Review

Review of reliability and treatment integrity trends in autism-focused research

Leslie Neely *, Heather Davis, John Davis, Mandy Rispoli

Texas A&M University, United States

A B S T R A C T

This review summarizes interobserver reliability and treatment integrity trends across five journals in autism-focused single-case research: *Focus on Autism and Other Developmental Disabilities, Journal of Applied Behavior Analysis, Journal of Autism and Developmental Disorders, Research in Autism Spectrum Disorder* and *Education and Training in Autism and Developmental Disabilities.* All articles published in the selected journals for the years 1992, 2002, and 2012 were screened based on predetermined inclusion criteria. A total of 119 articles met inclusion criteria and were evaluated in terms of whether they collected reliability data on dependent and independent variables, whether they collected treatment integrity data, as well as the degree to which the data met accepted quality criteria. Trends indicate that the collection and quality of the interobserver agreement data collection for dependent variables is improving. However, the percentage of studies meeting the minimum requirement for collection of interobserver agreement in each phase of the study remains low. Trends for the collection of the treatment integrity data and the quality of the treatment integrity data remains low but is demonstrating an increasing trend. Trends for the interobserver agreement data for the independent variables remain stable and low. Limitations of this review as well as implications for future research are discussed.

© 2014 Elsevier Ltd. All rights reserved.

Contents

* Corresponding author at: Department of Educational Psychology, Texas A&M University, 4225 TAMU, College Station, TX 77843, United States. Tel.: +1 979 862 1120.
    E-mail address: cockeril@neo.tamu.edu (L. Neely).

1. Introduction

The prevalence of autism spectrum disorder (ASD) is rising with a current estimate of 1 in 68 children diagnosed with the disorder (Center for Disease Control [CDC], 2014). Perhaps due to the nature of the disorder, or lack of a cure, consumers of ASD therapies and treatments are particularly susceptible to fad and controversial treatments (Metz, Mulick, & Butter, 2005). Fad treatments for ASD tend to not only be detrimental to the individual, but a waste of time and resources (Metz et al., 2005). To combat fad treatments there is both a practical and ethical need to evaluate ASD treatments with rigorous scientific experimentation and measurement.

Rigorous measurement is the cornerstone of data-based decision making for evaluating outcomes of student behavior in school settings. High stake decisions, such as changes in instructional placement for students demonstrating challenging behavior, often rely on standardized measures of the behavior (Parker, Vannest, Davis, & Clemens, 2012). While most standardized behavior assessment tools are developed upon the central tenants of classical test theory (e.g., reliability and validity), not all behaviors can be evaluated using standardized measures (AERA, APA, & NCME, 1999). In particular students identified with ASD often demonstrate idiosyncratic behaviors, which may be outside the scope of most standardized measures. These behaviors may be more precisely measured through systematic and direct observation methods (Odom et al., 2005).

When conducting research concerning the behaviors of individuals with ASD, single-case research methodology may be utilized in order to investigate individual differences (Horner et al., 2005; Kratochwill et al., 2013). Single-case methodology has been recently recognized as important in evaluating low-incidence issues, such as behaviors in individuals with ASD (Horner et al., 2005). This methodology is identified by the repeated measurement of the dependent variable on an individual level (Horner et al., 2005; Kratochwill et al., 2013). Data can be collected for individual participants or groups of participants serving as a unit. Since single-case studies are often designed to meet the specific behavioral concerns of individuals, replications and measurement of the dependent and independent variables are required to establish internal validity within designs (Kratochwill et al., 2013).

To facilitate high quality research IES founded the What Works Clearinghouse (WWC). A recent focus by the WWC on single-case methodology has led to the establishment of single-case design standards (Kratochwill et al., 2013). These standards assist researchers in evaluating the validity of the study designs and inform researchers on best practices in conducting high-quality research. These standards were intended to guide researchers and interventionists in developing, selecting, and implementing evidenced based practices within single-case research (Kratochwill et al., 2013).

The recognition of specific evidence-based practices by WWC highlights the importance of collecting and reporting IOA data for dependent variables in single-case experiments. The reporting of IOA data for single-case research is a pivotal concern as the measurement of each outcome variable must be conducted by independent assessors in an effort to ensure the accuracy of the data (Kratochwill et al., 2013). The use of two raters observing and recording the behavior simultaneously to assess the reliability and validity of the measurement allows for the calculation of inter-observer agreement (IOA; Foster, Bell-Dolan, & Burge, 1988). While leaders in the field of single-case research converge on the importance of the inclusion of reliability data and calculation of IOA (Council for Exceptional Children [CEC], 2014; Kratochwill et al., 2013), there has been an ongoing debate on the best method to collect and report IOA (Artman, Wolery, & Yoder, 2012; Kratochwill & Wetzel, 1977).

There is a strong tradition of utilizing either percent agreement or Cohen's *kappa* coefficient in single-case research (Cohen, 1960; Hartmann, Barrios, & Wood, 2004; Parker, Vannest, & Davis, 2013). Simple percent agreement indices are calculated by dividing the total number of agreements by the total number of recorded observations (Suen, 1988). Percent agreement is calculated to assess exact agreement between observers. One criticism of this method is that the measurement does not take into consideration the possibility of "chance agreements", which in turn, can overstate the calculated IOA index (Suen, 1988). Limitations identified in the percent agreement index led researchers to identify alternatives to calculating IOA indices. Recommendations for the use of Cohen's *kappa* (Cohen, 1960) have been made as an alternative. Cohen's *kappa* adjusts for the possibility of "chance agreements", which makes *kappa* a more conservative reliability index when reporting IOA data (Cohen, 1960; Hintze, 2005).

While Cohen's *kappa* is often considered an improved reliability index over percent agreement, published studies continue to rely on percent agreement indices in single-case research (Watkins & Pacheco, 2000). A recent review by Artman et al. (2012) published in the *Journal of Applied Behavior Analysis* found 100% of the 58 articles that reported IOA data used percent of agreements estimates, while only one reported Cohen's *kappa*. While there is a clear tendency to utilize percent agreement methods in reporting IOA in single-case research, the applications of both percent agreement and Cohen's *kappa* methods are subject to review based on minimum quality standards.

The WWC single-case standards clearly outline the quality requirements for the collection of IOA data on the dependent variable in single-case studies. As described by Kratochwill et al. (2013), in order to meet the minimum quality standards, IOA data should be collected within each phase in the study, for at least 20% of the data points in each phase. Current minimal