# A COMPARISON OF TREATMENT INTEGRITY ASSESSMENT METHODS FOR BEHAVIORAL INTERVENTION

by

Seong A Koh

---

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF DISABILITY
AND PSYCHOEDUCATIONAL STUDIES

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPY
WITH A MAJOR IN SPECIAL EDUCATION

In the Graduate College

THE UNIVERSITY OF ARIZONA

2010

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by Seong A Koh

entitled A Comparison of Treatment Integrity Assessment Methods for Behavioral Intervention

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy

_____ Date: 04.16.10
Dr. John Umbreit

_____ Date: 04.16.10
Dr. Carl Liaupsin

_____ Date: 04.16.10
Dr. Jolenea Ferro

_____ Date: 04.16.10
Dr. Stephanie Z. C. MacFarland

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 04.19.10
Dissertation Director: Dr. John Umbreit

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Seong A Koh

# ACKNOWLEDGEMENT

Since I enrolled in the EBD program of the University of Arizona, I have met wonderful people. Even though I conducted this study, this dissertation is the outcome of the collaboration and support of many people. I'd like to express my gratitude to: Dr. John Umbreit, my committee chair and the mentor in this field, who has given me endless support; Dr. Darrell Sabers, a great supporter who unfailingly shared his time and knowledge even though he was not one of my committee members. Dr. Sabers introduced and taught me G-theory when I was in the midst of a conflict analyzing the data. Additional thanks goes out to Dr. Carl Liaupsin, who generously contributed his time and energy in helping me to conduct this study successfully; Dr. Jolenea Ferro, who encouraged me to think differently when I confronted a problem, and whose constructive criticism allowed me to clarify the context of this paper; and Dr. Stephanie MacFarland, who continuously encouraged me, believing that I could overcome all difficulties.

Also, I would like to thank those people who opened the gate of my new exploration: Dr. Kwang-Sun Cho Blair, Dr. James Chalfant, Dr. Kathleen Lane, and Dr. June Maker. In addition, I offer my gratitude to my cohort, Donna M. Janney, Brenna K. Wood, Linda Reeves, and Matthew Hoge. Because of their friendship I have had a joyful and meaningful time in a foreign country. Dr. Kay Nelson and Dr. Martha Underwood, I will not forget your efforts to transform my paper from Korean English to American English. Lastly I thank you, my raters. Because of your participation I was able to conduct this research successfully. All of you are in my daily prayer.

# DEDICATION

I dedicate this dissertation to my parents, Kim Hyo-soon and Koh Jung-myung, to Anna Koh Varilla and her family, to Koh Seok-ho and his family, to my community, the Sisters of Charity in Korea and the United States, and finally to my students and colleagues.

*Ad Majorem Dei Gloriam!*

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

# LIST OF TABLES

Table

LIST OF TABLES – Continued

# LIST OF FIGURES

Figure

# ABSTRACT

The purpose of this study was to examine the similarity of outcomes from three different treatment integrity (TI) methods, and to identify the method which best corresponded to the assessment of a child's behavior. Six raters were recruited through individual contact via snowball sampling. A modified intervention component list and 19 video clips were derived from Stahr's (2005) study, "An Intervention for Children with Autism Spectrum Disorders (ASD) who Have Food Selectivity." The raters, randomly and evenly assigned to three dyads. Each dyad received an average of six hours training and reached 85% interobserver agreement (IOA) with a 0.60 kappa score. After training, each dyad watched 5 video clips per day and measured both the child's behavior and TI. The percentages of IOA, kappas, and indices of dependability for assessment of the child's behavior and TI were analyzed. The data revealed that all raters reached over 80% IOA and the whole interval (WI) and yes/no (Y/N) dyads reached .60 kappa, but the two raters in the Likert-type (LIK) dyad could not reach .60 kappa. The indices of dependability indicated that the six raters consistently observed and rated both the child's behavior and TI, but there was a discrepancy in scores (i.e., percentages of TI) between the two raters in the two indirect measure dyads (i.e., Y/N and LIK). An analysis of the percentages of total variance showed that the two indirect TI methods may affect the discrepancy between the two raters' rating scores. A comparison of the three different TI methods and correlation between the child's behavior and TI were examined using the *PASW Statistics 18* software program. There was no significant difference between the WI and the Y/N

dyads, while the assessments from the LIK dyad indicated a significant difference from the other two dyads. Both the WI and the Y/N dyads showed correlations between the degree of the child's behavior and the degree of TI, but there was no significant difference between the two correlation coefficients. Questions about reliability with the indirect TI measures suggest one should be careful in considering these results.

# CHAPTER 1

## INTRODUCTION

Treatment integrity (TI), also known as treatment fidelity or procedural reliability, can refer to at least three different processes (Noell, 2008). First, it can refer to the degree to which a researcher or a consultant adheres to an established research or consultation model. Second, it can refer to the extent to which procedures in the established model are implemented as designed by the researcher or consultant. Lastly, TI can be used to show the degree to which an intervention is implemented as planned. However, TI is generally used to refer to the degree to which an intervention is implemented as planned (Billingsley, White, & Munson, 1980) and to reflect whether the practitioner accurately and consistently implemented each component of an intervention (Lane, Bocian, MacMillan, & Gresham, 2004).

Various researchers have emphasized the importance of assessing treatment integrity in the applied behavior analysis research field (e.g., Gresham, 1989; Horner, Carr, Halle, McGee, Odom, & Wolery, 2005; Wilkinson, 2006). TI indicates that a change in the dependent variable (e.g., problem behavior) is directly related to the independent variable (e.g., intervention or treatment) rather than to any other factors (Lane, Bocian, et al., 2004; Wood, Umbreit, Liaupsin, & Gresham, 2007). For example, Gresham, Gansle, and Noell (1993) found that, if the intervention is not fully implemented, the internal and external validity are threatened. Similarly, Shadish, Cook, and Campbell (2002) stressed that unreliable implementation of treatment threatens statistical conclusion validity.

Not only in experimental design, but also in the practical world, the importance of

assessing TI is more critical than in the past. McIntyre, Gresham, DiGennaro and Reed (2007) pointed out that, because of recent legislation such as the No Child Left Behind Act (NCLB) and the Individuals with Disabilities Education Improvement Act (IDEIA) of 2004, practitioners and teachers should show if they accurately implement interventions or instruction plans over time. Moreover, the U.S. Department of Education's Institute of Education Sciences (IES) requires not only that "applicant should describe how they will assess the degree to which the training on the identified practices occurs as planned (fidelity of implementation of the intervention or transfer process)" (IES, 2009, p. 9) but also that "applicants should describe how they will establish the reliability and validity of the measure" (IES, 2009, p. 10) if they need to develop a new measure of TI. Therefore, researchers regard assessing TI as an essential element of school-based behavioral intervention research (Gresham et al., 1993); this includes function-based intervention (Lane, Umbreit, & Beebe-Frankenberger, 1999).

However, how does one conduct this assessment? TI can be assessed through either direct observation or indirect observation. For direct observation, observers (i.e., raters) check whether each step or component of intervention correctly occurred during observation by using a component checklist (e.g., Campbell & Anderson, 2008). However, Umbreit, Ferro, Liaupsin, and Lane (2007) suggested a different method that uses whole intervals to measure TI for direct observation. In this method, raters simultaneously assess TI while assessing a student's behavior. The assessment form is the same as typical interval measurement.

For indirect observation, measurement can be done in several ways: (a) a yes/no

component checklist after observation, (b) behavior rating scales (e.g., Likert-type scale), (c) self-report by using either yes/no component checklist or Likert-type component checklist, (d) permanent products (Umbreit, et al., 2007), and (e) interview (Cochrane, & Laux, 2008). Lane and Beebe-Frankenberger (2004) recommended the use of a four-point Likert-type scale (ranging from low integrity to high integrity), but currently some researchers have used a three-point scale in their studies (e.g., Lane, Kalberg, Bruhn, & Driscoll, 2008; Lane, Weisenbach, Phillips, & Wehby, 2007; Robertson, & Lane, 2007). The examination of permanent products is a more recently applied indirect method of assessing TI (Lane & Beebe-Franenberger, 2004). If an outcome of an intervention creates permanent products, such as the completed worksheet or printed document, the information in the products can be used as evidence that the intervention was implemented as planned (Umbreit, et al., 2007). Although interviewing the intervener is rarely used by educational researchers, school psychologists frequently use this method to assess TI. Indeed, Cochrane and Laux (2008) reported that 60.6% of 426 respondents who assessed TI in school sites used the interview method.

In conclusion, determining a TI measurement method is essential. As Peterson, Homer, and Wonderlich (1982) warned, unreliable and inadequate assessment of the independent variable can lead to a false functional relationship between the dependent variable (i.e., target behavior) and the independent variable (i.e., behavioral intervention). Therefore, such experts as Lane and Beebe-Frankenberger (2004), Lane, Kalberg, Bruhn, Mahoney, and Driscoll (2008), and Umbreit and colleagues (2007) have recommended carefully considering which TI measurement method to use.

## Statement of the Problem

Bare, Wolf, and Risley (1968) claimed that researchers should ensure that behavioral changes between experimental phases reflect alternation of the participant's behavior, and not the behavior of the instructors (e.g., teacher or experimenter) and/ or observers. Bakeman and Gottman (1997) suggested that, with systematic observation, properly trained observers should produce the same stream of behavior when given identical protocols. Therefore, the researchers concluded that "the personal qualities of the observers should not matter" (p. 3). However, currently, researchers not only report discrepancies between TI results produced by different TI measurement methods in the same observation, but also bring up issues related to measurement methods.

Specifically, Smith, Daunic, and Taylor (2007) argued that because self-report often inflates the degree of implementation relative to direct observation findings, direct observation of TI would be considered stronger evidence than indirect measures such as self-report. Also, Lane, Kalberg, et al. (2008) reconfirmed that teachers rated their TI higher than the degree of TI which a research assistant rated and stressed the importance of considering the TI measurement method. However, Lane, Weisenbach, et al. (2007) reported that teacher participants could attain a high level of TI inter-observer agreement (IOA) in function-based assessment and intervention. That is, when teachers and outside observers assessed TI by using the same rating scale checklist after observation, TI IOA could reach 100%.

Wood, Umbreit, Liaupsin, and Gresham's research (2007) demonstrated that, even though direct observation was the primary form of data collection, the results of correct

implementation depended on which form was used. Wood et al. (2007) argued that a whole interval assessment would reflect the precise effects of an intervention better than a component checklist completed after observation. These researchers suggested that more research would be needed to determine whether whole-interval assessment is more useful than a component checklist. Zvoch (2009) pointed out a critical limitation of a yes/no component checklist. Because the checklist requires "observers to treat the delivery of various program components as binary. . . . even the most knowledgeable and best trained observers are often forced to make a series of difficult 'all or none' decisions" (p. 56).

In summary, even though TI and its assessment methods have been discussed in the literature, it is hard to know whether any gap between two different TI measurement methods is caused by lack of training of raters before collecting TI data or by systematic variance in the assessment methods. Also, much of the literature regarding intervention for students with emotional and behavioral disorders (EBD) reports neither treatment integrity data (Mooney, Epstein, Reid, & Nelson, 2003) nor assessment methods (McIntyre, et al., 2007). Therefore, there is a need to investigate TI assessment methods in intervention research.

## Research Questions

The purpose of this study was (a) to examine whether different TI measurement methods produced similar TI data, and (b) to identify which method best corresponded to the dependent measures (i.e., changes in the child's behavior). This study examined TI data generated by two indirect measures (i.e., yes/no and Likert-type) and one direct observation measurement (whole-interval).

For the first inquiry (i.e., comparing three different TI assessment methods), specific questions addressed were:

1. Do raters assess the child's behavior similarly?
2. Do the three different TI assessment methods produce similar TI outcomes?

For the second inquiry, the specific questions were:

3. Is there a relationship between the degree of the child's on-task behavior and the degree of TI indicated by each TI measurement method?
4. If there is a correlation between the degree of the child's on-task behavior and the degree of TI indicated by each method, is there any difference among the level of correlations?

# CHAPTER 2

# REVIEW OF THE LITERATURE

Hagermoser Sanetti and Kratochwill (2008) stressed that researchers who use experimental designs cannot make valid conclusions about the functional relationship between a dependent variable and an independent variable without formative treatment integrity assessment. Peterson et al. (1982) emphasized that, although the true value of independent variable is unknown, if description of the independent variable is clear, the observed value should closely approximate the true value. As a result, researchers can assume that an independent variable's true value is the value specified by the experimenter. Therefore, several issues related to TI measurement have been discussed in the literature (e.g. Yeaton & Sechrest, 1981; Billingsley et al, 1980; Gresham, Gansle, & Noell, 1993; Gresham, Gansle, Noell, & Cohen, 1993; Wolery, 1994). Horner, Carr, Halle, McGee, Odom and Wolery (2005) identified TI as a quality indicator in single-subject research and required reporting TI "either through continuous direct measurement of the independent variable, or an equivalent" (p. 168) based on Gresham, Gansle, and Noell's (1993) recommendation. However Gresham (1989) criticized that TI has been assumed rather than assessed and empirically demonstrated.

## Previous Literature Review Outcomes

Hagermoser Sanetti and Kratochwill (2008) have provided comprehensive information about current issues in TI research and have also discussed these issues with respect to the behavior consultation model. The researchers summarized current TI

research outcomes in the following areas: (a) rationale of TI report in research, (b) TI measurement, (c) the relationship between TI and treatment outcomes, and (d) enhancing the degree of TI. As a result of reviewing articles, the researchers concluded that TI measurement is "one of the greatest challenges" (p. 98) in understanding TI.

In terms of TI measurement, researchers have been working in at least two directions. One is to review studies for identifying how commonly TI data were reported in intervention literature; the other is to explore any relationship between the degree of TI and treatment/intervention outcomes. While Peterson et al (1982) were first, several researchers (e.g., Gresham & Gansle, 1993; Gresham, Gansle, & Noell, 1993; Gresham, MacMillan, Beebe-Frankenberger, 2000; Griffith, Hurley, & Hagaman, 2009; McIntyre et al, 2007; Wheeler, Baggett, Fox, & Blevins, 2006; Wheeler, Mayton, Carter, Chitiyo, Menendez, & Huang, 2009) have investigated whether researchers report TI data in the literature in several different disability areas.

Gresham et al. (1993) and McIntyre et al. (2007) conducted literature reviews on research articles published in the *Journal of Applied Behavior Analysis* (JABA) from 1980 to 1990 and 1991 to 2005 to extend the research of Peterson, et al. in 1982. Since Peterson and the colleagues investigated TI data reporting percentages in JABA for behavioral intervention, the overall percentage of TI report had not changed very much, from 16% (Peterson et al., 1982) to 15.8% (Gresham et al., 1993), and then to 30.3% (McIntyre et al., 2007). In comparison, the rate of reporting operational definitions of independent variables dramatically increased, from less than 20% (Peterson et al., 1982) to 34.2% (Gresham et al., 1993) to 95% (McIntyre et al., 2007).

Peterson and colleagues (1982) warned that inadequate assessment of the independent variable could contaminate any conclusion on relationship between independent and dependent variable; however, in literature reviews, it is hard to find information such as TI measurement method, the percentage of sessions of TI data collection, and TI observer training. With the exception of a study by Griffith, Hurley, and Hagaman (2009), researchers who conducted literature review studies did not report data which were related to these issues. Griffith and colleagues categorized data-collecting methods into two categories – self-report and observation. The self-report category included rating scales, yes/no checklists, and interviews. The researchers coded observation either as direct or as indirect observation: direct observation was conducted at the time of the intervention, whereas indirect observation was conducted later by using video or audio recordings. The researchers reported that the main method used to assess TI was observation ($n$ = 17; 74%) and that 81% of observation was direct observation. Only one study used self-report, and 4 studies combined both observation and self-report. However, the researchers did not report what methods were used to collect TI data during observation. Also, these researchers neglected to mention one common indirect TI measurement method – the yes/no component checklist – after conducting observation.

In addition, there was no information how many sessions included TI data collection. Yet, the number of observation sessions is directly related to the representation of observational data (Gresham et al., 1993). Indeed, McIntyre et al. (2007) considered that a small number of TI assessment would constitute a high risk for treatment inaccuracy. Peterson and colleagues (1982) criticized researchers' double standard toward gathering

data on dependent variables but not on the independent variable. The “curious double standard” (Peterson, et al., 1982, p. 478) has also occurred in reporting IOA on TI. Even though researchers collect data for IOA on the dependent variable, it is not common to report TI IOA.

As subset inquiries, researchers also have paid attention to examining the relationship between the degree of TI and the effect size of outcomes (e.g., Gresham & Gansle, 1993; Griffith, et al., 2009), as well as changes in reporting TI data in the literature over time (Gresham, et al., 1993; Griffith et al., 2009; McIntyre et al., 2007; Wheeler et al., 2006; Wheeler et al., 2009). Gresham and Gansle (1993) reported effect sizes of reviewed studies when examining the relationship between the level of TI and intervention outcomes by using Cohen’s *d* for group studies (*n* = 23) and the percentage of nonoverlapping data points (PNOL; Mastropieri & Scruggs, 1985) method for small *N* studies (*n* = 158). Effect size estimates for the 23 studies of group design ranged from .32–3.00 with a mean effect size of .839 (*SD* = .731). In the study, 158, single-subject experimental studies were used to compute PNOL. The range of PNLO was from 40% to 100% and the mean of PNLO was 92.84% (*SD* = 10.48%). Correlation between the degree of TI and the treatment outcome was calculated using percentage of TI, effect size for group design studies, and PNLO for single subject design studies. The result was $r = .51$ ($p < .05$) between percentage of TI and effect size, and $r = .58$ ($p < .05$) between percentage of TI and PNLO. Based on the results, the researchers concluded that higher treatment integrity was associated with larger effect sizes.

Griffith and colleagues (2009) reported similar results about the relationship

between the degree of TI and the effect size of treatment outcomes, to Gresham and Gansle's (1993) report. The researchers obtained effect sizes from 38 studies in the learning disabilities literature by using Cohen's *d* for group studies (*n* = 4) and Swanson and Sachse-Lee's (2000) method that is based on Rosenthal's formula (cited in Swanson & Sachse-Lee, 2000) for small *N* studies (*n* = 34). Effect size ranged from .42 to 2.93, with a mean of 1.74. Also, the researchers found that almost of all of the reviewed studies reported higher levels of TI (i.e., greater than 90%). However, the researchers suggested that small variability in the TI data made it difficult to examine the relationship between the degree of TI and intervention outcomes.

Even though the two reviews of the literature (i.e., Gresham & Gansle, 1993; Griffith et al., 2009) infer that treatment outcome correlates to degree of TI to a certain level, several studies that tried to explore the relationship between the degree of TI and the treatment outcome by intentionally manipulating level of TI showed mixed results. Further analysis of these studies indicated the following: (a) there is a direct correlation between TI and treatment outcomes (e.g., Dib & Sturmey, 2007; DiGennaro, Martens, & Kleinmann, 2007; Henggerler, Melton, Brondino, & Scherer, 1997; Rhymer, Evans-Hampton, McCurdy, & Watson, 2002; Vollmer, Roane, Ringdahl, & Marcus, 1999; Wilder, Atwell, & Wine, 2006)), (b) there are inconsistent results among participants (e.g., Holcombe, Wolery, & Snyder, 1994; Noell, Witt, Cilbertson, Ranier, & Freeland, 1997; Sansosti & Powell-Smith, 2006), (c) there is no relationship between TI and treatment outcomes (e.g., Gansle & McMahon, 1997; Northup, Fisher, Kahang, Harrell, & Kurtz, 1997), and (d) there is an unclear relationship between the level of TI and the degree of

behavior change (Hagermoser Sanetti & Kratochwill, 2009).

However, in case of inconsistent results among participants, some researchers tried to answer the question as to why a particular student did not respond to intervention where other students improved their behaviors under high treatment integrity. As a result, researchers suggested causes such as (a) the student's high rate of absence (Noell et al., 1997), (b) caregiver's failure to implement an intervention at home (Sansosti & Powell-Smith, 2006), and (c) a history of unsuccessful learning with low level of treatment integrity, multi-treatment interference and the end of the school year (Holocombe et al., 1994).

Also, Taylor and Miller's (1997) study gives a reason for mixed research results. The researchers examined the impact of the function of student behavior problem on treatment efficacy and treatment integrity. Based on experiment results, the researchers concluded that even though the level of TI was one factor that affected treatment effectiveness, the function of a participant's problem behavior played a key role in impacting treatment effectiveness (e.g., Northup, Fisher, Kahang, Harrell, & Kurtz, 1997). That is, a high degree of TI does not ensure successful treatment outcomes (in the case of Gansle & McMahon, 1997) unless interventions are function-based. Therefore, the researchers suggested that through functional assessment, the interventionist can avoid an ineffective intervention component.

In addition, barriers related to TI measurement mentioned by Noell (2008) may contribute to understanding inconsistent results. Noell pointed out three main barriers. First, usually a behavioral intervention has several components, but whether each

component is equally important is undetermined. Second, there is no well-developed standard to calculate degree of TI. Lastly, each TI measurement method has its own limitation: direct observation can bring observation reactivity, permanent products do not represent all treatment components, and the data of self-reports by the treatment agent can be biased.

In conclusion, there is no arguing that there are a lot of uncertainties related to TI measurement and its relationship to the degree of TI and intervention outcome. The uncertainty may affect the report of TI in published research. Nevertheless, it appears that treatment integrity is "a crucial question in translating research to practice" (Hall, 1998, p. 294), that dependent variable measurement alone cannot reveal clear conclusions about the source of behavior change, and that "a curious double standard" (Peterson et. al., 1982, p. 478) is still working on behavioral intervention research.

## Literature Review for Treatment Integrity Assessment Methods

Based on outcomes and suggestions from previous research, TI measurement methods and related issues in the literature between 2006 and 2008 were examined. The procedures used by McIntyre et al. (2007) and Wheeler et al. (2006) were modified as criteria for review. Based on the recommendation of three experts in EBD field, five journals which might represent in EBD field were selected for analysis: *Behavioral Disorders* (BD), *Education and Treatment of Children* (ETC), *Journal of Applied Behavior Analysis* (JABA), *Journal of Emotional and Behavioral Disorders* (JEBD), and *Journal of Positive Behavioral interventions* (JPBI).

## *Review Methods*

### *Criteria for Inclusion*

*Criteria for initial inclusion.* All articles (excluding book reviews, remembrances, and reports) published between 2006 and 2008 in the five journals were reviewed by using a hand search to determine possible inclusion. For inclusion, six criteria which were modified from McIntyre, et al.'s (2007) and Wheeler, et al.'s (2006) criteria were considered.

First, the study should be experimental to examine the effects of intervention on behavior or academic support for people with disabilities. Studies without a clear baseline or control condition, such as an AB design or no control group, were excluded. Second, studies which were conducted in inpatient units were excluded, whereas outpatient center classrooms or day treatment classrooms were included. Even though the climate of juvenile correctional facility is similar to an inpatient unit, these studies were included because juvenile correction centers typically offer both academic and behavioral intervention. Third, if a study conducted a functional assessment or analysis followed by an intervention the study was included. However, articles that only included assessment were excluded. Fourth, although McIntyre et al. (2007) only reviewed school-based intervention studies, this review included studies with interventions for people with disabilities. Specifically, in the case of school-based intervention studies, interventions focusing not only on behavior problems but also on academic supports for students with EBD were included. In addition, studies on school-wide positive behavior support (SW-

PBS) were included. Fifth, research that utilized community-based interventions for people with disabilities who were over 19 years old were included. Lastly, indirect intervention studies which focused on training teacher, staff, or parents – that is, studies in which the level of intervention implementation was dependent variable – were excluded.

McIntyre et al. (2007) excluded articles of three or fewer pages in length, but there was no article that was three or fewer pages between 2006 and 2008. Therefore, the length of the articles was not considered as a criterion in this study.

*Criteria for the final inclusion*. Based on the initial inclusion, the selected studies were examined if the studies operationally defined independent variable and reported TI measures.

*Procedure for Coding*

*The initial coding procedures*. Once studies were selected based on the initial inclusion criteria, the selected studies were then coded according to two categories for selecting studies for the final inclusion: (a) operational definition of independent variable(s), and (b) assessment of treatment integrity.

- Operational definition of independent variable(s): Based on Gresham, Gansle, and Noell's criterion (1993), this was coded as "yes" or "no" according to the criterion in the following statement: "*If you could replicate this treatment with the information provided, the intervention is to be considered operationally defined*" (p. 258). In addition, whether the procedures specifically described the

treatment components, action (e.g., duration, group or individual, frequency of implementation) and any necessary materials was considered (Horner et al., 2005). If a study used manualized intervention and provided contact information or name of the program, the study was coded as “yes.”

- Treatment integrity assessment: The studies were coded based on reporting TI data. If a study reported both TI measurement method and degree of TI, the study was coded as “yes”; it was coded as “no” if it failed to report both. If a study reported TI measurement method but failed to report data or vice versa, it was coded as “monitored.”

*The final coding procedures*. Studies which included both operationally defined independent variable(s) and TI measurement were coded according to seven categories: (a) age level of participants, (b) intervention setting, (c) intervention structure (e.g., multi-components or single), (d) TI data collector, (e) TI measurement methods, (f) percentage of sessions for TI data collection, and (g) TI IOA report.

First, age level of participants were divided into 5 groups, such as preschool age (i.e., under 5 years old), kindergarten to eighth grade (6 yrs. to 14 yrs.), high school (15 yrs. to 18 yrs.), over 19 years old, and mixed age group. Second, the location of the intervention setting was divided into 5 categories such as school, home, clinic or center, community facility, and multiple places. Third, an intervention structure was identified based on two categories – intervention type and the possibility of multiple occurrences. Intervention type was potentially classified as antecedent based, consequence based, a teaching of

alternative behaviors intervention, or a multiple component intervention. If an intervention used one specific type, it was identified as a *single structure*; if an intervention combined different types (e.g., antecedent change and using positive reinforcement) and had several procedures, the intervention was identified as a *multiple structure*. If all intervention components or steps are occurred one time during a session, it was marked as "single" for possibility of occurrence. If several trials occur during a session or some components/ steps in an intervention could occur several times during a session, it was marked as "multiple."

Fourth, treatment integrity data collectors were divided into 6 categories: *outer observer, experimenter, intervention agent, multi-raters, inside rater* such as a class aid or school staff, and *no information*. Fifth, treatment integrity measurement methods were divided into 4 categories: *direct measure, indirect measure, self-report*, and *permanent products*. The use of an interval assessment form was coded as a direct measure. An interval form is one that observer(s) used if all required intervention components at an interval were correctly implemented. Either partial interval or whole interval was used in the selected studies. In the case of using a yes/no (Y/N) component checklist, the form was classified by two different styles. Either rater(s) checked whether each component was implemented as planned in general, or rater(s) checked each component implementation on a trial-by-trial basis. When researchers reported that they used a component checklist *during* observation, it was coded as direct measure.

Indirect measures were those in which rater(s) assessed TI *after* observation by using a Likert-type (LIK) rating scale or Y/N component checklist. LIK rating scales

were analyzed based on a “point” for each component. A rating scale could be 3-point, 4-point or 5-point based. In the case of the Y/N component checklists, there was no subcategory because this method checks general implementation for each step or component and then calculates the percentage of overall implementation of TI. When a rater used a Y/N component checklist and the researcher failed to report whether TI data were collected during observation, it was coded as indirect measure.

The last two categories of TI measurement methods were self-report and permanent products. In self-report, the intervention agent assesses the level of one’s implementation after implementing the intervention by using either a Y/N component checklist or a LIK rating scale. Additionally, permanent products, such as a point card, work sheet, or journal which correspond with an intervention component, also were marked as an TI measurement method when the researcher used a permanent product as a measurement method.

Based on the recommendations of Gresham et al. (1993), the percentages of sessions for TI were reviewed. The categories were divided into (a) 10% to less than 25%, (b) 25% - less than 50%, (c) over 50%, (d) 100%, (e) others, and (f) no report. Lastly, IOA of TI was reviewed. Criteria were (a) no report, (b) 70-79%, (c) 80-85%, (d) 86-90%, and (e) greater than 90%. Also, if TI IOA was reported, the percentage of sessions for TI IOA was reviewed together.

## *Review Results*

### *Articles Selected for Inclusion*

Twenty four studies in BD (*Behavior Disorders*), 24 studies in ETC (*Education and Treatment of Children*), 47 studies in JABA (*Journal of Applied Behavior Analysis*), 5 studies in JEBD (*Journal of Emotional and Behavioral Disorders*), and 26 studies in JPBI (*Journal of Positive Behavioral interventions*) met the criteria for the first inclusion. Among the 126 studies, only 62 reported both an operationally defined independent variable and TI data (see Table 1). These 62 studies were used for the final inclusion.

Table 1

*Articles Selected*

| Journal ($n$ = 126) | Independent Variable | Treatment Integrity | | |
|---|---|---|---|---|
| | | Yes | No | Monitored |
| BD (24) | OD[a] | 12 (50.0%) | 4 (16.7%) | 6 (25.0%) |
| | NOD[b] | 0 | 2 (8.3%) | 0 |
| ETC (24) | OD | 19 (79.1%) | 4 (16.7%) | 0 |
| | NOD | 0 | 1 (4.2%) | 0 |
| JABA (47) | OD | 13 (27.7%) | 31 (65.9%) | 1 (2.1%) |
| | NOD | 0 | 2 (4.3%) | 0 |
| JEBD (5) | OD | 3 (60.0%) | 2 (40.0%) | 0 |
| | NOD | 0 | 0 | 0 |
| JPBI (26) | OD | 15 (57.7%) | 9 (34.6%) | 2 (7.7%) |
| | NOD | 0 | 0 | 0 |

*Note.* Abbreviation of Journal name: BD (*Behavior Disorders*), ETC (*Education and Treatment of Children*), JABA (*Journal of Applied Behavior Analysis*), JEBD (*Journal of Emotional and Behavioral Disorders*), JPBI (*Journal of Positive Behavioral interventions*).

[a] OD = Operationally defined. [b] NOD = Not operationally defined.

The overall percentage of TI reported in the five journals is 49.2%. When the current TI reporting percentage was compared to previous reports (e.g., the 2007 study of McIntyre et al.), the TI reporting percentage in BD, ETC, JEDB, and JPBI was higher than that in previous studies. In the case of JABA, the current rate (27.7%) was only slightly lower than the previous level (i.e., 30.3%).

*Age Level of Participants*

All of the selected studies targeted school-age populations with the age level of 6 to 14 as the main target (66.1%; see Table 2). Among the initially selected 162 studies, 12 focused on behavioral interventions for adults, but there was no assessment of the degree of treatment integrity.

Table 2

*Age Level of Participants*

| Journal / Age level of participants (%) [a] | BD (*n* = 12) | ETC (*n* = 19) | JABA (*n* = 13) | JEBD (*n* = 3) | JPBI (*n* = 15) |
|---|---|---|---|---|---|
| Under 5 years old (17.7%) | 0 | 2 | 7 | 0 | 2 |
| 6 to 14 (66.1%) | 10 | 14 | 4 | 3 | 10 |
| 15 to 18 (12.9%) | 2 | 2 | 1 | 0 | 3 |
| Over 19 (0%) | 0 | 0 | 0 | 0 | 0 |
| Mixed (3.2%) | 0 | 1[b] | 1[c] | 0 | 0 |

[a] The percentages do not sum exactly to 100 due to rounding.

[b] 5th to 12th graders. [c] 3years to 8 years old.

*Intervention Setting*

Almost all of the studies were conducted in school setting, but some were conducted in different environments (Table 3). However, the TI reporting rate was very low among studies conducted in either community environments or clinic settings.

Table 3

*Intervention Setting*

| Journal / Intervention setting (%) | BD (*n* = 12) | ETC (*n* = 19) | JABA (*n* = 13) | JEBD (*n* = 3) | JPBI (*n* = 15) |
|---|---|---|---|---|---|
| School (82.3%) | 11 | 19 | 9 | 2 | 10 |
| Home (3.2%) | 0 | 0 | 1 | 0 | 1 |
| Center/ clinic (1.6%) | 0 | 0 | 1 | 0 | 0 |
| Community facility (4.8%) | 1[a] | 0 | 1[b] | 0 | 1[c] |
| Multi-places (8.1%) | 0 | 0 | 1[d] | 1[d] | 3[e] |

[a] Juvenile correctional facility. [b] Preschool summer camp. [c] Two grocery stores and two department stores 4 public settings [d] School and home. [e] School and vocational training center, school and home, clinic and home.

*Intervention Structure*

Almost the interventions (73.4%) included multiple types, multiple components or steps, and multiple opportunities for each component to occur (Table 4). In the case of multiple types, interventions included antecedent change, positive reinforcement, and/or teaching new behavior. An example of single type and single occurrence occurs with social story intervention. The experimenter simply required parents to read the story to their children and let the parents report whether they read the story (Sansosti, & Powell-Smith, 2006). Self-monitoring of an intervention is an example of single-multiple interventions. In this case, the student checked whether or not he/she was on-task in six 10-minute intervals (Gulchak, 2008).

Table 4

*Intervention Structure*

| Journal / Type - opportunity to occur (%) | BD ($n$ = 12) | ETC ($n$ = 19) | JABA ($n$ = 13) | JEBD ($n$ = 5[a]) | JPBI ($n$ = 15) |
|---|---|---|---|---|---|
| Single – Single (7.8%) | 0 | 2 | 0 | 1 | 2 |
| Single – Multiple (6.3%) | 0 | 0 | 3 | 0 | 1 |
| Multiple – Single (12.5%) | 1 | 1 | 2 | 1 | 3 |
| Multiple – Multiple (73.4%) | 11 | 16 | 8 | 3 | 9 |

[a] One study includes three different interventions.

*TI Data Collector*

As shown in Table 5, when TI data were collected, researchers used outer observers (*n* = 20; 32.3%), collected TI data by themselves (*n* = 18; 29.0%), or let at least two different types of raters (e.g., outer observer and teacher) collect the data (n = 11; 17.7%).

Table 5

*TI Data Collector*

| Data collector (%) \ Journal | | BD (*n* = 12) | ETC (*n* = 19) | JABA (*n* = 13) | JEBD (*n* = 3) | JPBI (*n* = 15) |
|---|---|---|---|---|---|---|
| **O**uter observer (32.3%) | | 4 | 9 | 5 | 1 | 1 |
| **E**xperimenter (29.0%) | | 4 | 2 | 4 | 1 | 7 |
| **I**ntervention **A**gent (11.3%) | Teacher (3.3%) | 1 | 0 | 0 | 0 | 1 |
| | Therapist (4.8%) | 0 | 0 | 1 | 0 | 2 |
| | Parent (1.6%) | 0 | 0 | 0 | 0 | 1 |
| | Participant (1.6%) | 0 | 1 | 0 | 0 | 0 |
| Multi-raters (17.7%) | O & E (4.8%) | 0 | 1 | 2 | 0 | 0 |
| | O & IA (6.45%)[a] | 1 | 2 | 0 | 1 | 0 |
| | E & IA (6.45%) | 1[b] | 2[c] | 0 | 0 | 1[d] |
| Inside rater (6.5%) | | 1 | 1 | 0 | 0 | 2 |
| No information (3.2%) | | 0 | 1 | 1 | 0 | 0 |

[a] Outer observer and teacher. [b] Experimenter and peer. [c] Experimenter and teacher. [d] Experimenter and parents.

*TI Measurement Methods*

As shown in Table 6, the percentage of direct TI measures (47.7%) is higher than that of indirect measures (35.8%). Y/N component checklists were generally used for both direct (*n* = 24; 35.8%) and indirect measure (*n* = 16; 23.9%). When LIK rating scales were used, almost all the studies used a 3-point scale.

Table 6

*TI Measurement Methods*

| TI measurement methods (%) \ Journal | | | BD (*n* = 16) [a] | ETC (*n* = 19) | JABA (*n* = 13) | JEBD (*n* = 3) | JPBI (*n* = 16)[a] |
|---|---|---|---|---|---|---|---|
| Direct measure (47.7%) | Interval (11.9%) | | 1 | 4 | 1[b] | 0 | 2 |
| | checklist (35.8%) | overall | 7 | 7 | 2 | 0 | 2 |
| | | trials | 1 | 0 | 5 | 0 | 0 |
| Indirect measure (35.8%) | Rating scale (11.9%) | 3-point | 2 | 2 | 0 | 2 | 1 |
| | | 4-point | 0 | 0 | 0 | 0 | 0 |
| | | 5-point | 0 | 0 | 1 | 0 | 0 |
| | Checklist (23.9%) | | 1 | 5 | 4 | 1 | 5 |
| Self-report | checklist (6.0%) | | 1 | 0 | 0 | 0 | 3 |
| | Rating scale (6.0%) | | 3[c] | 0 | 0 | 0 | 1[d] |
| Permanent product (4.5%) | | | 0 | 1 | 0 | 0 | 2 |

[a] When experimenter(s) used multiple raters or interventions in a study, different methods were used. [b] Frequency. [c] One 5-point rating scale, two 3-point rating scale.
[d] 4-point rating scale.

*Percentage of Data Collecting Sessions for TI*

As shown in Table 7, nearly one third of the studies (33.9%) collected TI data in 25% to 50% of observation sessions. The percentages of TI data collection were similar to that of IOA for the dependent variable(s). Even though 14.5% of selected studies collected TI data during all intervention phases, one fourth of the studies either scarcely collected TI data during intervention (16.1%) or did not report the number of TI data collection sessions (8.1%). Therefore, as McIntyre et al. (2007) pointed out, high risk of treatment inaccuracy is apparent in the studies.

Table 7

*Percentage of Data Collecting Sessions for TI*

| Journal / Percentage of sessions (%) [a] | BD ($n$ = 12) | ETC ($n$ = 19) | JABA ($n$ = 13) | JEBD ($n$ = 3) | JPBI ($n$ = 15) |
|---|---|---|---|---|---|
| 10 - less than 25% (21.0%) | 4 | 2 | 2 | 1 | 4 |
| 25 - less than 50% (33.9%) | 4 | 4 | 5 | 1 | 7 |
| Over 50% (6.5%) | 1 | 1 | 2 | 0 | 0 |
| 100% (14.5%) | 1 | 6 | 0 | 1 | 1 |
| Others [b] (16.1%) | 1 | 5 | 2 | 0 | 2 |
| No report (8.1%) | 1 | 1 | 2 | 0 | 1 |

*Note*. The percentages in parentheses indicate the percentages of each category.

[a] The percentages do not sum exactly to 100 due to rounding.

[b] Under five times. One study (Yurick et al., 2006) used descriptive word “intermittently.”

*TI IOA report*

Among 62 studies, only 18 reported percentages of TI IOA (Table 8). In almost all (15 studies out of 18 studies), TI IOA reached over 90 percent. Five studies in BD reported TI IOA, but no study reported the percentage of sessions. In the cases of studies in ETC, two studies reported the percentage of data collecting sessions, and the studies collected TI IOA data during 25% of observation sessions. Three studies out of six in JABA reported the percentages of sessions, and the percentages were 25%, 32%, and 100%. Only one study in JPBI reported both TI IOA and the percentage of observation sessions for TI IOA. The researcher collected TI IOA data during 33% to 40% of observation sessions.

Table 8

*TI IOA Report*

| Journal / Percentage of TI IOA (%) | BD ($n$ = 12) | ETC ($n$ = 19) | JABA ($n$ = 13) | JEBD ($n$ = 3) | JPBI ($n$ = 15) |
|---|---|---|---|---|---|
| 70s | 0 | 0 | 0 | 0 | 0 |
| 80 – 85 (1.6%) | 1 | 0 | 0 | 0 | 0 |
| 86 – 90 (3.2%) | 0 | 0 | 2 | 0 | 0 |
| Over 90 (24.2%) | 4 | 6 | 4 | 0 | 1 |
| No report (71.0%) | 7 | 13 | 7 | 3 | 14 |

*Note.* The percentages in parentheses indicate the percentages of each category.

### *Summary and Discussion*

For almost three decades, researchers have stressed the importance of assessing TI and have discussed issues related to TI. However, many researchers and publications still failed to assess and report TI. The percentage of TI reported in the literature has not changed very much over the years, whereas the percentage of reporting operational definitions of the independent variable has increased dramatically. Furthermore, the percentages of TI reported vary from journal to journal. For example, the percentage of TI report in ETC was 79.1%; in JABA, it was 27.7%.

Nevertheless, studies that reported the degree of TI were inclined toward school-based interventions. It is noticeable that community-based behavioral intervention programs for both school age and adult population are not common. Although some have been conducted, they have not adequately reported TI data. Also, studies that were clinic-based inadequately reported the degree of TI, even though they reported positive outcomes. As Gresham (1989) warned, if any positive outcome occurs but there are no TI data, the conclusions are questionable.

Many interventions (73.4%) that were analyzed here included multiple components and steps or procedures. Even single-component interventions tended to include multiple opportunities or several trials. One positive outcome from this review was that several studies in the *Journal of Applied Behavior Analysis* assessed the degree of TI on a trial-by-trial basis during observation sessions. These attempts should be encouraged and examined for their adequacy in assessing the degree of TI compared to a Y/N checklist. Rating scales were used infrequently. When they were used, almost all of the studies used

3-point rating scales.

Peterson et al. (1982) commented that experimenter bias may affect the observation of independent variables. When collecting TI data, in over one fourth of studies, the experimenters themselves collected the TI data (29.0%). However, several studies (17.7%) used multiple raters so that the degrees of TI could be compared. Lane and Beebe-Frankenberger (2004) suggested assessing TI from multiple perspectives so that the accuracy of the TI measure might increase.

There is no consensus among researchers about the percentage of sessions in which TI data collection should occur. For example, one sample of a TI form (see Lane & Beebe-Frankenberger, 2004, p. 135) indicates that TI data should be collected everyday along with data collection on the dependent variables. Umbreit et al. (2007) suggested that, even though TI data should be assessed daily, researchers should consider practical factors that might lead observers to collect TI data once or twice a week rather than in every session. McIntyre and colleagues (2007) suggested that if IOA data are collected in 35% of observations, 15% of sessions could be used for TI assessment, and data on IOA for the dependent variable could be collected in the other 20% of observations. Consequently, there is no consensus about the appropriate percentage of sessions for collecting TI data to ensure representative of data. Over one fourth of studies ($n$ = 18; 29%) reported TI IOA, and most of all studies ($n$ = 15) reported over 90% TI IOA.

# CHAPTER 3

# METHODS

This study analyzed different TI assessment methods by using existing behavioral intervention video clips, taken from "An Intervention for Children with Autism Spectrum Disorders (ASD) who Have Food Selectivity" conducted by Stahr in 2005. The video clips were recorded under IRB approval granted by Vanderbilt University, where the study was conducted. Before conducting this research, material access permissions were obtained from the principal investigator (PI) of the previous study, as well as from the parents whose child participated in the food selectivity research.

## Participants and Setting

### *Participant Recruiting Procedure*

The PI used snowball sampling, and contacted potential participants via e-mail. Through this process, in total, there were six eligible raters, one alternate, and one outer observer recruited for this research. The alternate and the outer observer served as procedural fidelity raters. All eight participants provided written consent for voluntary participation in the study.

*Attrition.* During the training sessions for the Likert-type rating scale dyad, one rater dropped out. Therefore, the alternate participated in this study as a rater. Only two extra training days were required to reach 85% of IOA with the new rater in the Likert-type dyad.

*Grouping*

The six raters picked their identification (ID) numbers (i.e., as primary or secondary observer for the dyad) and their TI measurement method by lot. The first dyad chose the yes/no (Y/N) TI method, the second dyad the Likert-type (LIK) TI method, and the last dyad the whole-interval (WI) TI method. Table 9 presents information about the participants.

Table 9

*Information of Participants*

| | | Participants | |
|---|---|---|---|
| ID | Dyad | Earned Degree | Current currier status |
| S1 | WI | M.A. | Doctoral student in Ed. Psych |
| S2 | WI | Ph. D | Assistant professor |
| S3 | Y/N | M.A. | Teacher of ED private program |
| S4 | Y/N | M.A. | Doctoral student in Special Ed. |
| S5 | LIK | M.A. | Doctoral student in Special Ed. |
| S6 | LIK | M.A. | Doctoral student in Ed. Psych. |

*Setting*

This study was conducted in a conference room in the College of Education building in which a projector, tables, and chairs were available. The schedule for training and rating sessions were decided jointly by raters and the PI.

## Definitions

*Treatment Integrity (TI)*

In this study, the term of “treatment integrity” is used to refer to the extent to which a behavioral intervention is implemented as planned.

*Procedural fidelity (PF)*

In this study, procedural fidelity refers the degree to which the research procedures are implemented as designed. That is, through procedural fidelity, training sessions and rating sessions for the three dyads were evaluated as to whether or not each session was procedurally conducted as planned.

## Variables

Based on the questions, variables can be differently considered.

*Variables for Question 1: Do Raters Assess the Child’s Behavior Similarly?*

*Dependent variable.* The dependent variable was the child’s replacement behavior, modified from the original food consumption intervention. The replacement behavior was on-task behavior which was defined as (a) staying in the table and seat area, (b) waiting for the instructor’s task direction, (c) picking up the spoon (or food if fingers were used) within 5 seconds when a new food item was given, (d) taking the spoon or food (if fingers were used) to the mouth within 5 seconds, (e) chewing food or touching the tongue to the food (if the instructor asked for it) with or without physical assistance, or (f) following the instructor’s direction.

*Independent variable*. The independent variable was the intervention designed to improve food consumption. The original intervention focused only on food consumption. No systematic procedures were developed to address off-task behavior in the original study. The interventionist used various practices to get the child to continue participating in the food study sessions. However, for the present study, it was necessary to identify a standard set of procedures for responding to off-task behavior that the raters could evaluate. These procedures (steps 9-11, below) were not known to the interventionist because they were created post hoc for the TI assessments conducted in this study.

The eleven components (or steps) of the intervention, adopted and modified from the original food study for the purpose of this study, were as follows:

1. When child is ready (i.e., seated or standing in seat area; mouth empty), place a bite on the plate and give task direction “Take a bite” or “OK, let’s try this one.”
2. If child takes a bite, give praise (e.g., “good job”) him.
3. If child does not pick up the spoon or food within 5 seconds, use hand-over-hand physical assistance to place the child’s hand on the spoon or to let him pick up food.
4. If child does not move the spoon or food to his mouth within 5 seconds, use hand-over-hand physical assistance to guide the spoon or food to his mouth.
5. If child does not open his mouth, return the spoon or food to the plate.
6. After any refusal, begin new trial with 1/2 bite and repeat 1-5.
7. If child refuses to take 1/2 bite, begin new trial with 1/4 bite and repeat 1-5.

8. If 1/4 bite is not consumed, verbally prompt child to use his tongue to touch the bite (“Touch tongue to bite”).
9. If child escapes from table, briefly redirect (short statement) him with minimal interaction.
10. Give up to 3 redirections (10 sec. between warnings). If child does not return to table after 3 redirections, physically bring child back to table and then repeat 1-5 with a smaller bite.
11. If child plays with a cracker, ignore him.

*Variables for Question 2: Do the Three Different TI Assessment Methods Produce Similar TI Outcomes?*

*Dependent variable.* The dependent variable was TI outcome using each method.

*Independent variables.* The independent variables were the TI measurement method and video clip used for observation. The ten video clips showed behaviors of the same child and of the same instructor at different phases in the original study. The video clips included assessment, baseline, intervention, and probes that were randomly sequenced

*Variables for Question 3 & 4: Which Method Best Corresponds to the Changes in the Child’s Behavior?*

*Two variables for correlation.* The child’s on-task behavior and TI outcomes were two variables. By using the percentages of intervals of the child’s behavior and the percentages of TI, the PI calculated correlation between the degree of child’s behavior and the degree TI for each dyad.

# Materials

*Procedural fidelity Checklist (PFC)*

The PI developed two types of procedural fidelity checklist for the training (Appendix B.1 and B.2) and the rating (Appendix B.3) sessions, each session lasted 60-90 min. All PFCs were Y/N step-by step checklists.

*Data Recording Sheets*

The PI designed four instruments for recording data, basing each on previously existing materials (more information, see Umbreit, et al., 2007, and Lane, & Beebe-Frankenberger, 2004). These instruments were *Interval Data Recording Form* (Appendix C.1) for WI dyad, *Child Behavior Data Recording Form* (Appendix C.2), *[Yes/No] Component Checklist* (Appendix C.3), and *Likert-type Rating Scale* (Appendix C.4).

The *Interval Data Recording Form* included a 24-interval table for both the child's behavior and intervention. The PI separated the child's behavior table from the Y/N component checklist and LIK checklist so that the child's behavior recording sheet could be distributed before observation, whereas the Y/N checklist or LIK rating scale could be distributed *after* each observation for Dyad 2 and 3. In order to control for confounding variables, which can affect rating results for the indirect observation methods, the raters in Dyads 2 and 3 were asked not to take any notes regarding treatment integrity during their observation of each video clip.

*Video Clips*

As Baer, Wolf, and Risley (1965) acknowledged, the majority of behavioral interventions are complex and include several different components. The video clips used in this study presented a complex intervention that allowed raters to assess a wide range of procedures. The food consumption intervention included several components related to antecedent, reinforcement, prompting, and extinction procedure. It also included several trials in each session, and the number of trials in the video clips differed. As a result, several different numbers of possibilities for implementing each component occurred. Therefore these characteristics allowed for checking TI in different ways.

From the available original 34 intervention video clips (i.e., 3 assessments, 4 baselines, 3 probes, and 24 intervention sessions), the PI chose a total of 27 video clips, including all assessment and baseline sessions and probes as well as 17 randomly selected video clips from the intervention phase. The length of the original video clips varied from 90s to 20 min. The PI edited the selected video clips to a standard length of 6 min so there would be (a) sufficient opportunity to observe variability in implementation and (b) a sufficient number of intervals to establish variability. Using these guidelines, it was possible to construct a total of six non-intervention video clips and 14 intervention video clips ($n$ = 20).

Three non-intervention video clips and 7 intervention video clips were randomly assigned for training or rating sessions. As a result, there were 10 video clips for training and 10 for rating. However, one of the training video clips was excluded from the final selection because it included substantial variance from the intervention protocol, leaving

nine video clips for training and 10 video clips for rating sessions. Because the video clips were randomly assigned and labeled, there was no time order, which minimized the likelihood of auto-correlation. For training sessions, the video clips were labeled from T 1 to T 9. For the rating sessions, the video clips were labeled from VC 1 to VC 10. Table 10 shows video clip arrangement for the training and rating sessions.

Table 10

*Video Clip Arrangement*

| Training sessions | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| T 1 | T 2 | T 3 | T 4 | T 5 | T 6 | T 7 | T 8 | T 9 | |
| I | I | I | P | A | I | I | I | I | |
| **Rating sessions** | | | | | | | | | |
| VC 1 | VC 2 | VC 3 | VC 4 | VC 5 | VC 6 | VC 7 | VC 8 | VC 9 | VC 10 |
| I | B | I | I | I | B | I | I | A | I |

*Note*. I = Intervention phase; B = Baseline; A = Assessment; P = Probes

When edited, the video clips included twenty four 15-second intervals. Signals were inserted in each video clip to mark these intervals, the observation starting point, and the finish. Those video clips were shown to the raters according to clip label number. However during the training sessions, when a dyad needed extra training sessions after watching all 9 video clips, the primary observer picked the number of video clip to be used among 9 video clips for training by lot on each training session.

## Experimental Design

This research was conducted in two phases.

### *Phase I. Training for Observation*

In this phase, nine video clips were available. Interobserver agreement (IOA) on both the child's behavior and the treatment integrity (TI) was collected. Training continued until IOA reached an average level of 85% and 0.65 for Cohen's *k* for three consecutive observations on both the child's behavior and TI. Rater S1, S3, and S5 were the primary rater in each dyad. The two raters in each dyad had to reach the IOA standard, and each rater and the PI functioning as the primary rater had to reach the standard as well (Figure 1).

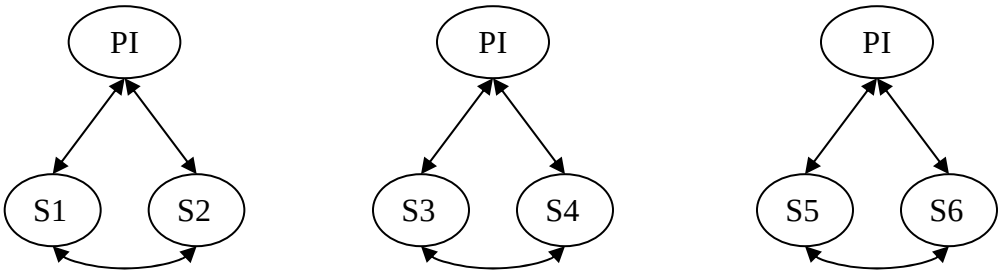


*Figure 1*. Training for TI and the child's behavior IOA between two raters and the PI

Each dyad scheduled its own training sessions with the PI. Basically, each dyad received three training trials on Training session I (Appendix B.1) and on the average, five training trials on Training session II (Appendix B.2). Each trial included watching,

rating, and discussing each video clip. Table 11 presents the steps of the first-day training (i.e., Training session I)

Table 11

*First Day Training Process (Training Session I)*

| Steps |
|---|
| 1. Confirm the consent form |
| 2. Confirm CITI completion: submit CITI completion report |
| 3. Identify dyad, primary observer, and second observer by lot |
| 4. Introduce intervention in video clips |
| 5. Watch a video clip |
| 6. Discuss replacement behavior and intervention procedures |
| 7. Distribute and explain assessment form and measurement method including how to calculate IOA |
| 8. Training 1: Watching video clip T 1 |
| 9. Calculate IOA for the child's behavior and treatment integrity |
| 10. Discuss the result, Q & A, and collect assessment sheets. |
| 11. Training 2: Distribute assessment sheets and watching video clip T 2 |
| 12. Calculate IOA for the child's behavior and treatment integrity |
| 13. Discuss the result, Q & A, and collect assessment sheets. |
| 14. Training 3: Distribute assessment sheets and watching video clip T 3 |
| 15. Calculate IOA for the child's behavior and treatment integrity |
| 16. Discuss the result, Q & A, and collect assessment sheets. |
| 17. Break (10 minutes) or Finish |

*Dyad 1 (WI)*. This dyad was trained to rate both the child's behavior and treatment integrity at the same time while watching each video clip by using the 15-s whole-interval method. Specifically, at the end of each interval, the raters scored a plus ("+") if

all required intervention components were correctly implemented throughout the entire interval. If the interventionist failed to implement any part of the intervention at any point during an interval, it was scored as a minus ("-"). Raters used the same measurement system to score the child's replacement behavior.

*Dyad 2 (Y/N).* First, this dyad recorded the child's behavior using the same 15-s whole-interval method described for the first dyad. Then, *after* watching a video clip, the raters immediately assessed TI by using a Y/N component checklist for each intervention component.

*Dyad 3 (LIK).* This dyad also recorded the child's behavior using the same 15-s whole-interval method described for the first two dyads. Dyad 3 assessed TI by using a four-point (0 to 3 and n/a) Likert type scale for each step immediately *after* watching the video clips.

*Phase II. Rating*

In this phase, data were collected for both TI and the child's behavior as they were for the training sessions. IOA for both the child's behavior and treatment integrity also was assessed as before. There were two rating sessions for each dyad.

Each rater in a dyad independently rated TI and the child's behavior during or after watching video clips, based on the particular TI measurement method being used. A total of 5 video clips were watched per session. If one rater finished the assessment earlier, the rater waited quietly until the other rater finished rating. When both were done, the raters submitted their measurement sheets immediately and then had a 4-5 min break before

rating another video clip. There was no discussion during the rating session and the raters were asked not to share their rating results with each other during a break time.

## Data Collection and Analysis

### *Procedural fidelity*

The PI was the primary rater for procedural fidelity (PF), collecting the data during the entire training and rating sessions for each dyad by using the PF checklists. An outer observer independently checked the procedural fidelity by using the same checklist for IOA. The level of the procedural fidelity is presented as a percentage of procedure steps implemented. The degree of procedural fidelity was computed by dividing the number of steps present by the total number of steps in the procedure and then multiplying by 100.

### *The Replacement Behavior of the Child*

All raters collected data on the replacement behaviors during *Phase II* by using the whole interval measure. The percentage of intervals for the replacement behavior was then calculated.

*Agreement of two raters within a dyad.* Agreement over the entire ratings of the child's behavior can be evaluated by using Pearson's correlation (Kazdin, 1982). Therefore, Pearson's product-moment correlation coefficient was used first, and then the percentages of IOA and Cohen's *k* values were compared. An 80% IOA or 0.6 kappa score was considered acceptable (see Horner et al., 2005), but 85% IOA and 0.65 kappa score were set as the standard in this study.

- The percentages of IOA: The use of an interval assessment form allows assessment of whether there is agreement on each interval for the observed behavior (Kazdin, 1982). Therefore the degree of IOA can reveal whether two raters similarly observed and rated the child's replacement behavior. In this study, S1, S3, and S5 were the primary raters. Interval-by-interval agreement ratio was used for the percentage of IOA. The percentage of IOA was calculated as follows:

$$\% \text{ of IOA} = \frac{\text{N. of agreements}}{\text{N. of agreements + N. of disagreements}} \times 100$$

- Cohen's *k*: A common problem with the percentage score of IOA is that some agreement could occur just by chance alone (Bakeman & Gottman, 1997). There is also a high possibility of a high level of agreement if a target behavior occurs with a relatively high frequency. Kazdin (1982) suggested an advantage of kappa is that it offers an estimate of agreement between two observers corrected for chance. Cohen's Kappa is calculated as follows:

$$\kappa = \frac{P_{obs} - P_{\exp}}{1 - P_{\exp}}$$

where $P_{obs}$ is the relative observed agreement among the rater, and is computed by summing up the tallies representing agreement and dividing by the total number of tallies. $P_{exp}$ is the probability of each observer randomly choosing each category; it is computed by multiplying the first column by the first row total, adding this to the second column total multiplied by the second row total, and then dividing the resulting sum of

the column-row products by the total number of tallies squared.

That is, $P_{obs} = \frac{\sum_{i=1}^{n} x_{ii}}{N}$ and $P_{\exp} = \frac{\sum_{i=1}^{n} x_{+i} x_{i+}}{N^2}$

*Similarity among the raters: Index of dependability*. In the case of comparing two raters, the data of the first observer were used as the standard for each dyad. However, when the comparison involved six raters, because there was no “true” value as a standard, it was hard to decide whether any discrepancy among raters was significantly different. Also, Peterson et al. (1982) claimed that an agreement among observers did not automatically ensure that the observation results were reliable. Some of observers could be sharing a bias or limitation which could systematically influence their measurement, while other observers might be measuring the behaviors they were supposed to measure. Therefore, it is necessary to use the concept of dependability in Generalizability theory (G-theory).

G-theory is a statistical theory that can be used to evaluate the dependability of behavioral measurements (Shavelson & Webb, 1991). Dependability refers to the accuracy of generalizing from an observed score to the average score that can be rated under all possible conditions. G-theory provides a coefficient that reflects the level of dependability; the coefficient is similar to a reliability coefficient.

To examine whether six raters consistently measure the child’s behavior, a one-facet crossed design ($i \times R$) in D-studies (Decision studies) was used. In this study, raters were treated as the main source of the error which was the cause of any difference among the

rating results and the child's replacement behavior in intervals of the whole interval assessment form was treated as the object of measurement (see Figure 2).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | + | + | - | + | + | - | - | - | - | + | - | - | + | + | - |
| S2 | + | + | - | + | + | - | - | - | - | + | - | - | + | + | - |
| S3 | + | + | - | + | + | - | - | - | - | + | - | - | + | + | - |
| S4 | + | + | + | + | + | - | - | + | - | + | - | + | + | + | - |
| S5 | + | + | - | + | + | - | - | - | - | - | - | + | + | + | - |
| S6 | + | + | - | + | + | - | - | - | - | - | - | - | + | + | - |

*Figure 2*. An example of rating results (VC 1) on the child's behavior of six raters.

The main source of error (i.e., rater; $\sigma_r^2$) and all interactions ($\sigma_{ir\cdot e}^2$) between other sources and rater indicate the variability due to differences in agreement among raters, i.e., the absolute error variance. Overall, the intervals can be estimated when comparing data from six raters via interval-by-interval agreement. That is, through this analysis, the correlation of rating across intervals and raters can be estimated (D. Sabers, personal communication, January 24, 2010). For this purpose, two types of values were calculated. One was estimated absolute error variance ($\sigma_{Abs}^2$); the other one was the index of dependability (coefficient; Φ). Estimated absolute error variance and index of dependability were calculated video clip by clip.

An estimated absolute error variance was calculated as follows:

$$\sigma_{Abs}^2 = \frac{\sigma_r^2}{n_r} + \frac{\sigma_{ir\cdot e}^2}{n_r}$$

Index of dependability was calculated as follows:

$$\Phi = \frac{\sigma_i^2}{(\sigma_i^2 + \sigma_{Abs}^2)}$$ where $\sigma_i^2$ = variance of intervals (i.e., systematic variance)

There is no rule of thumb for index of dependability to decide acceptable level. However, following the rules of thumb of George and Mallery (2003), 0.7 phi value was set as the guideline for acceptable level.

*Treatment Integrity*

*Degree of treatment integrity.* The level of TI was presented as the percentage of intervals with correct implementation (WI), the percentage of steps implemented (Y/N), and the percentage of potential points earned (LIK).

- The formula for percentage of intervals (WI) was:

$$\frac{\text{N. of intervals with correct implementation}}{\text{24 (N. of total intervals)}} \times 100$$

- The formula for percentage of items implemented (Y/N) was:

$$\frac{\text{N. of steps correctly implemented}}{\text{(N. of total steps – N. of N/A)}} \times 100$$

- The formula for percentage of potential points (LIK) was:

$$\frac{\text{Sum of points}}{\text{Potential points}} \times 100$$

where potential points = 3×(N. of total steps – N. of N/A)

*Agreement of two raters within a dyad.* First, Pearson correlation was calculated by using percentages of TI from two raters in a dyad. Then, the percentage of TI IOA and Cohen's *k* was used to examine agreement between two raters. There is no established guideline for TI IOA. By considering the recommendation of Horner et al. (2005), a slightly higher standard of 85% IOA and a 0.65 kappa score was set as the standard in

this study.

- Percentage of TI IOA in a dyad: In this study, S1, S3, and S5 were the primary raters in each dyad. The formula for percentage of TI IOA was:

$$\% \text{ of IOA} = \frac{N. \text{ of agreements}}{N. \text{ of agreements} + N. \text{ of disagreements}} \times 100$$

  where agreements were:

  - WI dyad: interval-by-interval agreement
  - Y/N dyad: the same response (i.e., yes-yes, no-no, or n/a-n/a) between two raters on a step.
  - LIK dyad: either the same response or the adjacent response such as 0 and 1 or 2 and 3 between two raters in a step. Because one (i.e., seldom) and 2 (i.e., often) or any number and n/a are not similar concepts, those matches were treated as different responses.

- Cohen's *k*: In the case of WI dyad, "+/-" were the two codes for the coding scheme, whereas "yes/no/n·a" were the three codes for Y/N dyad and "0/1/2/3/n·a" were the five codes for LIK dyad. After completing an agreement matrix, each Cohen's *k* was calculated using the formula given previously.
- Index of dependability: Because there were disagreements in several cases within each dyad, indices of dependability were calculated for each dyad. To decide whether two raters in a dyad consistently assessed TI in a video clip and across video clips, a two-facet crossed design ($v \times i / s \times R$) in D-studies was used. Raters were treated as the main source of the error; the video clips and each TI measurement method were set as the object of measurement which

related to systematic variance. A phi value of 0.70 was set as the guideline for acceptance.

*Percentage of total variance.* The result of the D-study in each dyad showed that raters fairly consistently observed and rated the instructor's intervention implementation behavior. It indicated that the main source of inconsistency was systematic variance. Therefore, a G-study was used to check the portion of the all possible sources of error which were related to the main source of error (i.e., rater) on total variance. For this, a two-facet crossed design ($v \times i \times R$ or $v \times s \times R$) in each dyad was used to compare the percentage of total variance on each variance.

*Differences in TI measurement methods outcomes.* Because the contributed portion of both rater and TI measurement method to total variance for each dyad was similar, comparing TI measurement methods outcomes was possible. As the first step, descriptive statistics were used to compare the three different TI measurement methods. Next, by using ANOVA mixed design (i.e., two-factor split-plot design) the outcomes from three TI measurement methods were compared, because the data were derived from repeated measurement. The two factors were the TI measurement method and the video clip. The main purpose of the study was to compare TI measurement outcomes from three different methods. Therefore, differences created by interaction or video clip were not analyzed. Percentages of TI were used as data. Because there was difference among the three methods, a post hoc test was conducted.

*Relationship between Degree of the Child's Behavior and Treatment Integrity*

*Pearson product-moment correlation coefficient.* Peterson and colleagues (1982) argued that if both independent and dependent variable were observed and recorded, the value of the observed independent variable could be compared to the value of dependent variable. If the meaning of "being observed" is direct measurement, it is impossible to compare the degree of TI to the degree of child's behavior when using indirect measures which are done after an observation. However, in the literature review, there was no evidence that researchers have examined the relationship between the independent variable and the dependent variable by using direct measure outcomes only. Therefore, in a conventional way, a relationship between the degree of child's behavior and the degree of TI was examined by using scatter plot first. Then, the value of Pearson moment-product correlation coefficient of each dyad was calculated.

*Significant difference between two correlation coefficients*. Because two dyads showed some degree of relationship between the child's behavior and TI, a significant difference between two correlation coefficients was tested by using the formula suggested by Glass and Stanley (1970). The null hypothesis and the alternative hypothesis are:

$$H_0 : \rho_{xy} = \rho_{xz}$$

$$H_1 : \rho_{xy} \neq \rho_{xz}$$

The statistic for testing $H_0$ against $H_1$ is:

Where *n* is the sample size,

$r_{xy}$ is the correlation of percentage of intervals of child's behavior and percentage of TI for the WI dyad,

$r_{xz}$ is the correlation of percentage of child's behavior and percentage of TI of the Y/N dyad, and

$r_{yz}$ is the correlation of percentage of the WI and of the Y/N,

$$z = \frac{\sqrt{n}(r_{xy} - r_{xz})}{\sqrt{(1-r_{xy}^2)^2 + (1-r_{xz}^2)^2 - 2r_{yz}^3 - (2r_{yz} - r_{xy}r_{xz})(1-r_{xy}^2 - r_{xz}^2 - r_{yz}^2)}}$$

The two critical values for the hypothesis test are ${}_{\alpha/2}z$ and ${}_{1-(\alpha/2)}z$

All data analysis was conducted using *PASW Statistics 18* software program.

# CHAPTER 4

# RESULTS

## Training for Observation

Rater training sessions were given to each dyad until both raters in a dyad reached an average IOA level of 85% or 0.65 for Cohen's *k* for three consecutive observations for both the child's behavior and the TI between the two raters and between the each rater and the PI (Table 12).

Table 12

*Training Results*

| Dyad | | The child's behavior IOA (range) | Treatment Integrity IOA | |
|---|---|---|---|---|
| | | | % (range) | Cohen's *k* (range) |
| Dyad 1(WI[a]) | S1-S2 | 90.6 (79.2-100) | 85.4 (70.8-100) | 0.67 (0.43-1.00) |
| | PI-S1 | 94.0 (79.2-100) | 88.0 (75.0-100) | 0.70 (0.43-1.00) |
| | PI-S2 | 94.9 (87.5-100) | 90.7 (83.3-100) | 0.78 (0.52-1.00) |
| Dyad 2 (Y/N[b]) | S3-S4 | 93.2 (75.0-100) | 90.9 (33.3-100) | 0.85 (-0.15-1.00) |
| | PI-S3 | 93.9 (79.2-100) | 87.9 (44.4-100) | 0.80 (0.16-1.00) |
| | PI-S4 | 91.3 (66.7-100) | 87.9 (44.4-100) | 0.79 (0.38-1.00) |
| Dyad 3 (LIK[c]) | S5-S6 | 92.5 (75.0-100) | 85.5 (63.6-100) | 0.62 (0.26-1.00) |
| | PI-S5 | 91.7 (79.2-100) | 88.3 (72.7-100) | 0.62 (0.39-1.00) |
| | PI-S6 | 95.0 (87.5-100) | 93.7 (81.8-100) | 0.75 (0.54-1.00) |

[a]WI: Whole interval. [b]Y/N: Yes/No component checklist. [c]LIK: Likert-type rating scale.

The number of training sessions varied depending on each dyad's progress, but all three dyads viewed all nine training video clips in order to reach the IOA criteria. The WI dyad reached the guideline for both the percentage of IOA and the kappa after three training sessions. The Y/N dyad reached the required percentage of IOA for rating the child's behavior after the first training session, but did not reach the guideline for TI until the third training session. The two raters attained over 85% IOA for both the child's behavior and TI with the PI after the third training session.

In the LIK dyad, after the second training session, rater S5 stopped participating in this study. With a new rater, the dyad needed two more training sessions. The dyad reached the percentage of IOA guideline for both the child's behavior and TI, but did not attain the 0.65 kappa guideline. Because of time limitations, the PI adjusted the kappa guideline from 0.65 to 0.60, and the training sessions concluded with a 0.62 kappa which exceeded the recommendation by Horner et al. (2005).

*Procedural fidelity (PF) IOA*

PF data were collected by the PI during all training and rating sessions. The average percentage of procedural fidelity during training sessions for the WI dyad, the Y/N dyad, and the LIK dyad were all 100%. The percentages of PF of the rating sessions were also 100% for all the three dyads.

For PF IOA, the alternate and an outer observer collected data from different dyads on different days during training and rating sessions. TI IOAs were collected in 75 % of the training sessions for the WI dyad and 100% of training sessions for the Y/N and LIK

dyads, whereas TI IOAs were collected for100% of the rating sessions for the WI and Y/N dyads, and 50% of rating sessions for the LIK dyad. The PF IOAs for all training and rating sessions that were assessed were at 100%.

## Similarity of Treatment Integrity Data among TI measurement methods

To examine whether the different TI measurement methods produced similar TI data, two questions were posed in a hierarchical order. Question 1, "*Do raters assess the child's behavior similarly?*" had to be addressed first because implementation of the behavioral intervention was directly related to the child's behavior. That is, the premise that the raters observed the same behavior of the child had to be demonstrated before comparing their assessments of TI. Then, Question 2, "*Do the three different TI assessment methods produce similar TI outcomes?*" was addressed.

### *Question 1: Do Raters Assess the Child's Behavior Similarly?*

All three dyads used the whole-interval observation method to assess the child's behavior. To answer whether the raters assessed the child's behavior similarly, correlations between the two raters within each dyad as well as among the six raters were examined.

#### *Agreement between Two Raters in a Dyad*

To examine the similarity between each dyad of raters, Pearson's product-moment correlation coefficient was calculated. The examination revealed that the Pearson's

correlation coefficient of each dyad was quite high. Specifically, the correlation coefficient for the WI dyad was .941**, for the Y/N dyad it was .826**, and for the LIK dyad it was .955** (** $p < .001$, 2-tailed).

Next, the average percentage of IOA for the child's replacement behavior between the two raters and the value of Cohen's *k* were calculated and compared (Table 13).

Table 13

*Agreement between Two Raters within a Dyad on the Child's Behavior*

| Agreement / Video clip | % of IOA | | | Cohen's *k* | | |
|---|---|---|---|---|---|---|
| | WI | Y/N | LIK | WI | Y/N | LIK |
| VC 1 | 91.7 | 83.3 | 91.7 | .83 | .65 | .83 |
| VC 2 | 91.7 | 91.7 | 95.8 | .83 | .83 | .92 |
| VC 3 | 100.0 | 87.5 | 95.8 | 1.00 | .71 | .92 |
| VC 4 | 91.7 | 91.7 | 100.0 | .82 | .78 | 1.00 |
| VC 5 | 91.7 | 91.7 | 91.7 | .83 | .83 | .83 |
| VC 6 | 91.7 | 95.8 | 95.8 | .83 | .92 | .92 |
| VC 7 | 79.2 | 83.3 | 87.5 | .57 | .00 | .74 |
| VC 8 | 91.7 | 83.3 | 91.7 | .78 | .43 | .83 |
| VC 9 | 75.0 | 87.5 | 79.2 | .25 | .73 | .52 |
| VC 10 | 91.7 | 87.5 | 91.7 | .83 | .83 | .83 |
| Average | 89.6 | 88.3 | 92.1 | .76 | .67 | .83 |

Overall, the results showed that the averages of agreements between the raters in each dyad on the child's behavior were in acceptable ranges (i.e., over 85% IOA and over .65 kappa). Nevertheless, when specifically looking at the data, some points of concern

became evident. For example, in the case of the WI dyad, in two of the video clips (i.e., VC 7 and VC 9), the degree of agreement was quite low (i.e., 79.2%, 75.0%, respectively). Also, kappas were very low for these two video clips when compared to the other results. In the case of VC 7, kappa was 0.56; for VC 9, it was 0.25. However, the percentages of IOA for 7 out of 8 video clips were all 91.7%; the last one was 100%, and the mode of kappas was 0.83 ($n$ = 5 video clips).

Even though the Y/N dyad attained a relatively low average of IOA and of kappa, when compared to the results of the other two dyads, the variability of IOA scores was not as large. Agreement was 83.3% for three video clips (VCs 1, 7, and 8), was. 87.5% for three other video clips (VCs 3, 9, and 10), was 91.7% for three clips (VCs 2, 4, and 5), and was 95.8% for VC 6. Nevertheless, the range of kappas raises the question of whether some agreements between the two raters were made by chance. The LIK dyad showed some differences for VC 9. Except the VC9, the two raters kept high degree of IOAs and kappas.

*Similarity among the Raters*

To compare the ratings among the six raters, a box plot (Figure 3) was used. The means of percentage of intervals, indicated as "▬" on the box of the dyads, are closely located for several video clips (e.g., VCs 2, 5, and 10), but farther apart for others. Ranges of difference in the percentage of intervals between the two raters (indicated as the length of a box) also differ.

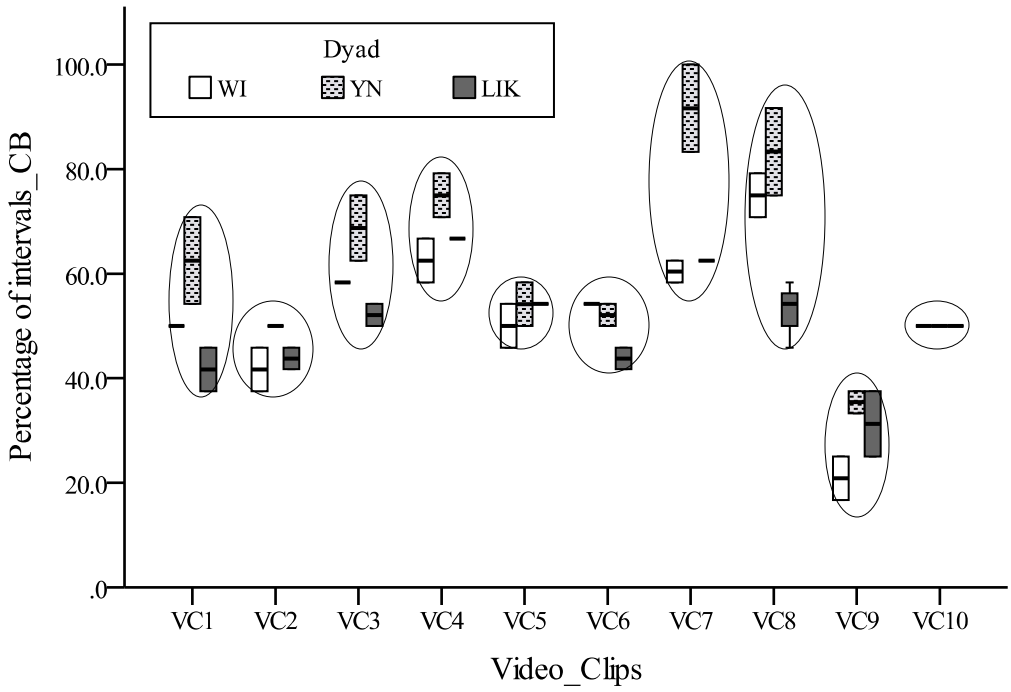


*Figure 3.* Means and ranges of percentages of intervals of the child's behavior among three dyads

As mentioned in Chapter 3, it was hard to determine whether the raters reliably and similarly assessed the child's behavior across the video clips or whether any significant difference existed among the raters. Moreover, on further observation of the box plot, it is hard to discern whether the percentage of differences (i.e., error variance) among the raters was mainly caused by raters themselves or by using different measurement methods. To examine the similarity among the six raters, G-theory was used. The raters were treated as the main source of error, and intervals (i.e., interval data recording system) were the object of measurement. Therefore, the sources of error were the rater facet and the interactions between the rater facet and intervals (see Table 14).

Table 14

*Indices of Dependability for the Child's Behavior*

| Video Clip | Component Estimates | | | $\hat{\sigma}^2(\Delta)$ | Φ |
|---|---|---|---|---|---|
| | $\sigma_i^2$ | $\sigma_r^2$ | $\sigma_{ri\cdot e}^2$ | | |
| VC 1 | .170 | .009 | .080 | .0148 | .920 |
| VC 2 | .208 | .000 | .048 | .0080 | .963 |
| VC 3 | .131 | .019 | .106 | .0208 | .863 |
| VC 4 | .187 | .003 | .036 | .0065 | .966 |
| VC 5 | .194 | -.001[b] | .065 | .0107 | .948 |
| VC 6 | .224 | .001 | .035 | .0060 | .974 |
| VC 7 | .093 | .025 | .097 | .0203 | .821 |
| VC 8 | .096 | .024 | .101 | .0223 | .811 |
| VC 9 | .110 | .003 | .100 | .0172 | .865 |
| VC 10 | .188 | -.003[a] | .072 | .0115 | .942 |

[a] Negative variance component estimates may occur. There are two approaches to treating small negative estimates. In this study, follow Cronbach et al. (1972, cited in Shavelson & Webb, 1991) and set the negative estimates to zero.

An absolute error variance, $\hat{\sigma}^2(\Delta)$, indicating error from raters ($\sigma_r^2$) and interaction ($\sigma_{ri\cdot e}^2$), was used to calculate the index of dependability, denoted as Φ (phi). A high coefficient indicates that the ratings were dependable or reliable (D. Sabers, personal communication, January 24, 2010).

The phi values for VC 7 and VC 8 were relatively low when compared to the other phi values, and the result matches to the graph in Figure 3. However, each phi value was within an acceptable range. Based on phi values, the question of whether raters similarly assessed the child's behavior could be answered – the raters consistently rated the child's behavior for each video clip across intervals.

*Question 2: Do the Three Different TI Assessment Methods Produce Similar TI Outcomes?*

Because it was determined that the six raters consistently rated the child's behavior, TI data were examined to answer whether the three different TI measurement methods resulted in similar outcomes.

*Agreement between Two Raters in a Dyad*

To examine the similarity between two raters within a dyad, each percentage of TI between the raters was compared in a line graph (see Figures 4, 5, and 6) and Pearson's product-moment correlation coefficient was calculated. Upon inspection of Figure 4, it is hard to see any distinct difference between two raters except VC 7. Also, Pearson's correlation coefficient was high ($r$ = .907**, ** p < .000, 2- tailed).

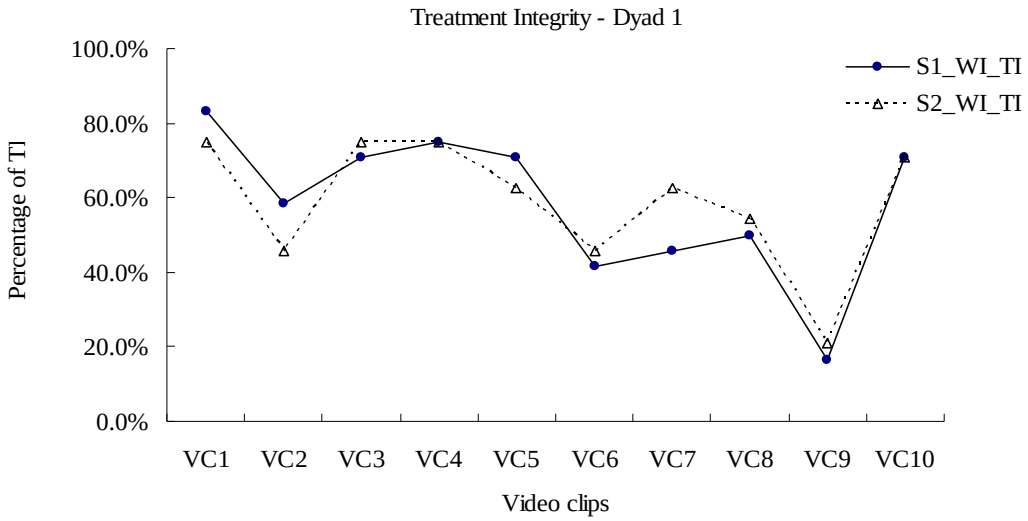


*Figure 4.* Percentage of TI on the WI dyad

In the case of the Y/N and LIK dyads (Figures 5 and 6), score gaps between two raters are apparent for several video clips. The Pearson's correlation coefficient was low and not significant (Y/N dyad, $r$ = .480; LIK dyad, $r$ = .471).

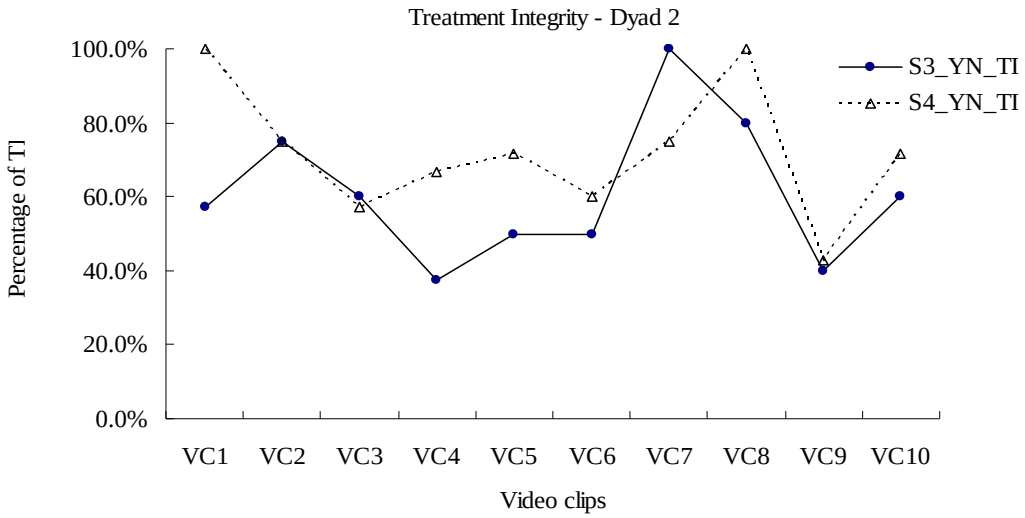


*Figure 5*. Percentage of TI on the Y/N dyad

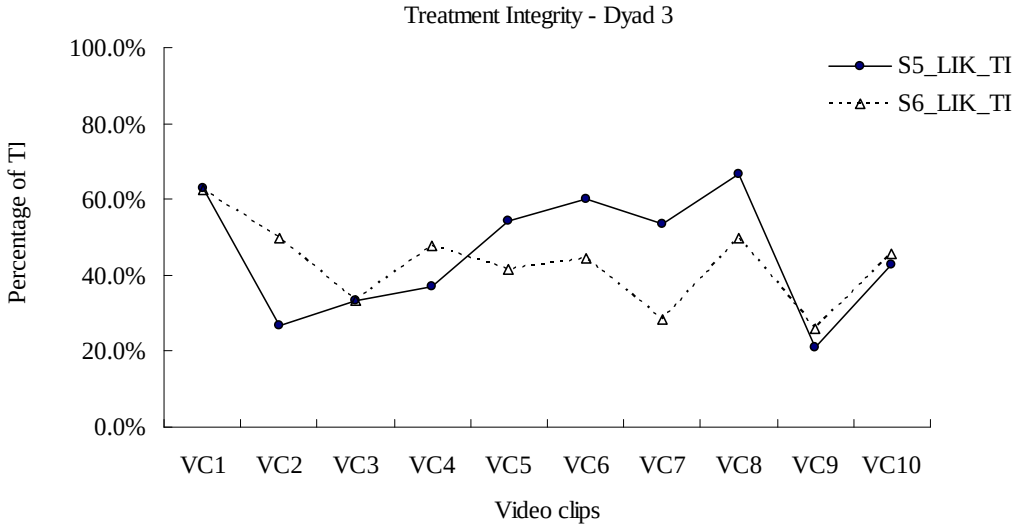


*Figure 6*. Percentage of TI on the LIK dyad

Table 15 shows the percentage of TI IOA and values of Cohen's *k* for each dyad. There is no set guideline for degree of TI IOA in the literature, but considering the guideline of IOA for the dependent variable, an 80% IOA or 0.6 kappa score was considered acceptable (see Horner et al., 2005). Additionally, Fleiss (1988, cited from Bakeman, & Gottman, 1997) characterized kappas of .40 to .60 as fair while Bakeman and Gottman regarded kappas less than .70 with some concern.

Table 15

*Agreement between Two Raters within a Dyad on TI*

| Agreement / Video clip | % of IOA | | | Cohen's *k* | | |
|---|---|---|---|---|---|---|
| | WI | Y/N | LIK | WI | Y/N | LIK |
| VC 1 | 91.7 | 72.7 | 72.7 | .75 | .57 | .33 |
| VC 2 | 87.5 | 100.0 | 81.8 | .75 | 1.00 | .62 |
| VC 3 | 95.8 | 72.7 | 63.6 | .89 | .58 | .05 |
| VC 4 | 100.0 | 72.7 | 81.8 | 1.00 | .61 | .35 |
| VC 5 | 83.3 | 81.8 | 100.0 | .63 | .73 | .61 |
| VC 6 | 87.5 | 90.9 | 90.9 | .67 | .84 | .22 |
| VC 7 | 75.0 | 81.8 | 72.7 | .51 | .58 | .52 |
| VC 8 | 87.5 | 81.8 | 63.6 | .75 | .65 | .10 |
| VC 9 | 79.2 | 81.8 | 90.9 | .32 | .72 | .80 |
| VC 10 | 91.7 | 72.7 | 90.9 | .80 | .58 | .43 |
| Average | 87.9 | 80.9 | 80.9 | .71 | .69 | .40 |

Even though the two raters in each dyad might produce similar overall TI scores, they might disagree on the scoring of individual intervals or components/steps, which would reduce their level of IOA. In the case of WI dyad, the scores from the two raters were similar and Pearson's *r* was high, yet there were some disagreements between the two raters when scores were examined interval-by-interval for several video clips. Specifically, in two video clips (VC 7 and VC 9), there were disagreements in at least 5 of the intervals. Y/N dyad showed no correlation between two raters' scores. Also, the two raters reached over 85% for only 2 video clips. They showed disagreements for either 2 or 3 components in the rest of video clips. The LIK dyad showed high agreements in rating the child's behavior. That is, even though the raters reached high agreements (i.e., over 90%) for 4 video clips, for the other 4 video clips, they varied noticeably in their evaluation of each component.

Overall, the three dyads reached Horner, et al.'s 80% guideline for IOA. However, only the WI dyad reached the 85% IOA guideline which was established for the training sessions, and the two TI indirect measurement methods merely reached 80% IOA. Because there were 11 steps for indirect measurement of TI, disagreement on one step created a 9.09% gap in the percentage of TI. Using the WI method, a disagreement on one interval created only a 4.2% difference. In the case of kappas, the WI dyad and the Y/N dyad reached Cohen's *k* guideline for the training sessions (i.e., $k = 0.65$), but the LIK dyad did not. Not only was the average of kappas low, but also almost none of the values of kappa ($n = 7$) reached 0.60.

In summary, when examining the overall average TI IOA and kappa scores for TI for all three dyads, the percentages of TI IOA were acceptable for all three dyads, and the kappas were acceptable for the WI and Y/N dyads. However, there were several cases in which the two raters showed a great degree of disagreement. As a result, it is difficult to assess whether the two raters in each dyad similarly observed and rated TI in a single video clip as well as across all ten video clips. Therefore, to examine whether the two raters consistently measured the same behaviors across all video clips, a one-facet D-study for each video clip and a two-facet D-study across 10 video clips was used.

*Index of dependability.* The rater facet was considered as a main source of variation, and the TI value for each measurement system (i.e., an interval for the WI, and a step for the Y/N and the LIK) and for a video clip was the object of measurement. The number of raters was set at $n = 2$. In the case of the WI dyad (Table 16), the two raters consistently assessed TI across video tapes ($\Phi = 0.825$) as well as in each video clip except two (i.e., VCs 7 and 9) when set 0.70 for the acceptance level. Moreover, these TI outcomes directly matched the percentage of intervals of both of the child's behavior and TI as well as kappas for TI.

If the raters observed the child's behavior differently, it should have been reflected by the measurement of TI for the instructor's intervention. In other words, if the two raters showed a discrepancy for TI in particular video clips (i.e., VCs 7 and 9) the two raters should have shown a large disagreement for TI in VCs 7 and 9. The disagreement matched phi values also. Phi values for VCs 7 and 9 were below 0.70, whereas the other phi values were more than 0.70.

Table 16

*Estimated Error Variances and indices of Dependability for WI dyad on TI*

| Video Clip | Component Estimates | | | $\hat{\sigma}^2(\Delta)$ | Φ |
|---|---|---|---|---|---|
| | $\sigma_i^2$ | $\sigma_r^2$ | $\sigma_{ri\cdot e}^2$ | | |
| VC 1 | .130 | .001 | .020 | .021 | .861 |
| VC 2 | .199 | .0025 | .0285 | .031 | .865 |
| VC 3 | .185 | .000 | .011 | .011 | .944 |
| VC 4 | .196 | .000 | .000 | .000 | 1.000 |
| VC 5 | .147 | .000 | .042 | .042 | .778 |
| VC 6 | .192 | .000 | .032 | .032 | .857 |
| VC 7 | .136 | .0045 | .058 | .0625 | .683 |
| VC 8 | .196 | .000 | .032 | .032 | .860 |
| VC 9 | .051 | .000 | .054 | .054 | .486 |
| VC 10 | .172 | .000 | .022 | .022 | .887 |
| Across 10 VCs *(v×i×R)* [a] | | | | .0305 | .842 |

[a] For more details, see Table 19.

The Y/N dyad reached more than 80% IOA for the child's behavior for all video clips and showed disagreements for VCs 7 and 8 in terms of kappa scores. However, the two raters showed unacceptable percentages of TI IOA for VCs 1, 3, 4, and 10, and unacceptable kappas of TI IOA for VCs 1, 3, 7, and 10. Moreover the phi values revealed that agreements between the two raters were under 0.70 in VCs 3, 4, 7, and 10 (see Table 17). Therefore, there are unmatched results between IOA for the child's behavior and TI IOA, and among the percentage of IOA, kappas, and phi values. Especially, VC 10 was just one of several video clips for which the two raters showed a discrepancy when comparing the percentages of IOA or kappas, but the phi vale was the lowest among other

phi values.

Even though Pearson's correlation coefficient ($r$ = .480) indicated that there was no significant correlation between the two raters' observation values on TI, the overall phi value ($\Phi$ = .797) indicated that the two raters relatively consistently rated the instructor's TI across 10 video clips.

Table 17

*Estimated Error Variances and Indices of Dependability for Y/N dyad on TI*

| Video Clip | Component Estimates | | | $\hat{\sigma}^2(\Delta)$ | $\Phi$ |
|---|---|---|---|---|---|
| | $\sigma_s^2$ | $\sigma_r^2$ | $\sigma_{rs\cdot e}^2$ | | |
| VC 1 | .945 | .032 | .218 | .250 | .791 |
| VC 2 | 1.364 | .000 | .000 | .000 | 1.000 |
| VC 3 | .845 | .0135 | .418 | .4315 | .662 |
| VC 4 | .927 | .068 | .3635 | .4315 | .682 |
| VC 5 | 1.091 | .0135 | .2135 | .227 | .828 |
| VC 6 | 1.509 | .000 | .091 | .091 | .943 |
| VC 7 | .473 | .0275 | .268 | .2955 | .615 |
| VC 8 | 1.000 | .009 | .1045 | .1135 | .898 |
| VC 9 | 1.373 | .0275 | .268 | .2955 | .823 |
| VC 10 | .309 | .0955 | .4045 | .5000 | .382 |
| Across 10 VCs *(v×s×R)* [a] | | | | .269 | .797 |

[a] For more details, see Table 20.

In the case of the LIK dyad (Table 18), the phi value was under .70 for VCs 3, 4, and 7, the two raters could not reach 80% IOA on VCs 1, 3, 7, and 8, and kappa was less than 0.60 for VCs 1, 3, 4, 6, 7, 8, and 10. Surprisingly, this LIK dyad reached high

indices of dependability on most of video clips and the raters consistently measured TI across 10 video clips (Φ = .849), even though the kappas were poor.

Table 18

*Estimated Error Variances and Indices of Dependability for LIK dyad on TI*

| Video Clip | Component Estimates | | | $\hat{\sigma}^2(\Delta)$ | Φ |
|---|---|---|---|---|---|
| | $\sigma_s^2$ | $\sigma_r^2$ | $\sigma_{rs\cdot e}^2$ | | |
| VC 1 | 1.936 | .000 | .3045 | .3045 | .864 |
| VC 2 | 4.691 | .000 | .200 | .200 | .959 |
| VC 3 | 1.064 | .0135 | 1.3045 | 1.318 | .447 |
| VC 4 | 2.045 | .1275 | .8725 | 1.000 | .672 |
| VC 5 | 3.536 | .0135 | .0545 | .068 | .971 |
| VC 6 | 3.800 | .0635 | .118 | .1815 | .954 |
| VC 7 | 2.618 | .2045 | 1.0225 | 1.227 | .681 |
| VC 8 | 2.109 | .1045 | .691 | .7955 | .726 |
| VC 9 | 4.209 | .000 | .2045 | .2045 | .954 |
| VC 10 | 3.645 | .0045 | .1545 | .159 | .958 |
| Across 10 VCs *(v×s×R)* [a] | | | | .550 | .849 |

[a] For more details, see Table 21.

The results of a one-facet and a two-facet crossed D-study showed that the main source of inconsistency was not caused by the raters. Therefore, the causes may be the systematic variance which is from video clips and the TI measurement method. To answer this inquiry, a two-facet crossed design in G-study was used.

*Percentage of total variance*. The three main sources of variances were rater, TI measurement method, and video clip. The rater facet was set as the source of error

variance whereas video clip and TI measurement method were the sources of systematic variance. The number of raters was set at 2 ($n_r$ = 2).

In the case of WI dyad (see Table 19) the percentage of variance related to raters which accounts for the total variance was 14.09%. In comparison, the interval (i.e., TI measurement method), video clips, and interaction between interval and video clip which formed systematic variance accounted for a much larger portion of the total variance (85.91%). Specifically, the interaction between interval and video clip made up the largest percentage of the total variance (68.36%).

Table 19

*G- study (v×i×R) on WI Dyad for TI*

| Source of variation | Estimated Variance component | random effects variance components ($n_r$ = 2) | Percentage of total variance | |
|---|---|---|---|---|
| | | | error | systematic variance |
| Rater | .000 | .000 | .00 | |
| Interval | .012 | .012 | | 5.54 |
| Video clip | .026 | .026 | | 12.01 |
| Rater × Interval | .002 | .001 | 0.46 | |
| Rater × Video clip | .001 | .0005 | 0.23 | |
| Interval × Video clip | .148 | .148 | | 68.36 |
| Rater × Interval × Video clip | .058 | .029 | 13.40 | |
| | $\hat{\sigma}^2(\Delta)$ | .0305 | | |
| | $\Phi$ | .842 | | |

In the case of the Y/N dyad (see Table 20), the percentage of variance for the raters was 20.35% whereas the percentage of systematic variance was 79.65%. This percentage of variance for raters is slightly higher than the percentage for the WI dyad (i.e., 14.09%). However, among the sources of variance, the percentage of total variance which was directly related to the Y/N checklist (i.e., "step") was 42.89%. Moreover, when adding the portion of interaction between step and video clip, the amount is almost the same as the percentage of the systematic variance. Therefore, even though the percentage of variance of raters accounts for 20% of the total variance, the main cause of disagreement between the two raters in the Y/N dyad was not the raters but systematic variances such as the TI measurement method itself.

Table 20

*G- study (v×s×R) on Y/N Dyad for TI*

| Source of variation | Estimated Variance component | random effects variance components ($n_r = 2$) | Percentage of total variance | |
|---|---|---|---|---|
| | | | error | systematic variance |
| Rater | -.009[a] | .000 | .00 | |
| (Intervention) Step | .567 | .567 | | 42.89 |
| Video clip | .070 | .070 | | 5.29 |
| Rater × Step | .023 | .012 | 0.91 | |
| Rater × Video clip | .066 | .033 | 2.50 | |
| Step × Video clip | .416 | .416 | | 31.47 |
| Rater × Step × Video clip | .447 | .224 | 16.94 | |
| $\hat{\sigma}^2(\Delta)$ | | .269 | | |
| $\Phi$ | | .797 | | |

[a] In this study, the negative estimates is set to zero.

The LIK dyad showed consistency in rating the child's behavior during both the training and the rating sessions. However, when TI was assessed, the two raters were unable to reach the established level of agreement. The main cause of the inconsistency was the systematic variance rather than raters (see Table 21). The percentage of variance from the two LIK raters was 15.07%. Similar to the Y/N dyad, TI measurement method itself accounts for 39.34% of the total variance with 41.94% of the total variance derived from the interaction between two sources, TI measurement method and video clip.

Table 21

*G- study (v×s×R) on LIK Dyad for TI*

| Source of variation | Estimated Variance component | random effects variance components ($n_r = 2$) | Percentage of total variance | |
|---|---|---|---|---|
| | | | error | systematic variance |
| Rater | -.013[a] | .000 | .00 | |
| (Intervention) Step | 1.435 | 1.435 | | 39.34 |
| Video clip | .133 | .133 | | 3.65 |
| Rater × Step | .019 | .0095 | 0.26 | |
| Rater × Video clip | .113 | .0565 | 1.54 | |
| Step × Video clip | 1.530 | 1.530 | | 41.94 |
| Rater × Step × Video clip | .967 | .484 | 13.27 | |
| $\hat{\sigma}^2(\Delta)$ | | .550 | | |
| $\Phi$ | | .849 | | |

[a] Negative variance component estimates may occur. It is treated as the value of zero.

In summary, the results from both the G-study and the D-study indicated that the raters consistently measure TI, and the percentage of variance from sources related to the

rater accounted for 14.09% (WI dyad), 20.35% (Y/N dyad), and 15.07% (LIK dyad), respectively. For the WI dyad, the video clip variable was the main source of variance. For the two indirect TI measurement methods (i.e., Y/N and LIK), TI measurement method was the main source of variance. However, the portion of error variance and that of systematic variance were similar among three dyads.

*Similarity of TI measurement methods outcomes*

To examine whether the means of percentages of TI from the three methods were similar, these means of percentages of TI were compared first by visual inspection.

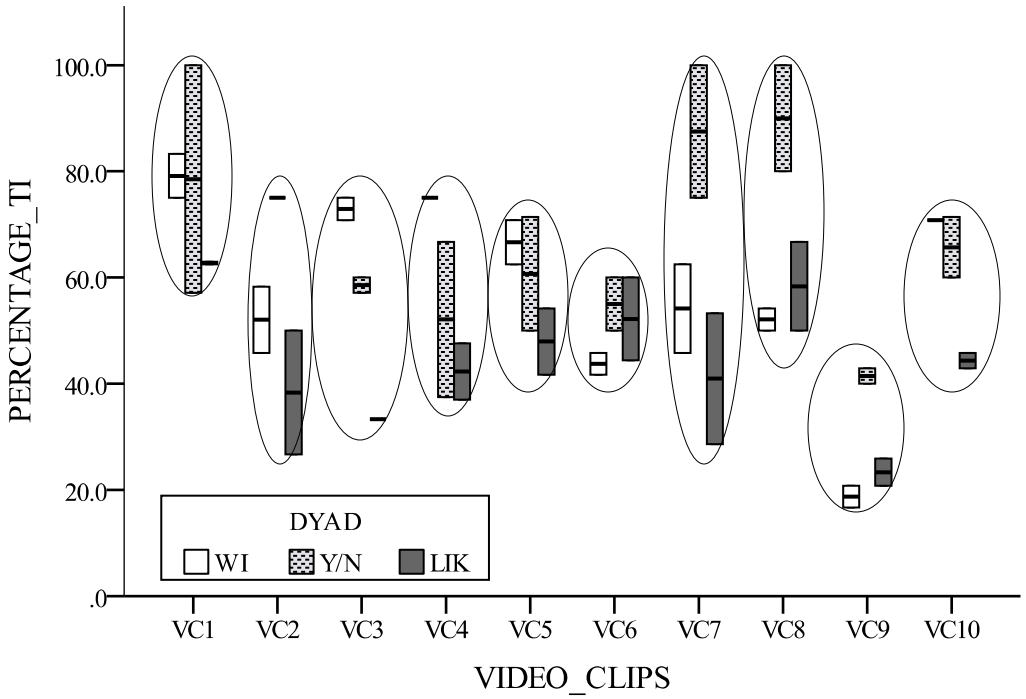


*Figure 7*. Means and ranges of percentages of TI among three dyads

Figure 7 indicates that there is a discrepancy in the TI measurement outcomes among the methods. Specifically, both the WI dyad and the Y/N dyad rated the degrees of TI high, whereas LIK dyad obtained the lowest scores among the three dyads. This is more obvious in the average of the TI percentages among the three dyads (Table 22). A comparison of the average means revealed that the WI dyad and the Y/N dyad shared the number of cases of greatest mean scores (i.e., 5 out of 10 video clips), whereas the LIK dyad had the least. However, when comparing standard deviation of each dyad across 10 video clips, fluctuation among the values of *SD* for the two indirect TI measurement methods was severe. That of the Y/N was the greatest.

Table 22

*Mean and Standard Deviation of Each Dyad Reporting TI*

| Video Clip | Whole Interval | | [Yes/No] | | Likert-type | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| VC 1 | 79.150 | 5.869 | 78.550 | 30.335 | 62.750 | .354 |
| VC 2 | 52.050 | 8.838 | 75.000 | .000 | 38.350 | 16.476 |
| VC 3 | 72.900 | 2.970 | 58.550 | 2.051 | 33.300 | .000 |
| VC 4 | 75.000 | .000 | 52.100 | 20.648 | 42.300 | 7.495 |
| VC 5 | 66.650 | 5.869 | 60.700 | 15.132 | 47.950 | 8.839 |
| VC 6 | 43.750 | 2.899 | 55.000 | 7.071 | 52.200 | 11.031 |
| VC 7 | 54.150 | 11.809 | 87.500 | 17.678 | 40.950 | 17.466 |
| VC 8 | 52.100 | 2.970 | 90.000 | 14.142 | 58.350 | 11.809 |
| VC 9 | 18.750 | 2.899 | 41.450 | 2.051 | 23.350 | 3.606 |
| VC 10 | 70.800 | .000 | 65.700 | 8.061 | 44.350 | 2.051 |

To examine whether the three different TI measurement methods produced similar outcomes, an ANOVA mixed design was used. There was a significant difference in TI measurement method ($F = 8.536$, $p = .002$), Video clips ($F = 2.949$, $p = .024$), and interaction ($F = 2.495$, $p = .013$). Because the difference among video clips and the interaction between TI measurement method and video clip was not considered as analysis subject, only Tukey's HSD on TI measurement methods was performed as a post hoc test.

The result of the post hoc test is shown in Table 23. The degrees of TI on the LIK dyad were significantly different than both the WI dyad and the Y/N dyad, but there was no significant difference between the WI dyad and the Y/N dyad. Even though the two indirect measurement methods share the same format except for the rating scale, the mean difference between two methods was significant at the level of .05.

Table 23

*Post hoc Comparisons*

| | (I) TI_Method | (J) TI_Method | (I-J) Mean Difference | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| Tukey HSD | WI | Y/N | -7.925 | 3.4262 | .069 | -16.371 | .521 |
| | | LIK | 14.145* | 3.4262 | .001 | 5.699 | 22.591 |
| | Y/N | WI | 7.925 | 3.4262 | .069 | -.521 | 16.371 |
| | | LIK | 22.070* | 3.4262 | .000 | 13.624 | 30.516 |
| | LIK | WI | -14.145* | 3.4262 | .001 | -22.591 | -5.699 |
| | | Y/N | -22.070* | 3.4262 | .000 | -30.516 | -13.624 |

*Note.* Based on observed means. The error term is MS (Error) = 117.385.

* The mean difference is significant at the .05 level.

## The Best Corresponding TI Measurement Methods for the Dependent Variable

The second part of inquiry relates to examining the relationship between level of child's behavior and level of intervention implementation. This inquiry consists of two questions.

### *Question 3: Is There a Relationship between the Degree of Child's Behavior and the Degree of TI Indicated by Each TI method?*

Figures 8, 9, and 10 are scatter plots of the WI, Y/N, and LIK dyads. A positive correlation was found between the degree of TI and the degree of the child's behavior with both the WI and the Y/N dyads. No relationship was found for the LIK dyad. To determine whether the visual inspection is reasonable, Pearson correlation coefficient of the three dyads was calculated.

The WI dyad's mean of child's behavior was 52.290, and the standard deviation was a 14.27. The mean of TI was 58.53 with a standard deviation of 18.31. As shown in Figure 8, the Pearson's coefficient was .508 with a significant correlation at the 0.05 level (2-tailed, $p = .022$). That is, 25.8% ($r^2 = 0.258$) of the proportion of variability on the degree of the child's behavior can be determined by the relationship with the degree of TI.

The mean score of the child's behavior in the Y/N dyad was 62.290 with a standard deviation of 17.760. Their mean TI was 66.455 with a standard deviation of 18.87. The Pearson's coefficient was .573 and the correlation was significant at the 0.01 level (2-tailed, $p = .008$). The data revealed that 32.8% ($r^2 = 0.328$) of total variability on the degree of the child's behavior could be determined from the relationship with the degree of TI.

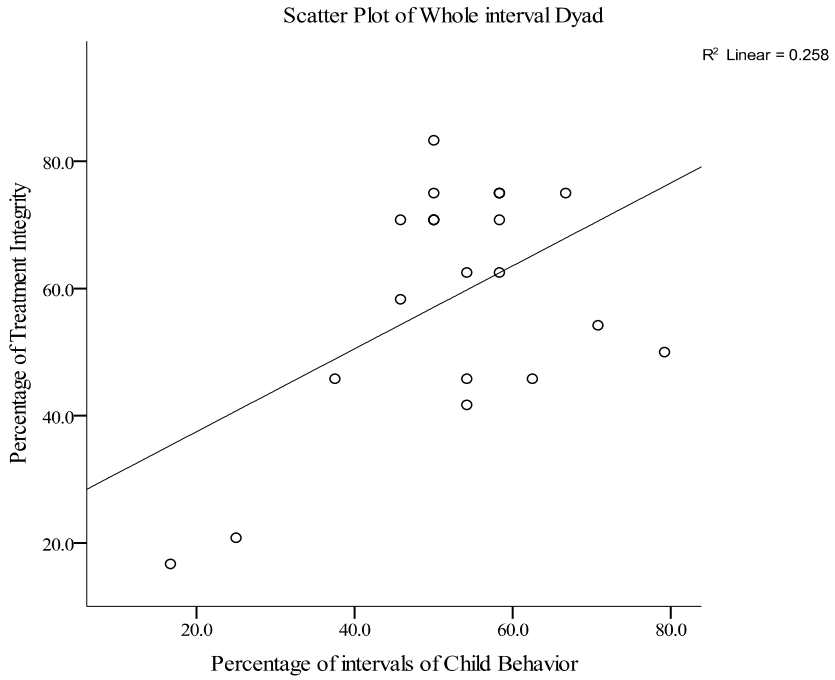


*Figure 8.* Scatter plot of the WI dyad (*N* = 20)

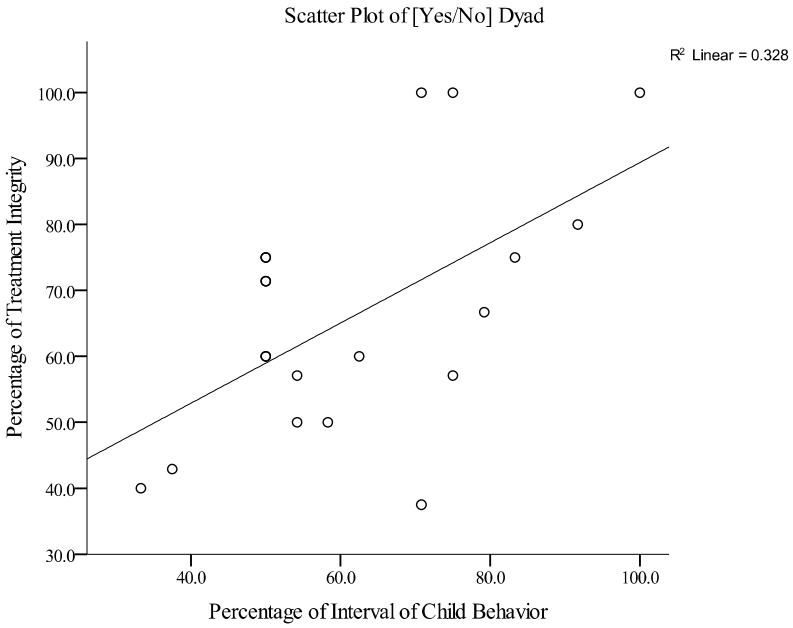


*Figure 9.* Scatter plot of the Y/N dyad (*N* = 20)

With the LIK dyad, the mean of the child's behavior was 49.380 with a standard deviation of 10.238. The mean for TI was 44.385 with a standard deviation of 13.450. The Pearson's coefficient was .105, indicating no significant correlation between two variables ($p$ = .659).

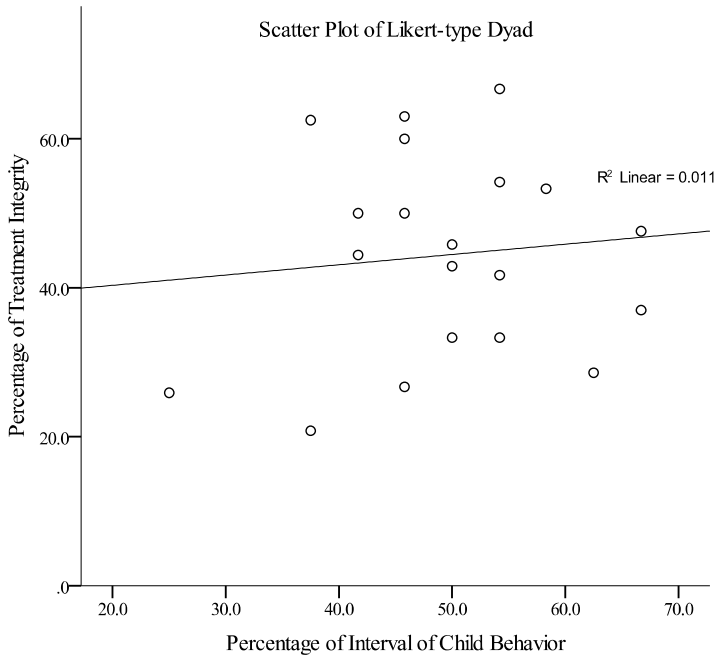


*Figure 10.* Scatter plot of the LIK dyad ($N$ = 20)

In summary, the WI and the Y/N dyads showed some correlation between the degree of the child's behavior and the degree of treatment integrity. The LIK dyad did not show a significant level of correlation.

*Question 4: Is There any Difference among the Level of Correlations?*

Both the WI dyad and the Y/N dyad showed some level of correlation between the degree of the child's behavior and the degree of TI. By using Glass and Stanley's (1970) formula, a difference between correlation coefficients from WI and Y/N dyads was examined. When $r_{xy}$ (correlation coefficient for the WI dyad) is $r_{xy}$ =.508, $r_{xz}$ (correlation coefficient for the Y/N dyad) is $r_{xz}$ =.573, and $r_{yz}$ (correlation coefficient between TI of the WI dyad and TI of the Y/N dyad) is $r_{yz}$=.153, the *z* score is 1.146. When set $p = .05$, the two critical values for the hypothesis test alphas are 0.025 ($-1.96 \leq z \leq 1.96$). The obtained value of *z* (= 1.146) falls between two critical values. Therefore, the null hypothesis is retained, indicating no significant difference between the two correlation coefficients.

# CHAPTER 5

# DISCUSSION

## Findings and Relationship to the Existing Literature

This study examined whether three different TI measurement methods produced similar TI outcomes and which, if any, set of TI data best correlated to the data regarding assessment of a child's behavior. From the onset, there were some issues regarding the definitions of specific terms related to TI measurement which needed to be addressed in this study. At the planning stages of the study, great confusion was uncovered about the specific terms, *direct observation,* and *component* or *steps*.

The majority of researchers use direct observation to refer to the data collected by observers during specific observation periods. Several researchers, who examined the collection of TI data using a component checklist, regarded this type of data collection as a precondition (Billingsley et al., 1980; Gresham, 1989). Gresham and Gansle (1993) particularly emphasized that each component of an intervention must be measured by direct observation. Furthermore, Lane and Beebe-Frankenberger (2004) classified TI assessment methods into five categories and determined a "direct observation procedure" among the five categories, differentiating behavior rating scales from direct observation procedures. Lane and Beebe-Frankenberger deemed the use of a behavior rating scale as a less direct method, because observers completed the rating scale *after* the observation. However, some researchers still regarded completing a rating scale after observation clearly as a direct observation method (Lane, Kalberg, Bruhn, et al., 2008).

The confusion over this term continued as Griffith et al. (2009) categorized collecting data while watching video clips as *indirect observation,* while Gresham et al. (2000) included using videotaping in direct observation. This ambiguity obviously has created a great deal of misunderstanding. This was confirmed by this current study's analysis of recent research, finding that the researchers of 18 studies reported that TI data were collected through direct observation. However, it was not clear whether the raters of the 18 studies collected TI data during observation or after observation. Additionally, the process which the raters of these studies used to determine whether a component in an intervention occurred was also not well-defined.

In an attempt to address this confusion, outcomes from applied psychology studies related to performance rating are presented. Borman (1978) described a three-stage process for performance evaluation: (a) observe related behavior, (b) evaluate each of these behaviors, and (c) weigh evaluation to arrive at a single rating on a performance dimension. Per this information, it was likely that the some of the observers of the 18 studies might have completed their checklists at the end of their observation sessions. If so, the observers may have had to rely upon memory and the 18 studies would have to be re-categorized as *indirect observation methods* in the analysis presented in Chapter 2 in this study. However, since no clear methods of how the TI data were collected were described in any of the studies, there is no evidence to support this assumption.

In the last decade, some researchers realized the confusion about this term, and attempted to reclassify TI measurement methods. For example, Gresham et al. (2000) divided TI measurement methods into two categories: direct assessment and indirect

assessment. Direct assessment included the consideration of each component of an intervention to be recorded *during* observation whether seen in person or by video-tape. Based on the Gresham et al.'s classifications, two of the current study's TI measurement methods, the Y/N component checklist and the Likert-type rating scale were categorized as indirect measures, even though the raters immediately completed the checklist at the end of the observation.

The other terms which have perplexed researchers are *components* or *steps*. For example, Billingsley et al. (1980) used term "component" to refer to an antecedent, behavior, and consequence in an intervention. Also, Gresham (1989) used the word "component" in his research, but did not clearly define what he meant by this term. It might be assumed a component related to one of the steps in a task analysis. Moreover, Gansle and McMahon (1997) used the words "component" and "step" interchangeably. However, Umbreit et al. (2007) employed a new word, "element," used as a replacement term for "component" as it was used in the Billingsley et al.'s study.

Researchers typically plan behavioral interventions focusing on changes in antecedents, behaviors, and consequences. If referred to as "components," some researchers may concentrate on just one component, while others may focus on all three components when designing a behavioral intervention. Also, it may possible that one component includes several procedural steps. On the other hand, if a researcher plans a behavioral intervention based on a task analysis, the intervention may consist of "steps" and changing an antecedent, or teaching a new behavior, or changing a consequence, each of which could be included in the intervention as a step.

There may be no golden rule for resolving these confusions. However, carefully thought-out and thorough descriptions of all operational definitions, procedure-related independent variables, and procedures regarding treatment integrity data collection may help researchers identify the specific information needed to successfully conduct research about TI.

Secondly, during the training sessions for all six raters regarding the observation of the child's behavior, the same whole-interval assessment form was used. If two raters disagreed on an interval, there was discussion time to further analyze the child's behaviors in that interval. This process may have resulted in maintaining higher IOA during training sessions. Furthermore, this training method may have resulted in the WI dyad having higher IOAs for rating both the child's behavior and TI than the IOAs of other two dyads. This assumption is based on a suggestion of Billingsley, et al. (1980) that the use of an interval recording system gives the raters more opportunities to check whether the correct type of consequences were delivered, or whether delivery was appropriate. However, there is no definitive evidence to support this assumption.

Thirdly, Horner and colleagues (2005) suggested that measurement of the degree of TI was effective when using a direct measurement method or an equivalent. However these researchers did not specifically mention what an equivalent measure was. Lane and Beebe-Frankenberger (2004) suggested that, by virtue of assessing TI from multiple perspectives, researchers may have more accurate TI data. Specifically, different TI data sets which would be collected from several different sources, (i.e., direct observation, self-report, and/ or permanent products) may increase the accuracy of TI assessment.

Shavelson and Dempsen-Atwood (1976) suggested four possibilities to explain the reason that there were mixed conclusions about the relationship between teaching behavior and student's outcomes, particularly mentioning the limitations in measurement methods. Based on the information from these studies, one question arose which became an underlying query for this study: What if the TI measurement methods themselves are unreliable?

According to the results of this study, there was no difference between the WI dyad and the Y/N dyad, but the LIK dyad's TI data were significantly different from those of the other two. Of the three TI measurement methods, one was interval-based direct assessment and two were indirect assessment methods. However, when comparing assessment systems, the WI measurement method and the Y/N checklists used similar assessment schema (i.e., yes/no/not applicable), while the Likert-type rating scale used a different assessment schema (i.e., a 4-point scale indicating "never", "seldom", "often," "almost always," and "not applicable"). Thus, when a rater uses an interval-based measurement method, the rater not only decides which intervention components must be delivered, but also which components are appropriately delivered in that particular interval. Similarly, when a Y/N component checklist is used, the rater should determine which of the components should have been implemented and whether the intervener correctly delivered the required components. Both methods (WI and Y/N) follow the "yes," "no," or "n/a" answer format. However, one critical characteristic that differentiates the two methods is that the WI measurement method relies on an instant decision, while the indirect Y/N measurement method depends on the rater's memory.

On the other hand, applying a Likert-type rating scale highly depends on a rater's judgment. The LIK dyad provided the most consistent data among the three dyads in the assessment of the child's behavior. The LIK dyad's average percentage of IOA (92.1%) in the rating sessions was very similar to their IOA in the training sessions (92.5%). However, with regard to assessing the degree of TI, the two raters had difficulty in reaching and keeping a certain level of IOA. Frank-Major (1985) claimed that the rating format itself was less important than the cognitive effect on the observer. Thorndike (1920, cited from Frank-Major, 1985) suggested the possibility of a "halo" effect, referring to a tendency to follow a general impression of the rate rather than an assessment of the actual levels of behavior. Also, Feldman (1981) explained how people classified and organized information and put the information into general categories; once a person made an evaluation, the person consequently used the result as a basis for later evaluation. That is, initial judgments, rather than factual information, determined later evaluation.

The WI and Y/N measurement methods share a commonality in their measurement system, whereas the Y/N and LIK systems share a similar characteristic that accurate application of both methods rely heavily on the rater's memory. In this study, it appeared that the two indirect methods seemed to be unreliable, because the cognitive processes of the raters in these two dyads may have affected their decisions as raters. However, outcomes from this study do not significantly support this idea.

Lastly, when analyzing causes of differences between two raters, G-theory was used in this study. Seven sources of variance were identified. Through indices of dependability,

it was apparent that the six raters consistently observed and rated the same behaviors described in the intervention without observer drift. However, there were huge gaps on several video clips between the raters' scores in the Y/N dyad and LIK dyad. Differences between the percentages of TI for the two raters in each dyad did not necessarily reflect either the percentage of TI IOA or the degree of kappa. For example, even though two raters reached low TI IOA (e.g., IOA = 79% and $k$ = .32) they may have identified similar percentage of TI (e.g., 16.7% and 20.7%). Furthermore, there could be sizable differences between the percentages of TI (e.g., 54.2% and 70.8%) while TI IOA was still acceptable (e.g., IOA = 83.3% and $k$ = .65).

Moreover, the percentages of total variance indicated that the main cause of rating inconsistencies may not have been the raters, but the systematic variance. That is, the variance from the TI method affected the percentage of TI, and it resulted in the percentage gap between the two raters in the Y/N dyad and the LIK dyad (see *SD* in Table 22). There is a possibility that those data from the two raters may have been unreliable because the percentages of TI may not have shown the actual degree of treatment integrity but may have indicated a distorted degree of treatment integrity affected by characteristics of the TI measurement method. If so, it may not have been appropriate to compare TI outcomes from three dyads and to examine relationship between the level of the child's behavior and the level of TI.

This was the first attempt to directly compare three different TI measurement methods that are commonly used in behavioral intervention research. The study was based on the research outcomes of Wood et al. (2007), and of Zvoch (2009). In contrast

to the previous researchers' concerns, there was no difference between the WI dyad and Y/N dyad, not only in terms of the percentages of TI but also regarding the correlation coefficient between the degree of TI and the degree of the child's behavior. However, there was a critical question that arose during the analysis process. Even though two raters in the Y/N dyad consistently and similarly observed the behaviors of both the child and the instructor according to the phi values, the TI measurement method itself accounted for 42.89% of total variance with 31.47% of total variance deriving from the interaction between the TI measurement method and the video clips. Because of this, the impact of the TI measurement method on the actual percentage of TI and the degree of TI IOA may be unexpectedly strong. Also, it was obvious that there were several huge gaps between the percentages of the two raters from the Y/N dyad. Therefore, it is questionable whether the ANOVA and Pearson's correlation analyses are meaningful and reliable, as the result could be a false presentation of the TI measurement methods.

## Limitations of the Study

This study contains at least six limitations. First, even though this study focused on comparing three different TI measurement methods, there were just two raters in each dyad and only ten 6-min video clips from one study conducted in a home environment that were used for rating purposes. Therefore, it is not possible to generalize the study's results to other forms of the three TI measurement methods or to different intervention situations. The original study was conducted in a home-based environment, and was based on one-on-one instruction. However, in the results of analysis of the literature,

almost all the previous studies were conducted in school-based settings. There are numerous uncontrolled variables in a classroom environment, and there is a question whether the different TI measurement methods would show similar comparison outcomes in the home environment.

Second, because of the difficulty of scheduling the training sessions, the six raters could not be trained together. As a result there may have been some variation in the quality of the training sessions. The training sessions included observation of video clips and discussion of disagreements between two raters by using operational definitions of the child's behavior as well as descriptions of the intervention components. Beyond the basic training content, the main discussion topics were brought up by the two raters in each dyad. With regard to assessment of the child's behavior, the six raters used the same measurement method (i.e., whole interval measure method). Therefore, the raters had the same opportunities to discuss the observed behavior interval-by-interval and to discuss any differences or acceptable levels of variance in the observed behavior.

On the other hand, the discussion topics about TI differed greatly. The WI dyad used the same process as used with rating the child's behavior. The raters checked and discussed the instructor's intervention implementation on an interval-by-interval basis. However, the two indirect measurement method dyads (i.e., Y/N and LIK) mainly focused on whether a certain step should have been implemented (i.e., n/a), how often a certain step took place, or whether the instructor "correctly" implemented each step of the intervention.

Moreover, because of time limitations, the PF IOA data for training sessions for the WI dyad could not be collected. This also may have affected the quality of the discussion. Time limitations also provided some disadvantages to the raters in the LIK dyad. Even though they could not reach the established 0.65 kappa level during the training sessions, the training session had to be terminated, although at the level of 0.62. This could have affected the degree of disagreement on TI during the rating sessions.

To ensure that the six raters similarly observed behaviors of both child and instructor, the six raters should have been trained together. IOA among all six raters and PI, and between the two raters in each dyad and the PI, should have reached 85% with a 0.65 kappa score, as shown in Figure 11.

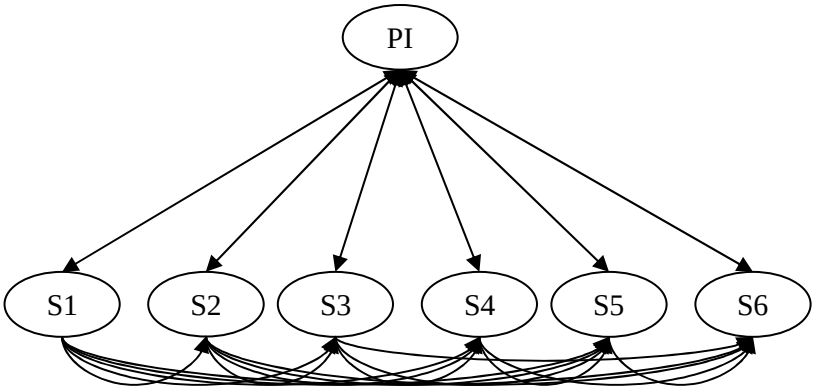


*Figure 11*. IOA among six raters and PI

Third, the original study focused on food consumption behavior and there were no established procedures for managing off-task behaviors. Various behavior management strategies were used by the instructor during the intervention phase. In the present study, specific procedures to manage off-task behavior (i.e., steps 9-11) were added to the

intervention protocol. From a measurement standpoint, this was an advantage in the present study. However, in most studies, all assessed intervention components are likely to be established in advance. Unless there is considerable inconsistency in implementation, there is likely to be a relatively high level of TI in other studies, which would reduce variability, making it more difficult to assess differences between various TI measurement methods.

Fourth, among the 34 video clips, 10 were edited and used for rating the child's behavior and TI. The number of video clips was chosen based on the percentage of observation sessions that is commonly used for IOA (i.e., 25% – 30% of observation sessions). However, it is not certain whether the percentage of sessions was enough to show any consistency in treatment integrity. The inconsistency across presentation of the video clips was an issue related to editing the video clips. Because the lengths of the original filmed sessions varied, all video clips were edited to six minutes in length. If the original length of a video clip was too short, footage from another video clip was added to make the clip the correct length. Consequently, at times, there were noticeable differences within one video clip. The WI dyad used whole-interval measurement for TI, which provides the most conservative possible estimates of behavior. The combined video clips may have affected the WI dyad's assessment of both the child's behavior and TI more than the two indirect TI measurement method dyads.

Fifth, during the rating sessions, raters were required to observe five video clips per day. Even though there were short breaks between observations of the video clips, a carry-over effect may have occurred with the two indirect measure dyads. Indirect

measurement highly relies on a rater's memory and capacity to recall events, possibly leading the raters to under- or over-estimate the degree of TI.

Lastly, there was a discrepancy between the analysis of the results of previous research and the current study's usage of a Y/N component checklist and a Likert-type rating scale. In this study, a Y/N component checklist was used as an indirect measurement method, whereas most previous researchers categorized this checklist as a direct measurement method. Also, the current study utilized a 4-point rating scale rather than a 3-point rating scale, as in prior research. Even though there was an obvious purpose for using the two indirect measurement methods, the results of this study did not directly indicate either any similarities or differences between the WI measurement method and Y/N checklist for direct TI assessment or between the direct measurement method and a commonly used 3-point rating scale as an indirect measurement method.

## Implications for Future Research

This study examined (a) the similarities among TI outcomes when using three different TI assessment methods and (b) whether any of the TI measurement methods best corresponded to the assessed degree of a child's behavior. However, this study was merely an initial attempt. There is still a great need for researchers to find mechanisms that ensure more reliable methods of TI measurement and the variables that produce more rigorous TI data, especially when a study is conducted using a single-subject design. Based on the outcomes and limitations of this study, some implications are suggested.

First of all, as Backeman and Gottman (1997) mentioned, training observers is a key

issue. However, most published studies neglected to report how and how long a researcher should train observers, especially with regard to the measurement of TI. There are two reasons why researchers should pay more attention to training method as well as the length of observer trainings. On the one hand, attaining higher IOA percentage scores directly relates to the consistency with which two raters observe and rate the same behavior based on clearly understood operational definitions of that behavior. Sometimes after training, observer drift may occur, possibly a result of a lack of training.

On the other hand, treatment implementation can deviate from a prescribed intervention plan. Baer et al. (1987) questioned the range of variance of assessed TI that still produced sufficient intervention effectiveness. There is no rule to decide how far one can deviate from an intervention plan and still obtain effective results (Gresham, et al., 2000). However, the rule of thumb may depend on observers' experiences, which are learned by analyzing and discussing the instructor's intervention and implementation behaviors. As demonstrated by the results in this study, a specific length of time was required until the two raters in each dyad reached a certain level of IOA. In this study, approximately six hours of training was required, and the length of time was not enough to maintain a high degree of IOA across rating sessions. Thus, an implication for future research is that for effective training to occur, researchers should examine appropriate guidelines for training, different lengths of training sessions, appropriate IOA levels, and training methods themselves.

Second, this study used video clips obtained in a modified home-based intervention. For extending and reexamining the outcomes of this study, the efficacy and reliability

among different TI measurement methods should be explored in school-based behavior intervention. As Yeaton and Sechrest (1981) pointed out, the complexity of a treatment is directly related to the degree of TI. Currently, almost all school-based studies use multiple-component interventions. Some TI methods may not be sensitive enough to assess the TI in a complex intervention. These inappropriate measurement methods may misrepresent the effect of an intervention. Therefore, TI measurement methods should be examined further in school-based environments so that those inappropriate methods for complex school-based interventions can be identified.

Third, currently a 3-point rating scale has been used in several studies for measuring TI for practical purposes. In the school psychology field, researchers who conduct behavioral interventions follow a behavioral consultation model. In the model, the researcher allows the consultation process to be a self-correcting process of the consultee (i.e., teacher). Therefore, teachers are actively involved in collecting data (Erchul & Schulte, 1996). Moreover, one of goals of applied behavior analysis is to build a school personnel's capacity to conduct function-based intervention (Lane et al., 2007). As a result, from both the teacher's and the researcher's perspectives, component checklist forms (especially Likert-type rating scales) have been used in several behavioral intervention studies. However, based on the results of this study, behavior rating scales may not give the consultee or the teacher correct performance feedback. Moreover, because the rating scale used in this study was 4-point, the efficacy and reliability of a 3-point rating scale should be examined.

Fourth, still there is no agreement among researchers about the appropriate percentage of observation sessions for monitoring TI. As already mentioned, collecting TI data with dependent variable data simultaneously is ideal, but as Umbreit and colleagues (2007) pointed out, researchers need to strike a balance between rigor and feasibility in obtaining accurate information about treatment integrity. In addition, insufficient data points for TI monitoring can not only bring a high risk of treatment inaccuracy, but can also threaten both the internal and external validity of the intervention. Therefore, researchers should investigate not only how many data points are needed but also what percentage of observation sessions is appropriate for examining and maintaining consistency of TI.

Fifth, this study was the first attempt to evaluate different TI measurement methods by analyzing the sources of variance which are causes of inconsistency between two raters. The analysis was helpful to recognize that a method itself can be a main source of variance. When creating a TI measurement method, it may benefit the researcher to examine whether systematic variance related to the method is the main source of inconsistency between two raters by using G-theory and to reduce the systematic variance which is caused by the TI measurement method before using it in practice.

Lastly, researchers who mainly use single-subject designs rely upon high percentage scores of IOA as a way to convince others of the accuracy and reliability of the recorded data (Bakeman & Gottman, 1997). However, as Johnston and Pennypacker (1993) mentioned, interobserver agreement actually does not provide any information about accuracy or reliability. In this study, three indices, IOA percentages, kappa scores and phi

values were used to examine whether raters consistently observed and rated behavior. However, the relationship among these three indices has yet to be explored. Researchers must continue investigating ways to ensure the accuracy and reliability of interobserver agreement, particularly when using single-subject designs.

# APPENDIX A: RATER INFORMATION

## RATER INFORMATION

1. Name: _____ Gender: F M

2. Current career status: _____

3. Degree: (a) Ph. D (b) MA (c) M. Ed in _____

   (d) Other _____

4. Courses you have taken (multiple choice) related to behavior principle, human behavior observation, assessment, & intervention

   (a) SERP 502

   (b) SERP 529 A

   (c) SERP 529 B

   (d) Other: _____

5. CITI completion date: _____

6. I'm familiar with collecting human behavior data

   (1) Yes (2) No

# APPENDIX B:

# PROCEDURAL FIDELITY CHECKLISTS

APPENDIX B.1

Procedural fidelity Checklist
(Training Session I)

Observer: _____ Date & Time: _____/_____/_____

Dyad: Whole interval / [Yes/No]/ Rating scale

| Steps | Yes | No |
|---|---|---|
| 1. Confirm consent form | | |
| 2. Confirm CITI completion | | |
| 3. Identify dyad, primary observer, and second observer | | |
| 3. Introduce intervention in video clips | | |
| 4. Watch a video clip | | |
| 5. Discuss replacement behavior and intervention procedures | | |
| 6. Distribute and explain assessment form & measurement method | | |
| 7. Training 1: watching video clip | | |
| 8. Calculate IOA | | |
| 9. Discuss the result and collect assessment sheet | | |
| 10. Training 2: watching video clip | | |
| 11. Calculate IOA | | |
| 12. Discuss the result and collect assessment sheet | | |
| 13. Training 3: watching video clip | | |
| 14. Calculate IOA and collect assessment sheet | | |
| 15. Break or Finish | | |

APPENDIX B.2

Procedural fidelity Checklist
(Training Session II)

Observer: _____ Date & Time: _____/_____/_____

Dyad: Whole interval / [Yes/No]/ Rating scale

| Steps | Yes | No |
|---|---|---|
| 1. Review intervention | | |

| Steps | Training 4 | | Training 5 | | Training 6 | |
|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No |
| 2. Distribute TI assessment sheet | | | | | | |
| 3. Watch a video clip | | | | | | |
| 4. Q & A | | | | | | |
| 5. Collect data sheet | | | | | | |
| 6. 4 min. break | | | | | | |

Observer: _____ Date & Time: _____/_____/_____

Dyad: Whole interval / [Yes/NO]/ Rating scale

| Steps | Yes | No |
|---|---|---|
| 1. Review intervention | | |

| Steps | Training 7 | | Training 8 | | Training 9 | |
|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No |
| 2. Distribute TI assessment sheet | | | | | | |
| 3. Watch a video clip | | | | | | |
| 4. Q & A | | | | | | |
| 5. Collect data sheet | | | | | | |
| 6. 4 min. break | | | | | | |

APPENDIX B.3

Procedural fidelity Checklist
(Rating Sessions)

Dyad: Whole interval / [Yes/No]/ Rating scale

<Sessions I>

Observer: _____ Date & Time: _____/_____/_____

| Steps | VC1 | | VC2 | | VC3 | | VC4 | | VC5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No |
| 1. Distribute TI assessment sheet | | | | | | | | | | |
| 2. Watch a video clip | | | | | | | | | | |
| 3. collect data sheet | | | | | | | | | | |
| 4. 4-minute break | | | | | | | | | | |

<Session II>

Observer: _____ Date & Time: _____/_____/_____

| Steps | VC6 | | VC7 | | VC8 | | VC9 | | VC10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No |
| 1. Distribute TI assessment sheet | | | | | | | | | | |
| 2. Watch a video clip | | | | | | | | | | |
| 3. collect data sheet | | | | | | | | | | |
| 4. 4-minute break | | | | | | | | | | |

# APPENDIX C

# BEHAVIOR DATA RECORDING FORMS

# APPENDIX C.1 Interval Data Recording Form

Observer: _____ Behavior: On-task (food consumption)

Interval Length: 15 sec. (whole interval) Observation length: 6 minutes

On-task behavior: *staying in the table and seat area, waiting for task direction, picking up the spoon (or food – if fingers are used) within 5 sec., taking the spoon (or food) to the mouth within 5 sec., and chewing food or touching the tongue to the food with or without physical assistance, or following direction.*

Intervention:

1. When the child is ready (i.e. seated or standing in seat area; mouth empty), place a bite on the plate and give task direction
2. If child takes a bite, praise (e.g., "good job") him.
3. if child does not pick up food or spoon within 5 sec., use hand-over-hand physical assistance to place child's hand on spoon
4. If child does not move spoon/food to mouth within 5 seconds, use hand-over-hand physical assistance to guide spoon to mouth
5. If child does not open his mouth, return the spoon to the plate
6. After any refusal, begin new trial with 1/2 bite and repeat 1-5
7. If child refuses to take 1/2 bite, begin new trial with 1/4 bite and repeat 1-5.
8. If 1/4 bite is not consumed, verbally prompt child to use tongue to touch bite
9. If child escapes from table, briefly redirect (short statement) with minimal interaction.
10. Give up to 3 redirections (10 sec between warnings). If child does not return to table after 3 redirections, physically bring child back to table and then repeat 1-5 with a smaller bite.
11. If child plays with a cracker, ignore him

<Video Clip # >

| Behavior Memo | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intervals | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Behavior | | | | | | | | | | | | |
| Intervention | | | | | | | | | | | | |
| Intervention Memo | | | | | | | | | | | | |

| Behavior Memo | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intervals | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| Behavior | | | | | | | | | | | | |
| Intervention | | | | | | | | | | | | |
| Intervention Memo | | | | | | | | | | | | |

On-task behavior: _____/_____ = _____ % Behavior IOA: ___/____ = ____%

Intervention: _____/_____ = _____ % Intervention IOA:_____/_____ = ____%

APPENDIX C.2

Child's behavior Data Recording Form

(DYAD 2 & 3)

Observer: _____ Behavior: On-task (food consumption)

Interval Length: 15 sec. (Whole interval) Observation length: 6 minutes

On-task behavior:

- *staying in the table and seat area, waiting for task direction, picking up the spoon (or food – if fingers are used) within 5 sec., taking the spoon (or food) to the mouth within 5 sec., and chewing food or touching the tongue to the food with or without physical assistance, or following direction.*

<Video Clip # >

The child's behavior:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| | | | | | | | | | | | |

Memo:

On-task behavior: _____/_____ = _____%

The child's behavior IOA:_____/_____= _____%

APPENDIX C.3

[Yes/No] Component Checklist

Observer: _____ Behavior: On-task (food consumption)

Interval Length: 15 sec. (Whole interval) Observation length: 6 minutes

<Video Clip # >

Intervention: *occurrence/ nonoccurrence*

| Components | Yes | No | N/A |
|---|---|---|---|
| 1. When the child is ready (i.e. seated or standing in seat area; mouth empty), place a bite on the plate and give task direction | | | |
| 2. If child takes bite, praise (e.g., "good job") him | | | |
| 3. If child does not pick up spoon/food within 5 seconds, use hand-over-hand physical assistance to place child's hand on spoon | | | |
| 4. If child does not move spoon/food to mouth within 5 seconds, use hand-over-hand physical assistance to guide spoon to mouth | | | |
| 5. If child does not open his mouth, return the spoon to the plate | | | |
| 6. After *any* refusal, begin new trial with 1/2 bite and repeat 1-5 | | | |
| 7. If child refuses to take 1/2 bite, begin new trial with 1/4 bite and repeat 1-5 | | | |
| 8. If 1/4 bite is not consumed, verbally prompt child to use tongue to touch bite | | | |
| 9. If child escapes from table, briefly redirect (short statement) with minimal interaction | | | |
| 10. Give up to 3 redirections (10 sec between warnings). If child does not return to table after 3 redirections, physically bring child back to table and then repeat 1-5 with a smaller bite | | | |
| 11. If child plays with a cracker, ignore him | | | |
| ***N of responses = N of Yes/ (N of components – N. of N/A)*** | | | |
| ***%*** | | | |
| ***Intervention IOA*** | ____/____ = ___% | | |

APPENDIX C.4

Likert-Type Rating Scale

Observer: _____ Behavior: On-task (food consumption)
Interval Length: 15 sec. (whole interval) Observation length: 6 minutes

<Video Clip # >

Intervention: *the level of correct implementation*

| Components | Never 0 | Seldom 1 | often 2 | Almost always 3 | N/A N |
|---|---|---|---|---|---|
| 1. When the child is ready (i.e. seated or standing in seat area; mouth empty), place a bite on the plate and give task direction | 0 | 1 | 2 | 3 | N |
| 2. If child takes bite, praise (e.g., "good job") him | 0 | 1 | 2 | 3 | N |
| 3. If child does not pick up spoon/food within 5 seconds, use hand-over-hand physical assistance to place child's hand on spoon | 0 | 1 | 2 | 3 | N |
| 4. If child does not move spoon/food to mouth within 5 seconds, use hand-over-hand physical assistance to guide spoon to mouth | 0 | 1 | 2 | 3 | N |
| 5. If child does not open his mouth, return the spoon to the plate | 0 | 1 | 2 | 3 | N |
| 6. After *any* refusal, begin new trial with 1/2 bite and repeat 1-5 | 0 | 1 | 2 | 3 | N |
| 7. If child refuses to take 1/2 bite, begin new trial with 1/4 bite and repeat 1-5 | 0 | 1 | 2 | 3 | N |
| 8. If 1/4 bite is not consumed, verbally prompt child to use tongue to touch bite | 0 | 1 | 2 | 3 | N |
| 9. If child escapes from table, briefly redirect (short statement) with minimal interaction | 0 | 1 | 2 | 3 | N |
| 10. Give up to 3 redirections (10 sec between warnings). If child does not return to table after 3 redirections, physically bring child back to table and then repeat 1-5 with a smaller bite | 0 | 1 | 2 | 3 | N |
| 11. If child plays with a cracker, ignore him | 0 | 1 | 2 | 3 | N |
| ***Average*** | /(3X ) | | | | |
| ***Intervention IOA*** | ____/____ = _____ % | | | | |

# REFERENCES

References marked with an asterisk (*) indicate studies that were used in final inclusion for analyzing TI measure in Chapter 2, Literature Review.

*Allen-DeBoer, R., Malmgren, K., & Glass, M. (2006). Reading instruction for youth with emotional and behavioral disorders in a juvenile correctional facility. *Behavioral Disorders, 32*, 18-28.

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press.

*Banda, D., & Kubina, R. (2006). The effects of a high-probability request sequencing technique in enhancing transition behaviors. *Education and Treatment of Children, 29*, 507-516.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Baer, D., Wolf, M., & Risley, T. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91-97.

Baer, D., Wolf, M., & Risley, T. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20*, 313-327.

Billingsley, F., White, O. R., & Munson, R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment, 2*, 229-241.

Borman, W. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*, 135-144.

*Bowman-Perrott, L., Greenwood, C., & Tapia, Y. (2007). The efficacy of CWPT used in secondary alternative school classrooms with small teacher/pupil ratios and students with emotional and behavioral disorders. *Education and Treatment of Children, 30*. 65-87.

*Bradshaw, C., Reinke, W., Brown, L., Bevans, K., & Leaf, P. (2008). Implementation of school-wide positive behavioral interventions and supports (PBIS) in elementary schools: Observations from a randomized trial. *Education and Treatment of Children, 31*, 1-26.

*Brown, K., & Mirenda, P. (2006). Contingency mapping: Use of a novel visual support strategy as an adjunct to functional equivalence training. *Journal of Positive Behavioral interventions, 8*, 155-164.

*Campbell, A., & Anderson, C. (2008). Enhancing effects of check-in/check-out with function-based support. *Behavioral Disorders, 33*, 233-245.

*Campbell, M., Helf, S., & Cooke, N. (2008). Effects of adding multisensory components to a supplemental reading program on the decoding skills of treatment resisters. *Education and Treatment of Education, 31*, 267-295.

*Chan, J., & O'Reilly, M. (2008). A Social Story™ intervention package for students with autism in inclusive classroom settings. *Journal of Applied Behavior Analysis, 41*, 405-409.

*Christenen, L., Young, R., & Marchant, M. (2007). Behavioral intervention planning: Increasing appropriate behavior of a socially withdrawn. *Education and Treatment of Children, 30,* 81-103.

*Cihak, D., Alberto, P., & Fredrick, L. (2007). Use of brief functional analysis and intervention evaluation in public settings. *Journal of Positive Behavioral interventions, 9,* 80-93.

Cochrane, W., & Laux, J. (2008). A survey investigating school psychologists' measurement of treatment integrity in school-based interventions and their beliefs about its importance. *Psychology in the School, 45,* 499-507.

*Delano, M., & Snell, M. (2006). The effects of social stories on the social engagement of children with autism. *Journal of Positive Behavioral intervention, 8,* 29-42.

*Dib, N., & Sturmey, P. (2007). Reducing student stereotype by improving teachers' implementation of discrete-trial teaching. *Journal of Applied Behavior Analysis, 40,* 339-343.

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

*Downs, A., Conley, R., Johansen, M., & Fossum, M. (2007). Using discrete trial teaching within a public preschool program to facilitate skill development in students with developmental disabilities. *Education and Treatment of Children, 30,* 1-27.

Erchul, W., & Schulte, A. (1996). Behavioral consultation as a work in progress: A reply to Witt, Gresham, and Noell. *Journal of Educational and Psychological Consultation, 7,* 345-354.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.

*Feng, H., Lo, Y., Tsai, S., & Cartledge, G. (2008). The effects of theory-of-mind and social skill training on the social competence of a sixth-grade student with autism. *Journal of Positive Behavioral interventions, 10,* 228-242.

Frank-Major, S. L. (1985). *Accuracy in performance appraisals: A comparison of two rater cognitive process models*. Unpublished master's thesis, Texas A&M University, Texas.

*Ganz, J., Bourgeois, B., Flores, M., & Campos, A. (2008). Implementing visually cued imitation training with children with autism spectrum disorders and developmental delays. *Journal of Positive Behavioral interventions, 10,* 56-66.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* 11.0 update (4th ed.). Boston: Allyn & Bacon.

*Gilbertson, D., Duhon, G., Witt, J., & Dufrene, B. (2008). Effects of academic response rates on time-on-task in the classroom for students at academic and behavior risk. *Education and Treatment of Children, 31,* 153-165.

*Gilbertson, D., Witt, J., Duhon, G., & Dufrene, B. (2008). Using brief assessments to select math fluency and on-task behavioral interventions: An investigation of treatment utility. *Education and Treatment of Children, 31,* 167-181.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall Inc.

*Gortmaker, V., Daly, E., McCurdy, M., Persampieri, M., & Hergenrader, M. (2007). Improving reading outcomes for children with learning disabilities: Using brief experimental analysis to develop parent-tutoring interventions. *Journal of Applied Behavior Analysis, 40,* 203-221.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18,* 37–50.

Gresham, F. M., & Gansle, K. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review, 22,* 254-273.

Gresham, F. M., Gansle, K., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis, 26,* 257–263.

Gresham, F. M., MacMillan, D., Beebe-Frankenberger, M., & Bocian, K. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practices, 15,* 198-205.

*Gresham, F. M., Van, M., & Cook, C. (2005). Social skills training for teaching replacement behaviors: Remediating acquisition deficits in at-risk students. *Behavioral Disorders, 31,* 363-377.

Griffith, A., Hurley, K., & Hagaman, L. (2009). Treatment integrity of literacy interventions for students with emotional and/or behavioral disorders. *Remedial and Special Education, 30,* 245-255.

*Gulchak, D. (2008). Using a mobile handheld computer to teach a student with emotional and behavioral disorder to self-monitor attention. *Education and Treatment of Children, 31,* 567-581.

*Hagan-Burke, S., Burke, M., & Sugai, G. (2007). Using structural analysis and academic-based intervention for a student at risk of EBD. *Behavioral Disorders, 32,* 175-191.

Hagermoser Sanetti, L., & Kratochwill, T. (2008). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy, 4*, 95-114.

Hagermoser Sanetti, L., & Kratochwill, T. (2009). Treatment integrity assessment in the schools: An evaluation of the treatment integrity planning protocol. *School Psychology Quarterly, 24*, 24-35.

Hall, J. (1998). Fidelity: A crucial question in translating research to practice. *Journal of Early Intervention, 21,* 294-296.

*Hawken, L., MacLeod, S., & Rawlings, L. (2007). Effects of the behavior education program (BEP) on office discipline referrals of elementary school students. *Journal of Positive Behavioral interventions, 9,* 94-101.

Henggeler, S., Melton, G., Brondino, M., & Scherer, D. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity. *Journal of Consulting and Clinical Psychology, 65,* 821-833.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education.

*Council of Exceptional Children, 71,* 165–179.

Institute for Education Sciences. (2009). Request for application: Education research and development center program. Retrieved September, 1, 2009, from http://ies.ed.gov/funding/pdf/2010_84305c.pdf

*Jameson, M., McDonnell, J., Johnson, J., Riesen, T., & Polychronis, S. (2007). A comparison of one-to-one embedded instruction in the general education classroom and one-to-one massed practice instruction in the special education classroom. *Education and Treatment of Children, 30,* 23-44.

*Johnson-Gros, K., Lyons, E., & Griffin, J. (2008). Active supervision: An intervention to reduce high school tardiness. *Education and Treatment of Children, 31*, 39-53.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

*Jones, E., & Feeley, K. (2007). Teaching spontaneous responses to young children with autism. *Journal of Applied Behavior Analysis, 40,* 565-570.

Kamps, D. Wendland, M., & Culpepper, M. (2006). Active teacher participation in functional behavior assessment for students with emotional and behavioral disorders risks in general education classrooms. *Behavioral Disorders, 31,* 128–146.

*Kay, S., Harchik, A., & Luiselli, J. (2006). Elimination of drooling by an adolescent student with autism attending public high school. *Journal of Positive Behavior Interventions, 8,* 24-28.

Kazdin, A. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

*Kern, L., Gallagher, P., Starosta, K., Hickman, W., & George, M. (2006). Longitudinal outcomes of functional behavioral assessment-based intervention. *Journal of Positive Behavioral interventions. 8*, 67-78.

*Kern, L., Starosta, K., Cook, C., Bambara, L., & Gresham, F. (2007). Functional assessment-based intervention for selective mutism. *Behavioral Disorders, 32,* 94-108.

*Kolko, D., Herschell, A., & Scharf, D. (2006). Education and treatment for boys who set fires: Specificity, moderators, and predictors of recidivism. *Journal of Emotional and Behavioral Disorders, 14,* 227-239.

*Lambert, M. C., Cartledge, G., Heward, W., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavioral interventions, 8*, 88-99.

Lane, K. & Beebe-Frankenberger, M. E. (2004). *School-based interventions: The tools you need to succeed.* Boston, MA: Allyn & Bacon.

Lane, K., Bocian, K., MacMillan, D., & Gresham, F. (2004). Treatment integrity: An essential – but often forgotten – component of school-based interventions. *Preventing School Failure, 48,* 36–43.

Lane, K., Kalberg, J. R., Bruhn, A. L., Mahoney, M., & Driscoll, S. (2008). Primary prevention programs at the elementary level: Issues of treatment integrity, systematic screening, and reinforcement. *Education and Treatment of Children, 31*, 465-494.

*Lane, K., Rogers, L., Parks, R., Weisenbach, J., Mau, A., Merwin, M., & Bergman, W. (2007). Function-based interventions for students who are nonresponsive to primary

and secondary prevention efforts: Illustrations at the elementary and middle school levels. *Journal of Emotional and Behavioral Disorders, 15*, 169-183.

*Lane, K., Thompson, A., Reske, C., Gable, L., & Barton-Arwood, S. (2006). Reducing skin picking via competing activities. *Journal of Applied Behavior Analysis, 39,* 459-462.

*Lane, K., Weisenbach, J., Little, A., Phillips, A., & Wehby, J. (2006). Illustrations of function-based interventions implemented by general education teachers: Building capacity at the school site. *Education and Treatment of Children, 29,* 549-571.

*Lane, K., Weisenbach, J., Phillips, A., & Wehby, J. (2007). Designing, implementing, and evaluating function-based interventions using a systematic feasible approach. *Behavioral Disorders, 32,* 122-139.

*Lannie, A., & McCurdy, B. (2007). Preventing disruptive behavior in the urban classroom: Effects of the good behavior game on student and teacher behavior. *Education and Treatment of Children, 30,* 85-98.

*Lee, S., Odom, S., & Loftin, R. (2007). Social engagement with peers and stereotypic behavior of children with autism. *Journal of Positive Behavioral interventions, 9,* 67-79.

*Liaupsin, C., Umbreit, J., Ferro, J., Urso, A., & Upreti, G. (2006). Improving academic engagement through systematic, function-based intervention. *Education and Treatment of Children, 29,* 573-591.

*Lo, Y., & Cartledge, G. (2005). FBA and BIP: Increasing the behavior adjustment of African American boys in schools. *Behavioral Disorders, 31,* 147-161.

*Maione, L., & Mirenda, P. (2006). Effects of video modeling and video feedback on peer-directed social language skills of a child with autism. *Journal of Positive Behavioral interventions, 8,* 106-118.

*Marckel, J., Neef, N., & Ferreri, S. (2006). A preliminary analysis of teaching improvisation with the picture exchange communication system to children with autism. *Journal of Applied Behavior Analysis, 39,* 109-115.

Mastropieri, M., & Scruggs, T. (1985-1986). Early intervention for socially withdrawn children. *The Journal of Special Education, 19*, 429-441.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. (2007). Treatment integrity of school-based interventions with children in the Journal of Applied Behavior Analysis 1991-2005. *Journal of Applied Behavior Analysis, 40*, 659-672.

Noell, G.. (2008). Research examining the relationships among consultation process, treatment integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation* (pp. 323-341). New York, NY: Taylor & Francis Group, LLC.

Noell, G., Witt, J., Gilbertson, D., Ranier, D., & Freeland, J. (1997). Increasing teacher intervention implementation in general education settings through consultation and performance feedback. *School Psychology Quarterly, 12*, 77-88.

Peterson, L., Homer, A., & Wonderlich, S. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15*, 477-492.

*Petursdottir, A., McComas, J., & McMaster, K. (2007). The effects of scripted peer tutoring and programming common stimuli on social interactions of a student with

autism spectrum disorder. *Journal of Applied Behavior Analysis, 40*, 353-357.

*Reeve, S., Reeve, K., Townsend, D., & Poulson, C. (2007). Establishing a generalized repertoire of helping behavior in children with autism. *Journal of Applied Behavior Analysis, 40*, 123-136.

*Regan, K., Mastropieri, M., & Scruggs, T. (2005). Promoting expressive writing among students with emotional and behavioral disturbance via dialogue journal. *Behavioral Disorders, 31*, 33-50.

Rhymer, K., Evans-Hampton, T., McCurdy, M., & Watson, T. S. (2002). Effects of varying levels of treatment integrity on toddler aggressive behavior. *Special Services in the Schools, 18*, 75-82.

Roane, H., & Ringdahl, J. (1999). Evaluating treatment challenges with differential reinforcement of alternative behavior. *Journal of Applied Behavior Analysis, 32*, 9-23.

*Robertson, E. J., & Lane, K. (2007). Supporting middle school students with academic and behavioral concerns: A methodological illustration for conducting secondary interventions within three-tiered models of support. *Behavioral Disorders, 33*, 5-22.

Sadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Co..

*Sansosti, F., & Powell-Smith, K. (2006). Using social stories to improve the social behavior of children with Asperger syndrome. *Journal of Positive Behavior Interventions, 8*, 43-57.

*Sansosti, F., & Powell-Smith, K. (2008). Using computer-presented social stories and video models to increase the social communication skills of children with high-functioning autism spectrum disorders. *Journal of Positive Behavioral interventions, 10*, 162-178.

Sasso, G., Conroy, M., Sticher, J.P., & Fox, J. (2001). Slowing down the Bandwagon: The misapplication of functional assessment for students with emotional or behavioral disorders. *Behavioral Disorders, 26*, 282–296.

Scott, T., Mclntyre, J., & Liaupsin, C. (2004). An examination of functional behavior assessment in public school settings: Collaborative teams, experts, and methodology. *Behavioral Disorders, 29*, 384–395.

Shavelson, R. J., & Dempsen-Atwood, N. (1976). Generalizability of measure of teaching behavior. *Review of Educational Research, 46*, 553-611.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE Publications, Inc.

Smith, S., Daunic, A., & Taylor, G. (2007). Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children, 30*, 121-134.

Stahr, B. (2005). *An intervention for children with autism spectrum disorders (ASD) who have food selectivity*. Unpublished master's thesis, Venderbilt University, Tennessee.

*Stahr, B., Cuhing, D., Lane, K., & Fox, J. (2006). Efficacy of a function-based intervention in decreasing off-task behavior exhibited by a student with ADHD. *Journal of Positive Behavioral interventions, 8*, 201-211.

*Staubitz, J., Cartledge, G., Yurick, A., & Lo, Y. (2005). Repeated reading for students with emotional or behavioral disorders: Peer- and trainer-mediated instruction. *Behavioral Disorders, 31*, 51-64.

*Sutherland, K., & Snyder, A. (2007). Effects of reciprocal peer tutoring and self-graphing on reading fluency and classroom behavior of middle school students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 15*, 103-118.

*Tarbox, R., Wallace, M., Penrod, B., & Tarbox, J. (2007). Effects of three-step prompting on compliance with caregiver requests. *Journal of Applied Behavior Analysis, 40*, 703-706.

*Taylor, B., & Hoch, H. (2008). Teaching children with autism to respond to and initiate bids for joint attention. *Journal of Applied Behavior Analysis, 41*, 377-391.

Taylor, J., & Miller, M. (1997). When time out works some of the time: The importance of treatment integrity and functional assessment. *School Psychology Quarterly, 12*, 4-22.

*Trussell, R., Lewis, T., & Stichter, J. (2008). The impact of targeted classroom interventions and function-based behavioral interventions on problem behaviors of students with emotional/behavioral disorders. *Behavioral Disorders, 33*, 153-166.

*Turton, A., Umbreit, J., Liaupsin, C., & Bartley, J. (2007). Function-based intervention for an adolescent with emotional and behavioral disorders in Bermuda: Moving across culture. *Behavioral Disorders, 33*, 23-32.

Umbreit, J., Ferro, J., Liaupsin, C., & Lane, K. (2007). *Functional behavioral assessment and function-based intervention: An effective, practical approach*. Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Underwood, M. A. (2007). *The efficacy of a systematic process for designing function-based interventions for adults in a community setting*. Unpublished doctoral dissertation, University of Arizona, Tucson.

*Vismara, L., & Lyons, G. (2007). Using perseverative interests to elicit joint attention behaviors in young children with autism: Theoretical and clinical implications for understanding motivation. *Journal of Positive Behavioral interventions, 9,* 214-228.

*Volkert, V., Lerman, D., Trosclair, N., Addison, L., & Kodak, T. (2008). An exploratory analysis of task-interspersal procedures while teaching object labels to children with autism. *Journal of Applied Behavior Analysis, 41*, 335-350.

*Walpole, C. W., Roscoe, E., & Dube, W. (2007). Use of a differential observing response to expand restricted stimulus control. *Journal of Applied Behavior Analysis, 40,* 707-712.

Wheeler, J., Baggett, B., Fox, J., & Blevins, L. (2006). Treatment integrity: A review of intervention studies conducted with children with autism. *Focus on Autism and Other Developmental Disabilities, 21*, 45-54.

Wheeler, J., Mayton, M., Carter, S., Chitiyo, M., Menendez, A., & Huang, A. (2009). An assessment of treatment integrity in behavioral intervention studies conducted with persons with mental retardation. *Education and Training in Developmental Disabilities, 44*, 187-195.

Wickstrom, K. (1995). *A study of the relationship among teacher, process and outcome variables with school-based consultation*. Unpublished doctoral dissertation, Louisiana State University, Baton Rouge.

Wilder, D., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis, 39,* 369-373.

*Wilder, D., Saulnier, R., Beavers, G., & Zonneveld, K. (2008). Contingent access to preferred items versus a guided compliance procedure to increase compliance among preschooler. *Education and Treatment of Children, 31,* 297-305.

*Wilder, D., Zonneveld, K., Harris, K., Marcus, A., & Reagan, R. (2007). Further analysis of antecedent interventions on preschoolers' compliance. *Journal of Applied Behavior Analysis, 40,* 535-539.

Wilkinson, L. (2006). Monitoring treatment integrity: An alternative to the 'Consult and hope' strategy in school-based behavioural consultation. *School Psychology International, 27,* 426–438.

Wolery, M. (1994). Procedural fidelity: A reminder of its function. *Journal of Behavioral Education, 4,* 381-386.

*Wood, B. K., Umbreit, J., Liaupsin, C. & Gresham, F. M. (2007). A treatment integrity analysis of function-based intervention. *Education and Treatment of Children, 30,* 105–120.

*Wood, J., & Poulson, C. (2006). The use of scripts toincrease the verbal initiations of children with developmental disabilities to typically developing peers. *Education*

*and Treatment of Children, 29,* 437-457.

Yeaton, W., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156-167.

*Yurick, A., Robinson, P., Cartledge, G., Lo, Y., & Evans, T. (2006). Using peer-mediated repeated readings as a fluency-building activity for urban learners. *Education and Treatment of Children, 29,* 469-509.

*Zaghlawan, H., Ostrosky, M., & Al-Khateeb, J. (2007). Decreasing the inattentive behavior of Jordanian children: A group experiment. *Education and Treatment of Children, 30,* 49-64.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence, status, and change. *American Journal of Evaluation, 30,* 44-61.