%</td>
<td>44.3%</td>
<td>13%</td>
<td>86–100%</td>
<td>97.2%</td>
<td>2%</td>
</tr>
<tr>
<td>JABA</td>
<td>17–100%</td>
<td>40.9%</td>
<td>9%</td>
<td>81–100%</td>
<td>95.8%</td>
<td>4%</td>
</tr>
<tr>
<td>MR (N = 3)*</td>
<td>19–100%</td>
<td>59.5%</td>
<td>33%*</td>
<td>90%*</td>
<td>90%</td>
<td>33%*</td>
</tr>
<tr>
<td>RDD</td>
<td>15–100%</td>
<td>46.9%</td>
<td>0%</td>
<td>70–100%</td>
<td>95%</td>
<td>0%</td>
</tr>
</tbody>
</table>
be developed that will provide observers with a mechanism to evaluate the implementation of treatment (Gresham et al., 2000; Wheeler et al., 2002) either through in-vivo observations or through videotape analysis.

In summary, the field needs to be more steadfast in promoting the inclusion of treatment integrity measures in behavioral treatment studies, as a measure of quality research that is evaluated before publication. If we are to truly advance the concept of evidenced-based practices, the work begins at home, in this case, with those who conduct and review research.

References


Received: 29 February 2008
Initial Acceptance: 20 April 2008
Final Acceptance: 10 July 2008