Sequential Analyses of Discrete-Trial Teaching

Denys Brand

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy in Psychology, The University of Auckland, 2015.

ABSTRACT

Discrete-trial teaching (DTT) is often used to teach academic and other skills to learners with intellectual disabilities. Traditionally, DTT data have been analysed using whole-sessions methods in which data collected across the duration of the study are aggregated into single-session bins. The aim of this thesis was to analyse DTT using sequential analysis, especially Markov transition matrices (MTM) and probability chains, to uncover any unsuspected sequential dependencies that may be affecting the quality of the teaching. Therapy sessions consisting of DTT were video-recorded in their typical setting (home or school) for 8 therapist-learner dyads (consisting of 8 therapists and 5 children with intellectual disabilities). From the video recordings all components of discrete trials were coded and time-stamped.

The aim of Study 1 was to identify, and then correct, therapists' within-trial treatment integrity errors. Between 110 and 1531 discrete trials per dyad were included. MTM identified treatment integrity errors for all dyads. Errors that were consistent across all dyads included learner self-corrections, response prompt errors, and incorrect application of error correction procedures. With 4 dyads, programme consultants were advised of the errors so that therapist re-training could be conducted. At follow-up, increases in treatment integrity were observed for 3 of the 4 dyads.

Study 2 had three aims. First, to extend the findings from Study 1 and investigate how within-trial error-correction treatment integrity errors affected learner responding on acquisition trials; second, to introduce a method of evaluating between-trial treatment integrity for the error-correction procedures that had a between-trials component; third, to compare the relative effectiveness of concurrent and interspersed task sequencing procedures using extended Markov chain analyses. The results showed that learners were less likely to make a correct acquisition trial response following within-trial error-correction errors than if

the procedure was administered as prescribed; that the between-trials error-correction treatment integrity was less than 60% for all dyads; and that learners may not be receiving the intended benefits of interspersed training. This thesis provided a demonstration of how clinically relevant information that can be obtained when DTT data are analysed on a within-sessions basis.

ACKNOWLEGEMENTS

I would like to thank my supervisors, Dr. Oliver Mudford and Associate Professor Douglas Elliffe for supervising my thesis. Thank you for all of your support, expertise, advice, corrections, proofreading, and most of all your patience with me during the course of this thesis. Thank you Dr. Mudford for introducing me to the world of applied behaviour analysis, and for giving me the opportunity to do research.

I want to thank all of the learners, therapists, consultants and organisations that participated in my research. Without you this project would not have been possible. Thank you to all of the second observers who did the IOA checks, and my team of proof-readers.

I want to thank my family who supported me over the course of my studies. Especially my parents, Kobus and Elaine Brand, for all of your love, support, understanding and counsel. It's been a long journey but in the end it has been worth it.

Lastly, thank you to Alexandria Fraser for all of your support and care, and my friend Pastor Greg Crawford for providing some much needed perspective and guidance during the difficult times.

Table of Contents Page

LIST OF TABLES

LIST OF FIGURES

Chapter I

Early intensive behavioural intervention (EIBI) is an intervention based on the principles of applied behaviour analysis (ABA), and is often used to teach functional and adaptive behaviours to learners with intellectual disabilities such as autism and global developmental delay. The intervention is intensive, and consists of between 20 to 40 hours per week (over the course of several years) of one-on-one therapy between a therapist and a learner (Lovaas, 1987). Discrete-trial teaching (DTT) is often an important part of EIBI. DTT consists of breaking down skills into smaller components and teaching each one of these smaller skills individually (Leaf & McEachin, 1999). Most skills taught during the initial stages of EIBI uses a DTT format.

The estimated costs of EIBI consisting of DTT ranges between $20 000 and $60 000 per learner per year (Chasson, Harris, & Neely, 2007; Jacobson, Mulick, & Green, 1998; Sallows & Graupner, 2005). Consultants and therapists have a professional obligation to provide high quality services to the learners receiving these services, and the parents/caregivers who pay for them, especially when considering both the cost of the intervention, and the time needed to complete it. One way to measure the quality of an intervention is to assess the integrity with which the procedure is administered (i.e., is the treatment being administered in a manner consistent with their design?) (Peterson, Homer, & Wonderlich, 1982). Research has shown that learning can be adversely affected when DTT is not administered as prescribed (e.g., DiGennaro Reed, Reed, Baez, & Maguire, 2011). Therefore, it is important to continuously develop and refine methods that can be used to assess the treatment integrity of DTT (on both a within and between-trial basis) when it is administered as part of a learner's regular teaching programme. The aim in developing such methods is to increase the overall quality of the therapy.

The first chapter of this thesis describes the different sequential analysis techniques which have been used in applied behaviour analysis (ABA) to investigate descriptive data for sequential associations between events. These include conditional probability analysis (Lerman & Iwata, 1993), lag sequential analysis (Sackett, 1979), Yule's Q (Yule & Kendall, 1957), contingency space analysis (Matthews, Shimoff, & Catania, 1987), and Markov probability chains (Gottman & Notarius, 1978). In this chapter the term *events* is used generically to refer to any event or behaviour that is recorded during a descriptive assessment and used in the analyses.

Descriptive Assessments

People engage in observable behaviour which can be measured and consequently studied. Johnston and Pennypacker (2009) described three fundamental dimensions by which behaviour can be measured; temporal locus, temporal extent, and repeatability. In other words, observable behaviour occurs at some point in time for a certain amount of time and repeatedly over time.

Descriptive assessments involve the direct observation and recording of operationally defined events in the client's natural environment (Bijou, Peterson, & Ault, 1968). Data collected during a descriptive assessment can be analysed to investigate the temporal and sequential relationships between events (Fahmie & Hanley, 2008). There are several ways in which a descriptive assessment can be conducted. These include ABC-recordings (Bijou et al., 1968), sequential recordings (Lerman & Iwata, 1993) and scatterplots (Touchette, MacDonald & Langer 1985). The analysis of descriptive data is known as a descriptive analysis.

Descriptive data can be graphed using a variety of different methods and this is an effective way of identifying and demonstrating the important temporal and/or sequential relationships within the data. Examples include cumulative records (Ferster & Skinner, 1957), occurrence graphs, and lag sequential analysis (Sackett, 1979). These methods depict the data within single time units (e.g., seconds, minutes) within a single session (Fahmie & Hanley, 2008).

Data collected during a descriptive assessment cannot be used to establish functional relationships between events (Thompson & Iwata, 2001). Descriptive analyses only provide correlational information about the association between events (Anderson & Long, 2002; Bijou et al., 1968; Rahman, Oliver, & Alderman, 2010; Woods, Borrero, Laud, & Borrero, 2010). For a functional relationship to be established an experimental functional analysis must be conducted (Woods et al., 2010).

Sequential Analysis

The term *sequential analysis* refers to a set of data analysis techniques used to identify sequential associations between events recorded during descriptive assessments. More specifically, these methods evaluate the degree to which the occurrence of one event predicts the occurrence of other events which immediately precedes or follows it (Yoder & Symons, 2010). So whereas a descriptive analysis involves the analysis of data collected during a descriptive assessment, a sequential analysis is a specific type of descriptive analysis.

One of the major goals of conducting a sequential analysis is to determine whether a sequence of events occurs more or less often than expected by chance. For the most part, the datum that is of interest in a sequential analysis is a two-event sequence (Martens, DiGennaro, Reed, Szczech, & Rosenthal, 2008; McComas, Moore, Dahl, Hartman, Hoch, & Symons, 2009). However, it is possible to have data units that involve more than two events. When analysing descriptive data using sequential analysis techniques, it is crucial to have detailed and accurate records of the time and sequence in which events occurred. Sometimes the onset and offset times may also be of importance (Yoder, Short-Meyerson, & Tapp, 2004).

All sequential analysis methods require that that coding be exhaustive (Yoder & Symons, 2010). Exhaustive coding refers to “all relevant units that occurred in an observation session” (Yoder & Symons, 2010, p. 96). In order for exhaustive coding to occur the observer must code events continuously across all units of time (for example every second or every specified interval). All sampling units (seconds/intervals) must be included in the analysis in order to produce interpretable results.

Conditional Probability Analysis. Probability is defined as the likelihood that an event occurs and can be a value ranging between 0 and 1. A probability value tending towards 0 indicates that a particular event is unlikely to occur, and a probability value tending towards 1 indicates that an event is likely to occur.

An unconditional probability is the probability of an event occurring irrespective of other events. Unconditional probabilities are calculated by dividing the number of observations in which an event was recorded by the total number of observations. For example, if an observation period was 20 s and the behaviour of interest occurred for 5 s of the 20 s then the unconditional probability of that behaviour occurring is .25 (5/20).
A conditional probability is defined as the likelihood of an event occurring given the occurrence of other events. A formal definition of a conditional probability is given by the equation:

$$p\ (A|B) = \frac{P\ (A \cap B)}{P\ (B)} \qquad (1)$$

Where *A* and *B* are two different types of events, p (A | B) is the probability of Event A occurring given the occurrence of Event B, and p (A ∩ B) are instances in which both Events A and B occurred.

Thompson and Iwata (2001) collected data across 90 10-s intervals and recorded events of interest which were problem behaviour and staff attention. The results showed that for one participant the problem behaviour occurred during 11 intervals and staff attention followed the problem behaviour in the next interval six times. Therefore, the conditional probability of staff attention given the problem behaviour was .55 (6/11).

Two approaches can be taken when conducting a conditional probability analysis (Martens et al., 2008; Samaha, Vollmer, Borrero, Sloman, Pipkin, & Bourret, 2009). One approach is to compare the conditional probabilities of each sequence. The sequence with the highest conditional probability is assumed to have the greatest sequential association with the event with which it has been coupled (Martens et al., 2008). This type of conditional probability analysis does not make reference to any unconditional probabilities.

Thompson and Iwata (2001) conducted a study in which the conditional probabilities of different social consequences following instances of problem behaviour were compared. The social consequences were attention from staff, escape from ongoing activities, and the presentation of tangible items. The problem behaviours were disruption, self-injurious behaviour (SIB), and aggression. The results showed that staff attention was the most likely social consequence for all forms of problem behaviour, escape from demands was less likely, and access to tangible items the least likely social consequence. In other words, the most likely response from staff members given the occurrence of problem behaviour was attention.

Interpretations from analyses where conditional probabilities have been calculated without reference to the unconditional probabilities can be misleading (Vollmer, Borrero, Wright, Van Camp, & Lalli, 2001). For example, if a conditional probability analysis shows that the conditional probability of escape from demands given the occurrence of tantrum is .70, then in absolute terms it can be concluded that the probability of escape from demands occurring given the occurrence of a tantrum is high because the value is close to 1. However, if the unconditional probability of escape from demands is .80, the conditional probability value of escape from demands given the occurrence of a tantrum is less than the unconditional probability value. Therefore, how the conditional probability is interpreted will change.

In the aforementioned example the conditional probability of escape from demands occurring contingent upon the occurrence of tantrums is less than the probability of escape occurring not contingent upon the occurrence of a tantrum. In other words, escape from demands occurs less often given the occurrence of a tantrum than if the tantrum had not occurred. It appears to be the case that the occurrence of the tantrums inhibits the occurrence of escape. Therefore, a comparison of conditional probability values in isolation without reference to the unconditional probabilities is incomplete since such a comparison may lead to different conclusions regarding the temporal or sequential relationship between two events (Vollmer et al., 2001).

Vollmer et al. (2001) compared the conditional probabilities of three potential reinforcers (attention, escape from demands, and access to tangible items) given the occurrence of problem behaviour. This was done in an attempt to identify either positive, negative, or neutral contingencies between responses or the consequences which follow them. A positive contingency exists if the conditional probability of the problem behaviour given the consequence is higher than the unconditional probability of the consequence. In other words, in a positive contingency the consequence is very likely to follow instances of problem behaviour. A negative contingency exists if the conditional probability of the problem behaviour given the consequence is less than the unconditional probability of the consequence. Negative contingencies are unlikely to maintain behaviour because the probability of the consequence occurring decreases when the problem behaviour occurs. A neutral contingency exists if the conditional and unconditional probabilities are approximately equal (Vollmer et al., 2001).

Lag Sequential Analysis. A method that is used frequently to evaluate temporal and sequential relationships between events is lag sequential analysis (Sackett, 1979). Lag sequential analysis involves calculating a different probability value for each individual lag resulting in dynamic transitional probabilities which changes from lag to lag. A lag is defined as either an event or a unit of time, depending on which is used on the x-axis of the graph. Lag sequential analyses can be time-based or event-based. For a time-based lag sequential analysis, a probability value is calculated for each time unit how the probability of a sequence changes over time is investigated. Event-based lag sequential analysis involves calculating a probability every time there is a transition from one event to the next. The procedure has evolved since its inception and its application is now widespread and includes areas such as psycholinguistics (Jose, 1988), food refusal (Woods et al., 2010), adults and children with intellectual and developmental disabilities (Borrero & Borrero, 2008), and the medical profession (Montague, Xu, Chen, Asan, Burrett, & Chewing, 2011).

Figure 1.1 is an example of a time-based lag sequential analysis (Borrero & Borrero, 2008). The event on y-axis is the target event, and the event occurring at 0 s on the x-axis is the criterion event. In Figure 1.1 the probability of the target event is calculated each second for the 50 s leading up to, and the 50 s following the criterion event. In other words, the analysis tracks changes in the conditional probability of the target event relative to the criterion across all specified lags. The conditional probabilities are called transitional probabilities in a lag sequential analysis, although they are calculated in the same way as in Equation 1. The lags are plotted on the x-axis. Figure 1.1 also contains the unconditional probability of the target event occurring (horizontal line). The unconditional probability is the probability of the target event on average. In the case of Figure 1.1, the unconditional probability is the average probability of the target event occurring during each 1-s interval.

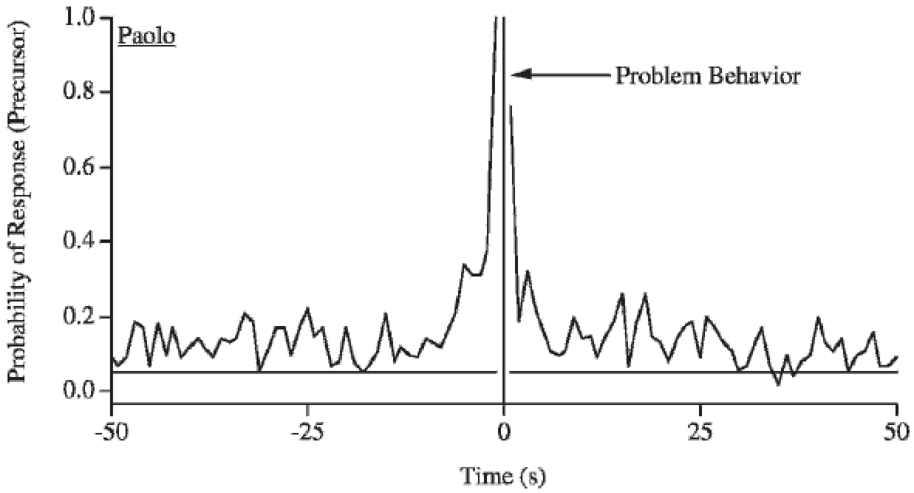

Figure 1.1: Time-based lag sequential analysis showing the transitional probabilities of the precursor behaviour in the 50 s leading up to and following the problem behaviour (Borrero & Borrero, 2008, p. 89).

With clinically important populations such as children and adults with intellectual disabilities, time-based lag sequential analysis has been used more often than event-based lag sequential analysis. One reason for this is that time is an important aspect of the data which can be modelled (as discussed in next paragraph). An event-based lag sequential analysis is time–independent. Examples of where time-based lag sequential analysis has been used to analyse data from clinically important populations includes identifying possible precursors to problem behaviours in two children diagnosed with autism (Borrero & Borrero, 2008), and evaluating the relationship between problem behaviours and environmental events with individuals with intellectual disabilities (Samaha et al., 2009).

A time-based lag sequential analysis is advantageous because it provides a molecular analysis of the data as transitional probabilities can be calculated on a second-to-second basis (Woods et al., 2010). Another advantage of a time-based lag sequential analysis is that it can identify features of the data which can be obscured by a conditional probability analysis. Borrero and Borrero (2008) conducted a study to identify possible precursors to problem behaviours and compared the results from a conditional probability analysis and a time-based lag sequential analysis. Precursor behaviours are defined as when the occurrence of one event reliably predicts the occurrence of another (Smith & Churchill, 2002). For example, if Event A reliably occurs in the moments preceding Event B then it can be concluded that an occurrence of Event A acts as a precursor for occurrence of Event B.

The results from Borrero and Borrero's (2008) conditional probability analysis showed that the problem behaviour reliably occurred given the occurrence of the precursor. However, the conditional probability analysis provided a static probability value based on data gathered across the entire observation period. The analysis obscured an important feature of the data which was how the probability of the precursor behaviour changed in the seconds immediately preceding and following the problem behaviour. A time-based lag sequential analysis was used to investigate changes in the transitional values of the precursor behaviour occurring in the 50 s leading up to and following the problem behaviour. Results from the analysis can be seen in Figure 1.1. The graph shows that the transitional probability of the problem behaviour was highest in the second immediately preceding the problem behaviour and decreases once the problem behaviour occurred. Therefore, a time-based lag sequential analysis provided important information about the temporal relationship between the two behaviours. This information was obscured in the conditional probability analysis (Borrero & Borrero, 2008).

One thing that is not clear from looking at a lag sequential analysis graph is the sample size of either the criterion (denominator in Equation 1) or target behaviour (numerator in Equation 1). The sample size of the criterion behaviour may not be equal at each lag and it is almost certainly the case that the target behaviour will not occur with equal frequency at each lag. However, this information is absent when examining a lag sequential analysis graph. Wilson (2008) attempted to address this issue by plotting a third axis on the graph which displayed the sample size of the criterion at each lag. This graph can be seen in Figure

1.2. As seen in Figure 1.2 the sample size of the criterion varies at each lag. It is suggested here that future analyses involving lag sequential analysis graphs should include a sample size axis as it provides important information.

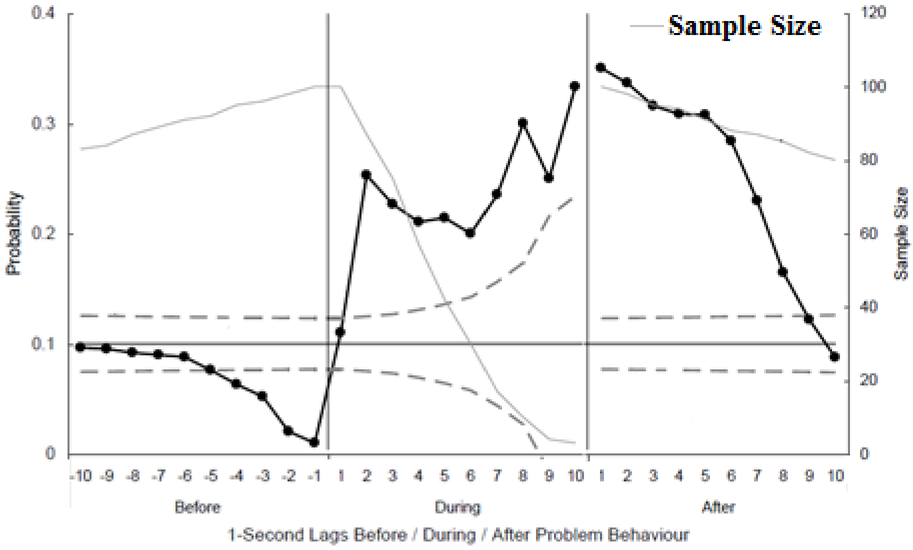

Figure 1.2: Time-based lag sequential analysis showing the sample size of the criterion for each lag on the right vertical axis (Wilson, 2008).

Lag sequential analysis graphs will often contain 95% or 99% confidence intervals. If transitional probabilities lie outside the confidence intervals the target event is occurring significantly more or less than what is expected (Sackett, 1979). *Z*-scores are often used to calculate the values of the confidence intervals (Alison & Liker, 1982). Gottman and Roy (1990) recommended a sample size of at least 20 when calculating confidence intervals using *z*-scores. Due to the large number of calculations involved in conducting a lag sequential analysis the technique is susceptible to Type-I errors (false positives), especially with an alpha value equal to .05. On the other hand, when calculating 99% confidence interval with an alpha value .01 the probability of committing a Type-II error (false negatives) increases.

Sackett (1979) provided some guidelines for interpreting statistically significant data points from a lag sequential analysis. Sackett suggested that if transitional probabilities are significant at specific reoccurring lags, or are grouped together then more meaningful conclusions can be drawn. For example, if probabilities are significant at specific reoccurring lags then that might suggest a cycle. Therefore, when interpreting statistically significant lags, the overall context of the analysis must be taken into consideration (Sackett, 1979).

A variation of the lag sequential analysis technique is the time-window sequential analysis (Yoder & Tapp, 2004). A time-window sequential analysis involves recording whether an event occurs within a certain time from the onset of a preceding event. The difference between a time-based lag sequential analysis and a time-window sequential analysis is in terms of the research question being asked. In a time-based lag sequential analysis the analysis involves calculating a probability value at each lag since the occurrence of the criterion (for example, what is the probability of the target behaviour in the second immediately following the criterion). A time-window analysis calculates the probability of the target event occurring at all during a specified period of time following the occurrence of the criterion event. For example, Chen, McComas, Hartman and Symons (2011) used a time-window sequential analysis to investigate whether peer rejection acts occurred within 15 s after the onset of physical aggression. It should be clear then that a time-window sequential analysis is less precise than a time-based lag sequential analysis (Yoder & Tapp, 2004).

Yule's Q. Another method for analysing sequential associations within descriptive data is Yule's Q (Yule & Kendall, 1957). Yule's Q values are "effect size metrics for categorical data" (Yoder & Symons, 2010, p. 126) and can be a number between -1 and 1. The numbers are interpreted in a manner similar to a Pearson correlation (Bonett & Price, 2007). Established effect size benchmarks for large, moderate, and small effects are .60, .43, and .20 respectively (Lloyd, Kennedy, & Yoder, 2013; Yoder & Symons, 2010). Positive values indicate a positive association and show that the observed frequency of the sequence is larger than what was expected by chance. Negative values indicate a negative association and show that the observed frequency of the sequence is less than what is expected by chance. A Yule's Q value of zero indicates no association between two events and the observed frequency of the sequence is at chance level (Yoder & Symons, 2010).

Yule's Q values are calculated by organising descriptive data into 2 x 2 contingency tables such as the one in Figure 1.3. As seen in Figure 1.3, the rows indicate whether *A* was the first event and the columns indicate whether *B* was the second event. Figure 1.3 can be used when considering both event and time lags (Bakeman, 2001).

As seen in Figure 1.3, cell *a* of the table indicates a sequence of events where *A* occurred first and *B* occurred second, cell *b* indicates where *A* occurred first and *B* did not occur second, cell *c* tallies the number of times *A* was not the first event but *B* was the second event, and cell *d* where *A* was not the first event and *B* was not the second event (Bakeman, 2001).

Lag 0 / Event 1 | Lag 1 / Event 2

	B	~B	Row Total
A	*a*	*b*	
~A	*c*	*d*	
Column Total			Total

Figure 1.3: 2 x 2 contingency table used in a Yule's Q analysis (Bakeman, 2001).

Once the data have been categorised into the cells of the contingency table the association between successive events can be quantified One such method is to calculate odds ratios. Given Figure 1.3 the odds ratios are given by:

$$odds\ ratio = \frac{a * b}{b * c} \qquad (2)$$

a, b , c and d are the four cells of the contingency table. Odds ratios can vary between zero and infinity. When the odds ratio is equal to one then there is no association between A and B. When the odds ratio is greater than one there is a positive association between A and B, and when the odds ratio is less than one there is a negative association between A and B (Yoder et al., 2004). However, odds ratios can become difficult to interpret because their range is in principle infinite (Bonett & Price, 2007). A Yule's Q statistic transforms the odds ratio to a number between -1 and 1 which makes it easier to interpret. The equation for Yule's Q is:

$$Yule's\ Q = \frac{(a*d) - (b*c)}{(a*d) + (b*c)} \qquad (3)$$

In order to produce interpretable Yule's Q values the expected frequency of each cell in the contingency table must be at least 5. Contingency tables with expected values of less than 5 are known as sparse tables and are not fit for analysis (Yoder & Symons, 2010). Expected cell frequencies can be calculated by multiplying the row and column totals and dividing by the overall total.

The Yule's Q statistic is sensitive to changes in cell frequencies. In other words, when a cell frequency increases by one it can produce a large change in the Yule's Q value. In cases where low cell counts are observed results should be interpreted with caution (Bakeman, 2001). Another problem with Yule's Q is that it will always produce a value of 1 or -1 if one of the cell counts is equal to zero.

The Yule's Q statistic is starting to receive recognition in the applied literature. Oliver, Woodcock, and Humphreys (2009) conducted a study to investigate temper outbursts in individuals with Prader-Willi syndrome. Two time-based lag sequential analyses were conducted investigating sequences in the three minutes preceding and following the criterion event. Yule's Q values were calculated in order to see if the conditional probabilities were greater than what was expected. The criterion for an association was a Yule's Q value of .40

which corresponds to an odds ratio value of 2.33 (2.33 times more likely that the conditional probability was greater than expected). Using the Yule's Q statistic, several meaningful sequences were extracted. For example for one participant the criterion event was arguing. The events preceding arguing were frowning (Yule's Q = .80), and asking repetitive questions (Yule's Q = .60), and the event most likely to follow arguing was crying (Yule's Q = .90) (Oliver et al., 2009).

Contingency Space Analysis. A contingency space analysis, "displays relations among events in terms of the probability of one event given or not given another" (Matthews et al., 1987, p. 71). Figure 1.4 is a generic display of a contingency space analysis. As seen in Figure 1.4, the y-axis represents the probability of Event Y occurring given the occurrence of Event X, and the x-axis represents the probability of Event Y occurring given the non-occurrence of Event X (Matthews et al., 1987). Figure 1.4 also contains a diagonal line called the *unity diagonal* (Martens et al., 2008). The unity diagonal represents instances where the conditional probabilities are equal. If a data point falls above the unity diagonal then Event *Y* is more likely to occur given the presence of Event X, and if a data point falls beneath the unity diagonal then Event Y is more likely to occur given the absence of Event X. A data point falling close to the unity diagonal indicates that Event Y is just as likely to occur given the occurrence or non-occurrence of Event X (Martens et al., 2008). A value close to 1.0 suggests a rich reinforcement schedule because almost every occurrence of Event Y is followed by an occurrence of Event X. A value tending towards zero would be indicative of a lean reinforcement schedule because it is unlikely that Event Y is followed by Event X (Event Y is followed by the non-occurrence of Event X).

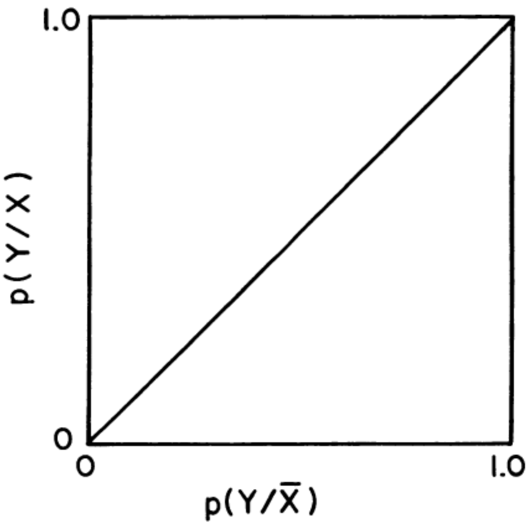

Figure 1.4: Contingency space analysis (Matthews et al., 1987, p. 71).

Figure 1.5 is an example of where a contingency space analysis has been used to graph some hypothetical data (Martens et al., 2008, p. 74). Figure 1.5 shows the conditional probabilities of a particular consequence (attention and escape) given both the occurrence and non-occurrence of a particular antecedent behaviour. As seen in Figure 1.5, the conditional probability of attention given the occurrence of the target behaviour is equal to .73, and the probability of attention given the absence of the target behaviour is approximately .10. The conditional probability of escape is equal (.11) regardless of whether the target behaviour occurred.

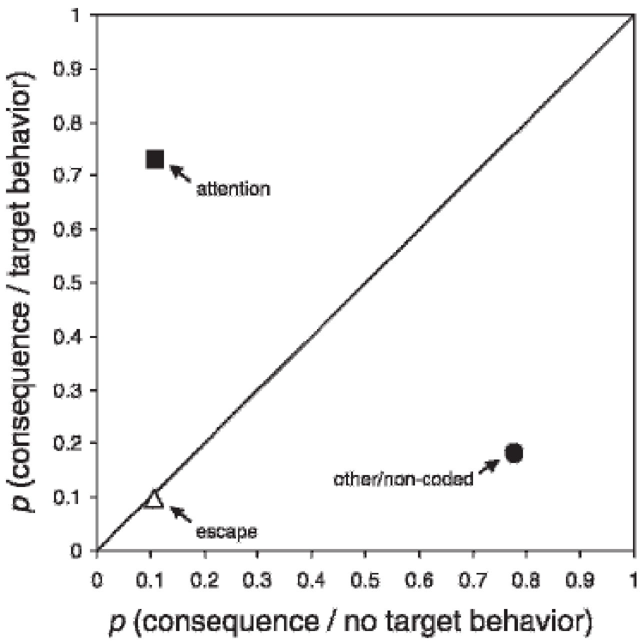

Figure 1.5: Contingency space analysis (Martens et al., 2008, p. 74).

Markov Probability Chains. Markov probability chains can also be used to model sequences within descriptive data. The analysis involves recording events of interest and constructing a transition matrix containing all possible transitions between all recorded events. The transition probabilities form the basis of a Markov chain analysis (Berchtold & Sackett, 2002).

The use of Markov probability chains to analyse sequences is best illustrated with an example. Table 1.1 is a frequency table containing some hypothetical data gathered during a descriptive assessment. These events are labelled A, B, and C. Table 1.1 also contains the unconditional probabilities of these events and were calculated by dividing the number of times each event occurred by the total number of events (Gottman & Notarius, 1978).

Table 1.1

Frequency table and unconditional probabilities for hypothetical data gathered during a descriptive assessment

	A	B	C	Sum
Frequency	115	122	87	324
Probability	.35	.38	.27	1.00

The next step in the analysis is to calculate the frequencies of all possible transitions between all events. This is done by constructing a transition frequency table (Table 1.2). The values in the frequency table are converted into probabilities by dividing the frequency count of a particular cell in the table by the total frequency count of the row. For example, the transitional probability of the sequence AB is .47 (54/115). Table 1.3 displays the transitional probabilities for the hypothetical data set. The right hand column of Table 1.3 is the sum of all the probabilities for a particular row and should add up to 1.0. The transitional probabilities can also be presented as a transition diagram (Figure 1.6).

Table 1.2
Frequency table of possible transitions between all possible events for the hypothetical data set

First Event	Second Event			
	A	B	C	Sum
A	29	54	32	115
B	41	49	22	122
C	33	33	21	87
				324

Table 1.3

Transition data from Table 1.2 expressed as probabilities

First Event	Second Event			
	A	B	C	Sum
A	.25	.47	.28	1.00
B	.34	.48	.18	1.00
C	.38	.38	.24	1.00

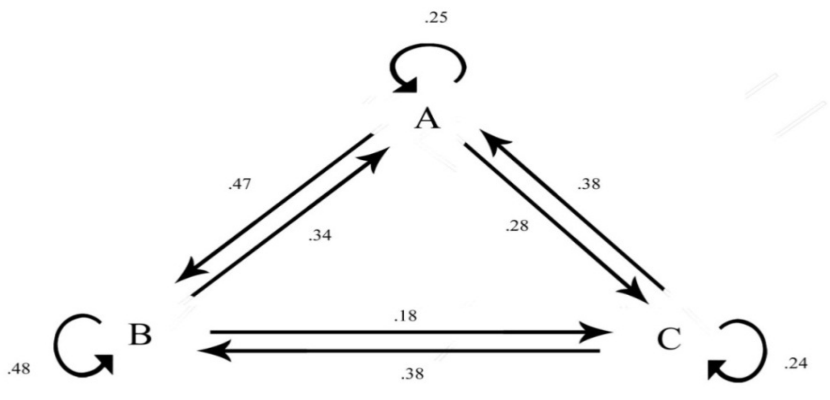

Figure 1.6: Transition diagram of transitional probabilities as seen in Table 1.3.

The transitional probabilities are used to calculate the probabilities of event sequences. Using the hypothetical data a sequence of interest may be ABC. In this example, we are interested in the probability of C given the transition from A to B and then B to C. In this case the probability of the sequence ABC is .08. This value is calculated by multiplying the transitional probability of AB (.47) with the transitional probability of BC (.18).

Markov probability chains have a stochastic property, meaning that it does not matter how a particular event came to be. The only thing that matters is that an event occurred and that there is set of probabilities that it will be followed by another event. In other words the system has no memory (Weingart, Prietula, Hyder, & Genovese, 1999). Consider the transition probabilities in Table 1.3. If B occurred then the probability of transitioning to either A, B, or C is as stated in Table 1.3. It does not matter how B came to be, the only thing that matters is that B has occurred and there now exists a set of probabilities that the next event will be either A, B, or C (Gottman & Notarius, 1978).

The transitional probabilities can also be used to construct a Markov probability tree which contains the probabilities of all possible sequences of a certain order. The order of the chain refers to the number of past events. For example, a Markov chain of order one means that the probability of the next event is determined by only the event immediately preceding it. A Markov chain of order two means that the probability of the next event is determined by the previous two behaviours or events and so forth (Gottman & Notarius, 1978). Figure 1.7 is a Markov probability tree showing all possible outcomes for a three-event sequence (data were taken from Table 1.3). As seen in Figure 1.7, there are nine possible outcomes. The probability values for each behavioural sequence are obtained by multiplying the probability values along each step in the chain. As seen in Figure 1.7, the most likely sequence is ABB ($p = .23$) and the least likely sequences are ACC and AAA ($p = .06$).

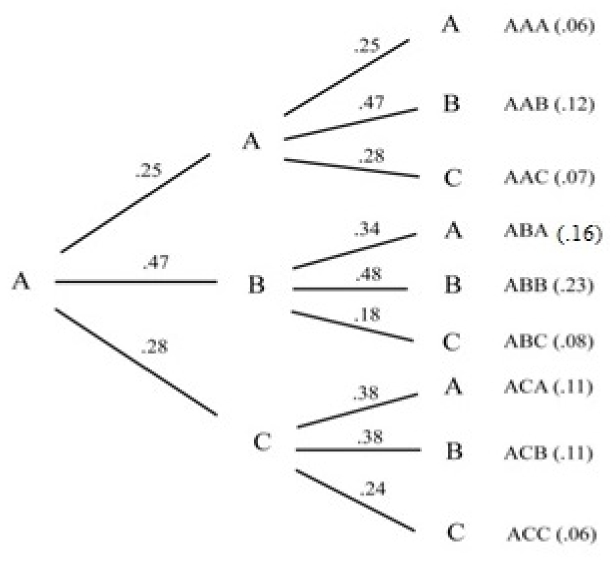

Figure 1.7: Markov probability tree for the hypothetical data using the transitional probabilities from Table 1.3.

Markov probability chains have also been used in a different manner from the method just described. Some areas of research, such as primatology, (e.g., Altman, 1965; Berchtold & Sackett, 2002), and certain statistical packages use Markov probability chains to calculate joint probabilities. Joint probabilities are calculated by dividing each cell in the frequency table by the total number of events in the frequency table resulting in the conditional probability of that sequence occurring. Using the data in Table 1.2, the conditional probability of the sequence AB is equal to .17 (54/324).09 (29/324). Expected values for each cell in the frequency table are calculated by multiplying the row and column totals and dividing that number by the total number of behaviours. For the sequence AB the expected value is given by (115 x 103) / 324. The expected values are then converted into the unconditional probabilities of the sequence occurring. The unconditional probability for the sequence AB is equal to .11 (36.56/324). Therefore, the conditional probability of the sequence AB is slightly less than the unconditional probability. Chi-square tests can be used to test for statistically significant differences between the conditional and unconditional probabilities of the sequences. The difference between the joint probability method and the more traditional method is that the joint probability method calculates the probability of sequences occurring by dividing the number of times a sequence occurred by total number of two-event sequences. The more traditional method only provides the probabilities of events following specific triggering events (Berchtold & Sackett, 2002).

Markov probability trees provide an exploratory approach to descriptive data given that it contains all the possible sequential outcomes for a data set no matter the length of the chain. However, an exploratory approach has the potential to result in large amounts of data outputs with difficult to interpret results, especially as the length of the chains increase (Sackett, 1979).

The fact that the analysis does calculate all possible sequences can be advantageous. An exploratory approach may reveal sequences occurring with probabilities that are higher or lower than expected and can reveal sequences that were not anticipated. From an applied perspective the discovery of these unexpected sequences may have clinical value. Perhaps certain sequential dependencies can impede learning, whereas others can facilitate it. For example, a sequential dependency that could impede learning would be one which compromises the integrity of an intervention. Research has shown a relationship between compromised levels of treatment integrity and impaired learner performance (e.g., Koegel, Russo, & Rincover, 1977). Once these problematic sequential dependencies have been identified they can be targeted for therapist retraining. A method such as Markov transition matrices is useful because the transitional probabilities between all components of a procedure can be calculated simultaneously, an especially useful feature when analysing data from interventions consisting of multiple components (e.g., discrete-trial teaching). It may be that a Markov chain analysis provides interesting initial results which can be explored further with other sequential analysis techniques.

Another advantage of Markov probability chains is that they can be used to analyse sequences with more than one transition without becoming too complicated mathematically. All of the sequential analysis methods discussed in this chapter focus primarily on the sequential association between two events. Given that events occur in a dynamic environment with constant transitions between events, this feature of Markov probability chains is advantageous and has potential clinical applications. It is possible that a particular type of learner response (problematic or otherwise) may be correlated not just with the event which immediately preceded the behaviour, but by the two or three which preceded it. For example, problem behaviour (such as learner self-harm) can be reliably predicted by the occurrence of stereotypy (i.e. hand flapping), which is followed by the occurrence of inappropriate vocalisations such as swearing. In this case, the therapist or caregiver can intervene as soon as the stereotypy occurs and does not have to wait for the inappropriate vocalisations to occur. The occurrence of stereotypy acts as a reliable predictor of the self-harm, even though it not usually the behaviour that immediately follows stereotypy.

However, the use of Markov transition matrices and chains in ABA has been sparse. One of the few examples is Stuart (1971). Stuart used Markov chains to evaluate the effectiveness of a behavioural contract for a 16-year old female exhibiting a number of problem behaviours such as promiscuity, exhibitionism, and drug abuse. A behavioural contract involves setting up contingencies for obtaining reinforcement. It involves setting out conditions which must be adhered to in order to gain the reinforcer, and specifies the reinforcer to be received for adhering to these conditions. The behavioural contract was set up to increase prosocial behaviour.

The results from the Stuart (1971) study can be seen in Figure 1.8. As seen in Figure 1.8 the behavioural contract appeared to be effective in increasing compliance to the conditions set out in the contract. Figure 1.8 shows that the probability of transitioning from a C+ (complying with conditions in contract) to another C+ increased over the three observation periods from .60 to .85 to .96. As a consequence the probability that a sequence of three C+'s occurring increased from .22 to .61 to .89. Using a Markov probability chain analysis was an effective method of evaluating the effectiveness of the behavioural contract (Stuart, 1971).

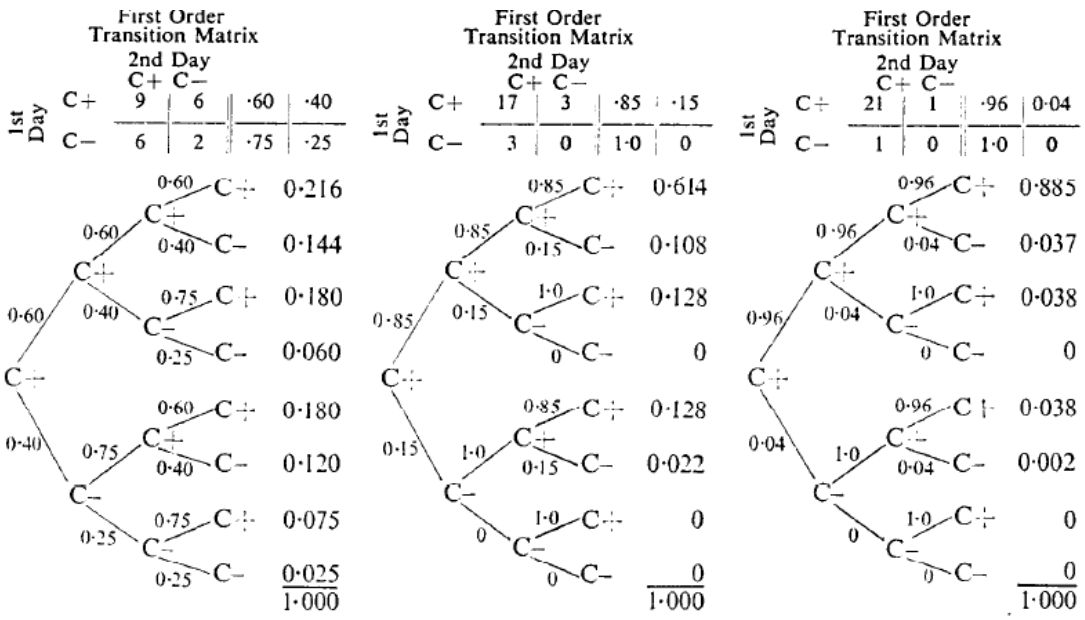

Figure 1.8: Markov probability chains analysis of the effectiveness of a behavioural contract (Stuart, 1973, p. 8).

Summary

The aim of this chapter was to describe the different sequential analysis methods which are used to analyse descriptive data in ABA. Researchers and analysts have several different sequential analysis methods available to them. The method used to analyse the data should be determined by the type of data collected, the important relations and associations within the data which the researcher wants to display, and the research question of interest (Fahmie & Haney, 2008). It is not necessarily a question of which sequential analysis method is the best, but rather which sequential analysis method is best suited to display the important sequential associations within the data. Whereas the clinical utility of other sequential analysis methods such as conditional probability analyses and time-based lag sequential analysis have been assessed and established, the clinical utility of Markov probability chains has not yet been assessed. One of the aims of the thesis is to provide such an assessment.

Chapter II

Discrete-Trial Teaching

Discrete-trial teaching (DTT) is a teaching strategy frequently used in ABA to teach functional skills to individuals (predominantly children) with intellectual disabilities (Smith, 2001). DTT is often a prominent feature of early intensive behavioural intervention (EIBI) programmes, and the effectiveness of EIBI programmes consisting of DTT has been demonstrated in several studies (e.g., Anderson, Taras, & O'Malley, 1996, Lovaas, 1987, McEachin, Smith, & Lovaas, 1993, Vismara & Rogers, 2010). Over the course of an EIBI programme a therapist can administer thousands of these discrete-trials, resulting in thousands of transitions between trials, and tens of thousands of transitions between all the components of the procedure. Some skills that have been taught to individuals with intellectual disabilities using DTT are the use of mands (Jennett, Harris, & Delmolino, 2008), learning nouns (Holding, Bray, & Kehle, 2001), language acquisition (Carr & Dores, 1981), prepositions (McGee, Krantz, & McClannahan, 1985), sign language (Carr, Kologinsky, & Leff-Simon, 1987), verbal play skills (Coe, Matson, Fee, Manikam, & Linarello, 1990), spontaneous use of colour adjectives (Miranda-Linne & Melin, 1992), generalized imitation skills (Young, Krantz, & McClannahan, 1994), spontaneous responding (Jones, Feeley, & Takacs, 2007), and identifying occupations from picture cards (Dogan & Tekin-Iftar, 2002).

DTT consists of five distinct components. These components are the discriminative stimulus (S^D), a prompt if necessary, a response (correct, incorrect or no response), a consequence determined by the type of response, and an intertrial interval (ITI). Discrete teaching trials have a distinct beginning and end. The presentation of the S^D constitutes the beginning of the trial and the start of the ITI the end (Smith, 2001; Leaf & McEachin, 1999). In this thesis the recipients of DTT are referred to as *learners*, and the teachers as *therapists*.

Traditionally, DTT takes place in a structured and controlled environment consisting of one-to-one teaching sessions between therapists and learners (Delprato, 2001). DTT involves breaking down particular skills into smaller parts, and teaching each one of these smaller skills individually (Leaf & McEachin, 1999). The goal of DTT is to increase unprompted correct learner responses and decrease incorrect responses (Smith, 2001). DTT is conceptually consistent with several operant conditioning principles. . These principles include the reinforcement of correct learner responses to promote subsequent correct responses, and shaping by successive approximations (LeBlanc, Ricciardi, & Luiselli, 2005; Leaf & McEachin, 1999; Lovaas, 2003; Smith, 2001).

This chapter consists of two parts. The first will provide a description of the DTT components. The second will outline recommendations for how each DTT component should be selected and applied. Therefore, this chapter contains both a descriptive and prescriptive account of the different DTT components.

Discrete-Trial Teaching Components: Descriptive

Discriminative Stimulus (S^D)

The S^D is an antecedent stimulus during which the therapist provides the teaching instructions to the learner and is the distinct beginning of any discrete trial (Smith, 2001). Teaching instructions include asking the learner questions such as “What is that?” or providing instructions such as “Point to red.” Trials administered during DTT can be classified as either acquisition, mastered or probe trials.

Acquisition trials are trials that contain stimuli for which the learner has not reached the mastery criterion. Acquisition stimuli are also called targets. Acquisition trials often require various levels of prompting to assist the learner in emitting a correct response (Smith, 2001).

Mastered trials are continued training after the mastery criterion has been met. A mastery criterion is a predetermined criterion used to determine whether a learner has mastered a target skill. The mastery criterion is determined by either the consultant who designed the learner's DTT programme, the therapist or the researcher. The ongoing teaching of mastered skills is known as maintenance and trials administered during maintenance are referred to as maintenance trials.

Several different types of mastery criteria exist. These include number of correct unprompted acquisition trial responses across sessions (Carr et al., 1987; Miranda-Linne & Melin, 1992), across consecutive trials (Dogan & Tekin-Iftar, 2002; Williams, Koegel, & Egel, 1981), across days (Jones et al., 2007) or across therapists (Summers & Szatmari, 2009). Some researchers use daily probe sessions to determine whether the learner has mastered a target skill (Leaf, Sheldon, & Sherman, 2010). How DTT data are presented and analysed will be discussed in Chapter III, but when data are presented as percentage correct all of the DTT research cited in this chapter requires a criterion of at least 80% correct unprompted learner responses across a certain number of trials or sessions. Najdowski et al. (2009) stated that it is important to have a mastery criterion that is not too lenient because it may lead to a premature conclusion that the learner has mastered a skill when in reality they have not.

The last type of DTT trial is a probe trial. Probe trials serve various purposes. Before a new stimulus or set of stimuli is taught, baseline probe trials are administered to evaluate learner performance. Target stimuli to be taught are determined by how the learner responded during the baseline probe trials (Leaf et al., 2010; Lerman, Dittlinger, Fentress, & Lanagan, 2011). Probe trials can also be used to determine if a learner has met the mastery criterion. Probe trials are also administered once a target skill has been mastered in order to evaluate ongoing learner performance for that particular target skill. If the learner performs poorly on the probe trials it would suggest that mastery of the target skill has not been maintained and that it may need to be re-taught (Leaf et al., 2010).

Learner Response

Learner behaviour following the S^D is called a response (Lovaas, 2003). The learner can emit either a correct response, an incorrect response, or no response. Correct responses are defined as responses which fulfil the requirements stated in the S^D. Incorrect responses are defined as responses which do not meet the requirements as stated in the S^D. No-responses are often considered to be incorrect responses (Carr et al., 1987; Cummings & Carr, 2009; Dogan & Tekin-Iftar, 2003; Schuster et al., 1992).

Consequence

Consequences are provided by the therapist contingent on the response that was emitted by the learner. In this thesis consequences for correct responses will be referred to as *correct consequences* and consequences for incorrect responses will be referred to as *error consequences*. Correct consequences include verbal praise, tokens, access to preferred items such as toys, edible items, and activities. Error consequences include removing all stimuli from the teaching area and withholding reinforcement (Lovaas, 2003), corrective feedback (Leaf & McEachin, 1999; Leaf et al., 2010) or the appropriate error-correction procedure. When a non-response is considered an incorrect response they are followed by the same consequences as incorrect responses.

Intertrial Interval

An ITI is a period of time that elapses between the end of one trial and the beginning of the next (Anderson et al., 1996). The ITI is what makes one trial discrete from any other trial. The ITI also allows for a period of time during which the therapist can record data.

Prompts

Prompts are additional antecedent stimuli provided by the therapist (if necessary) to assist the learner to emit a correct response in the presence of the S^D (Cooper, Heron, & Heward, 2007). Prompts can be divided into two types; stimulus prompts and response prompts. Stimulus prompts involve manipulations of the antecedents and are directly associated with the training stimulus itself. Examples of stimulus prompts include exaggerating relevant features of the training stimulus and positioning stimuli in such a way as to identify the correct one. This section will focus mainly on response prompts.

Response Prompts. Response prompts operate on the learner-response portion of a trial (Cooper et al., 2007). Several different types of response prompts exist. These include gestural, verbal, model, physical guidance, and shadow prompts (Fox, 1982; Lovaas, 2003). Gestural prompts involve a movement or gesture on the part of the therapist to assist the learner to emit a correct response (Anderson et al., 1996). For example, if the learner has to identify the correct stimulus when presented with several stimuli, a gestural prompt can include the therapist pointing towards the correct stimulus (Fox, 1982).

When a therapist uses words and the sounds of letters in order to evoke a correct learner response, a verbal prompt is in operation. Verbal prompts include questions, stating rules, providing instructions, or just providing the first letter of the required stimulus (Fox, 1982). Model prompts consist of the therapist providing a visual demonstration of the required response and the learner imitating the actions of the therapist (Anderson et al., 1996). For example, if a therapist teaches a learner to pick up a toy and put it in a box, the model prompt would consist of the therapist picking up the toy and putting it in the box while the learner is observing (Anderson et al., 1996). However, it is essential that the learner has already acquired general imitation skills. If not, then generalized imitation skills must be taught first (Ghezzi, 2007).

Physical guidance prompts involve the therapist physically guiding the movements of the learner (Fox, 1982). An example of a physical guidance prompt is hand-over-hand assistance. When administrating a discrete trial the therapist will take the hand of the learner and place it on the correct stimulus. A shadow prompt involves the therapist shadowing the movements of the learner (Fox, 1982). For example if a learner was required to touch a red card the hand of the therapist will shadow the movements of the learner's hand (without making physical contact) until the learner has touched the red colour card. As soon as the hand of the therapist makes contact with that of the learner the shadow prompt becomes a physical guidance prompt.

Prompt Hierarchies, Prompt Fading and Prompt Fading Procedures. Learners can become prompt-dependent. Prompt-dependent learners have come to rely on the presentation of the prompt to make a response (Fox, 1982; Lovaas, 2003). To minimize the likelihood of the learner becoming dependent on the prompt, prompts must be faded by gradually reducing the amount of prompting, a process known as prompt fading (Anderson et al., 1996; Fox, 1982; Smith, 2001). Prompt fading must occur so that there can be a transfer of stimulus control from the prompt to the S^D, and is required even if learners are not prompt dependent. Prompt fading is part of any effective DTT programme that includes prompts (Anderson et al., 1996; Lovaas, 2003; Smith, 2001).

In order to have a successful prompt fading procedure the establishment of a prompt hierarchy is necessary (Lovaas, 2003). Establishing a prompt hierarchy involves ordering prompts in order of intrusiveness. The term intrusiveness refers to the amount and intensity of assistance that is required for the learner to respond correctly (Lovaas, 2003). For example, a full physical guidance prompt is a more intrusive type of prompt than verbal or gestural prompts.

There are four prompt fading procedures (Anderson et al., 1996; Cooper et al., 2007). These procedures are most-to-least prompting, least-to-most prompting, graduated guidance (Fox, 1982), and time-delay procedures (Heckaman, Alber, Hooper, & Heward, 1998; Schuster, Griffen, & Wolery, 1992). Most-to-least prompting involves starting with the most intrusive type of prompt and progressing to less intrusive prompts. If the therapist is using physical guidance, gestural and verbal prompts, a most-to-least prompt hierarchy involving these prompts would be physical guidance-gestural-verbal. The therapist would progress from the physical guidance before fading to a gestural prompt and finishing with a verbal prompt. The verbal prompt would then be faded until behaviour was under the stimulus control of the S^D alone (Anderson et al., 1996).

Least-to-most prompting starts with the least intrusive type of response prompt and gradually progresses to more intrusive prompts. A least-to-most prompt hierarchy involving physical guidance, verbal and gestural prompts would be verbal-gestural-physical guidance. Initially, the learner has the opportunity to respond independently when using least-to-most prompting. The therapist will begin by providing a verbal prompt, progress to a gestural prompt before providing a physical guidance prompt (Anderson et al., 1996).

Graduated guidance involves the gradual fading of physical guidance (Anderson et al, 1996; Fox, 1982). Graduated guidance consists of three parts; full graduated guidance, partial graduated guidance, and shadowing (Fox, 1982). The therapist will begin by applying full graduated guidance. An example of full graduated guidance would be if the therapist placed their hand on the learner's arm and guided the movements of the learner until the response has been completed. The therapist then progresses to partial graduated guidance. The therapist will no longer have their hand fully placed on the arm of the learner but only place their index finger on the learner's arm. Once the learner is responding sufficiently during the partial graduated guidance the therapist will move on to shadowing. The therapist will shadow the movements of the learner without making physical contact with the learner. If the learner does not complete the response during the shadowing phase the therapist will return to partial graduated guidance (Fox, 1982).

Prompt fading in time-delay procedures involves fading the prompt by increasing the time between the S^D and the delivery of the prompt (Anderson et al., 1996). Time-delay procedures fall into two categories; constant and progressive (Cooper et al., 2007). A constant time delay procedure involves a constant time delay between the presentation of the S^D and presentation of the controlling prompt (Schuster et al., 1992). Progressive time-delay procedures involve progressively increasing the time between the presentation of the S^D and the presentation of the controlling prompt (Heckaman et al., 1998).

Other effective prompting strategies include simultaneous prompting and no-no prompting. Simultaneous prompting is a prompting procedure in which errors are minimized by implementing what is known as a controlling prompt (Leaf et al., 2010). The controlling prompt is presented as soon as the S^D has been presented. The controlling prompt assures that the learner emits a correct response on every trial. As a consequence the learner never makes an incorrect response and will access the reinforcer on every trial (Leaf et al., 2010; Wolery, Holcombe, Werts, & Cipolloni, 1993). Therefore, learners do not have the opportunity to respond independently when a simultaneous prompting procedure is being used. Simultaneous prompting has been used to teach discrimination skills (Leaf et al., 2010), word recognition (Schuster et al., 1992), and to identify different occupations from a series of picture cards (Dogan & Tekin-Iftar, 2002).

No-no prompting is a response-prompting procedure in which the learner is allowed to respond independently following an incorrect response (Leaf et al., 2010). If the learner emits an incorrect response then the therapist provides corrective feedback. Once the corrective feedback has been provided the trial is repeated. The repeated trial is called a remedial trial and consists of the same stimuli that were presented in the previous incorrect trial. If a learner emits another incorrect response during the remedial trial then the therapist will administer a controlling prompt (as in the simultaneous prompting procedure) in order to assure that a correct response follows two incorrect responses (Leaf et al., 2010).

Discrete-Trial Teaching Components: Prescriptive

The following section will discuss guidelines for the effective administration of the different DTT components.

Discriminative Stimulus (S^D)

Guidelines for the effective administration of S^D include that the S^D be discriminable, appropriate to the task being performed, uninterrupted, and it is critical that the learner attends to the S^D (Anderson et al., 1996; Koegel et al., 1977; Leaf & McEachin, 1999; Lovaas, 2003). If the learner is not attending to the S^D when it is administered it becomes less likely that they will discriminate the relationship between the S^D and the correct response. It is important to keep instructional phrases short and simple and to use as few words as possible (Anderson et al., 1996) so that it is clear to the learner what behaviour is required for them to access the correct consequence. For example, if the therapist wants the learner to touch a picture of a ball a short instructional phrase such as “Touch ball” is preferable to a longer phrase such as “Can you please be a good learner and touch the picture of the ball?” Lovaas (2003) recommended that the wording of the S^D should be consistent across trials in order not to confuse the learner. However, the wording of S^D may also be varied to promote generalisation. Anderson et al. (1996) stated that S^D should be presented only once per trial so that learner behaviour can come under the stimulus control of the S^D.

Learner Response

Learners should emit a response within a certain period of time from the presentation of the S^D. Different times have been recommended. Lovaas (2003) recommended that a response should be emitted within 1 – 3 s. Anderson et al. (1996) and Leaf and McEachin (1999), recommended that responses should occur within 3 – 5 s. If no response is emitted within the recommended time then the trial is scored as a non-response which is often considered an incorrect response.

Lovaas (2003) stated that it is important to keep response requirements consistent across therapists. For example, if the response requirement is for the learner to respond within 3 s of the S^D, the 3-s time-limit should be adhered to by all therapists. If the response requirement is not consistent across therapists the effectiveness of the correct consequence may be diminished. The behaviour the learner emits following a S^D should follow the programmed definition of a correct response if a correct consequence is to be provided (Lovaas, 2003).

It is important that the learner is able to perform the target behaviour (Ghezzi, 2007). It may be necessary to shape behaviour to ensure that the learner is able to perform the correct response and fulfill the requirements of the S^D before training of the S^D begins (Koegel et al., 1977; Leaf & McEachin, 1999). Koegel et al. (1977) advised that “Each reinforced response should be at least as good as the last one,” (p. 200).

Leaf and McEachin (1999) stated that occasionally learner self-corrections are acceptable, and if a learner does self-correct then the therapist must repeat the trial to ensure that the learner can emit the correct response without self-correcting. However, the majority of DTT training manuals have incorrect responses immediately followed by the appropriate consequence and not another opportunity to respond.

Consequence

Consequences should be provided contingent on learner responses (Koegel et al., 1977). Effective reinforcing consequences can be identified by performing preference assessments (Cooper et al., 2007). For example, Leaf et al. (2010) used a pair-stimulus preference assessment to identify highly preferred toys to serve as reinforcers. Reinforcer effectiveness should be assessed continuously in order to prevent reinforcer satiation (Leaf & McEachin, 1999; Lovaas, 2003). In order to avoid reinforcer satiation Lovaas (2003) recommended that the learner only have between 3 – 5 s access to the reinforcer. However, access time to the reinforcer will depend on the type of reinforcer that was provided.

Consequences should be applied as soon as possible once the learner has emitted a response (Koegel et al., 1977; Leaf & McEachin, 1999; Lovaas, 2003). Koegel et al. (1977) recommended that the consequence be delivered within 3 s of the response. This is because it is important to establish a functional relationship between the response and the consequence (Ghezzi, 2007).

Consequences should be unambiguous (Koegel et al., 1977; Leaf & McEachin, 1999). The therapist should ensure that it is clear to the learner when a correct or error consequence is being administered. For example, if a therapist is administering an error consequence by saying the word “No”, the word “No” should not be accompanied with laughter or smiling (Leaf & McEachin, 1999). With regard to correct consequences social praise should be provided every time a correct consequence is administered and secondary items such as access to preferred items, activities and edibles if needed (Leaf & McEachin, 1999; Sundberg, & Partington, 1998).

Koegel et al. (1977) recommended that consequences should be consistent and follow each correct and incorrect response. However, Leaf and McEachin (1999) recommended that correct consequences be provided on a continuous reinforcement schedule initially before being thinned to an intermittent reinforcement schedule as the learner progresses. The thinning of the reinforcement schedule serves a number of purposes. Behaviours that are reinforced on thinner schedules of reinforcement are harder to extinguish relative to behaviours that are reinforced on richer schedules (i.e., Partial reinforcement extinction effect; Mackintosh, 1974). The thinning of the reinforcement schedule serves to make the behaviour more resistant to extinction. It is also important for appropriate behaviour to continue when therapy is not taking place and the target response is administered without the appropriate consequence following. Thinning the schedule may also reduce possible reinforcer satiation effects.

It is recommended that therapists do not provide correct consequences for incorrect learner responses (Lovaas, 2003). Basic operant conditioning research with pigeons using conditional-discrimination procedures have shown that reinforcing incorrect discriminations can have detrimental effects on their ability to make correct discriminations (Davison & McCarthy, 1980).

Intertrial Interval

The general conclusion about the duration of the ITI is that shorter ITI are preferable to longer ITI (Anderson et al., 1996; Lovaas, 2003; Smith, 2001). Some authors have recommended specific ITI durations. Smith (2001) recommended that the ITI should be between 1 and 5 s in duration, and Lovaas (2003) recommended between 1 and 3 s. Anderson et al. (1996) recommended between 3 and 5 s. Other authors have recommended that the ITI duration be catered to suit the learner so that an optimal teaching pace can be maintained (Leaf & McEachin, 1999). Belfiore, Fritts, and Herman (2008) stated that if the ITI is too short then the learner may not be able to discern between the end of a trial, and when to respond on the following trial. If the ITI is too long then the learner may engage in off-task behaviour. Belfiore et al. recommended between 2 and 5 s.

Studies by Koegel, Dunlap, and Dyer (1980), Valcante, Roberson, Reid, and Wolking (1989) have been inconclusive as to whether there is a functional relationship between ITI duration and learner performance during DTT. Koegel et al. (1980) conducted a study with children diagnosed with autism in which the length of the ITI was systematically varied.

Koegel et al. arranged short ITI conditions and long ITI conditions. The duration of the ITI varied between 1 and 3 s across the different short ITI conditions, and varied between 5 and 26 s across the different long ITI conditions. The study used both a multiple baseline and reversal design. In the multiple baseline design the long ITI condition was followed by the short ITI. During the reversal design, trials were administered during one ITI duration condition before changing to the other ITI condition. The conditions were then reversed again.

The results from the Koegel et al. (1980) study (multiple baseline design) showed that during the long ITI conditions learner performance was erratic and responding did not improve over time. The effect was especially evident for Learner 3 who received over 1000 trials without any marked improvement. Once the short ITI condition was introduced there was an immediate increase in learner performance and the mastery criterion was reached rapidly. The results from the reversal design showed that the percentage of correct unprompted learner responses increased during the short ITI conditions and decreased during the long ITI conditions.

Koegel et al. (1980) stated that the results did not suggest that shorter ITI durations will always be superior to longer ITI durations as other variables such as task difficulty and the individual characteristics of the learner need to be considered. However, the results are compelling and suggest a possible functional relationship between ITI duration and learner performance.

The finding that shorter ITI durations resulted in improved learner performance is at odds with findings from experimental operant conditioning research with pigeons. White and Wixted (1999) conducted a study in which they compared discriminability (log *d*, Davison & Tustin, 1978) in a delayed matching-to-sample (DMTS) task in which ITI durations were either 1 s or 15 s. The DMTS consisted of the presentation of a red or green sample stimulus.

A keypeck to the sample stimulus initiated a delay period after which the red and green sample stimuli appeared. A keypeck to one of the sample stimuli resulted in the appropriate consequence (reinforcement or blackout) before the ITI.

White and Wixted's (1999) results showed that discriminability was better in the long ITI duration condition (higher log *d* values) compared to the short ITI duration condition. One possible explanation for this finding is that there is less proactive interference from previously learned stimuli. Proactive interference occurs when previously learned stimuli interferes with the learning of new stimuli. Therefore, the results from White and Wixted do not agree with those of Koegel et al. (1980) who found better learner performance with shorter ITI. Having a shorter ITI in DTT may have a practical purpose because it may serve to reduce the likelihood of off-task learner behaviour, although such a hypothesis has not been tested.

Valcante et al. (1989) conducted a study in which both the ITI duration and the duration of time which had to elapse between the presentation of the S^D and the learner response (wait-time) were manipulated. Short ITIs were 1 s in duration and the long ITIs were 10 s in duration. When the wait-times were short the therapist waited one second for the learner to respond and when the wait-times were long the therapist waited 10 s for a learner response. The different wait-times and ITI durations were presented in all possible pairwise comparisons.

The results from the Valcante et al. (1989) study showed that longer wait-times were superior to shorter wait-times regardless of ITI duration. Therefore, the Valcante et al. results failed to support those of Koegel et al. (1980) as there were no detectable differences in learner performance regardless of ITI duration. Valcante et al. explained this discrepancy by stating that wait-time may be a more powerful independent variable than ITI duration.

Prompts

Prompts are usually administered during acquisition trials and trials with mastered stimuli should not be prompted (Lovaas, 2003). A prompted correct response should be rewarded with a lesser reinforcer (i.e., of smaller magnitude, or less preferred) than an unprompted response. The general consensus among authoritative DTT sources is that the initial prompting of correct responses should either be simultaneous with the S^D or immediately following the S^D (Anderson et al., 1996; Koegel, Glahn, & Nieminen, 1978; Leaf & McEachin, 1999; Lovaas, 2003; Smith, 2001). Lovaas recommended that if the prompt is not delivered simultaneously with the S^D then it has to be delivered within 1 s. If the prompt is not delivered in close temporal contiguity to the S^D then there is the risk that the prompt may lose its effectiveness and it may no longer lead to correct learner responses. Over time, as the learner progresses, prompts must be faded to avoid prompt dependency.

Prompts should be effective (Koegel et al., 1977; Leaf & McEachin, 1999). In other words, prompts should be followed by a correct learner response. If a prompt is consistently followed by incorrect responses then the prompt is ineffective and the therapist must consider using a different prompt (Leaf & McEachin, 1999).

Prompts should be selected that will result in correct learner responses without providing more assistance than is needed (Leaf & McEachin, 1999). Fox (1982) stated that, “You should begin with those prompts that will ensure that the desired behaviour will occur. These beginning prompts are likely to be very obvious to the student and involve your active participation.” (p. 83). Lovaas (2003) recommended that the effectiveness of the least intrusive type of response prompts should be assessed before progressing to a more intrusive type of prompt. Leaf and McEachin (1999) and Ghezzi (2007) also recommended that the least intrusive type of response prompt should be used first before progressing up the prompt hierarchy.

Summary

This chapter provided both descriptive and prescriptive accounts of the five DTT components. A summary of the information provided in this chapter can be seen in Figure 2.1. Figure 2.1 contains the five DTT components, the order in which these components should be administered, and a short summary of the prescriptive information provided. In this thesis when discussing within-trial treatment integrity the term within-trial treatment integrity refers to Figure 2.1. A trial with perfect treatment integrity will confirm to the guidelines outlined in Figure 2.1.

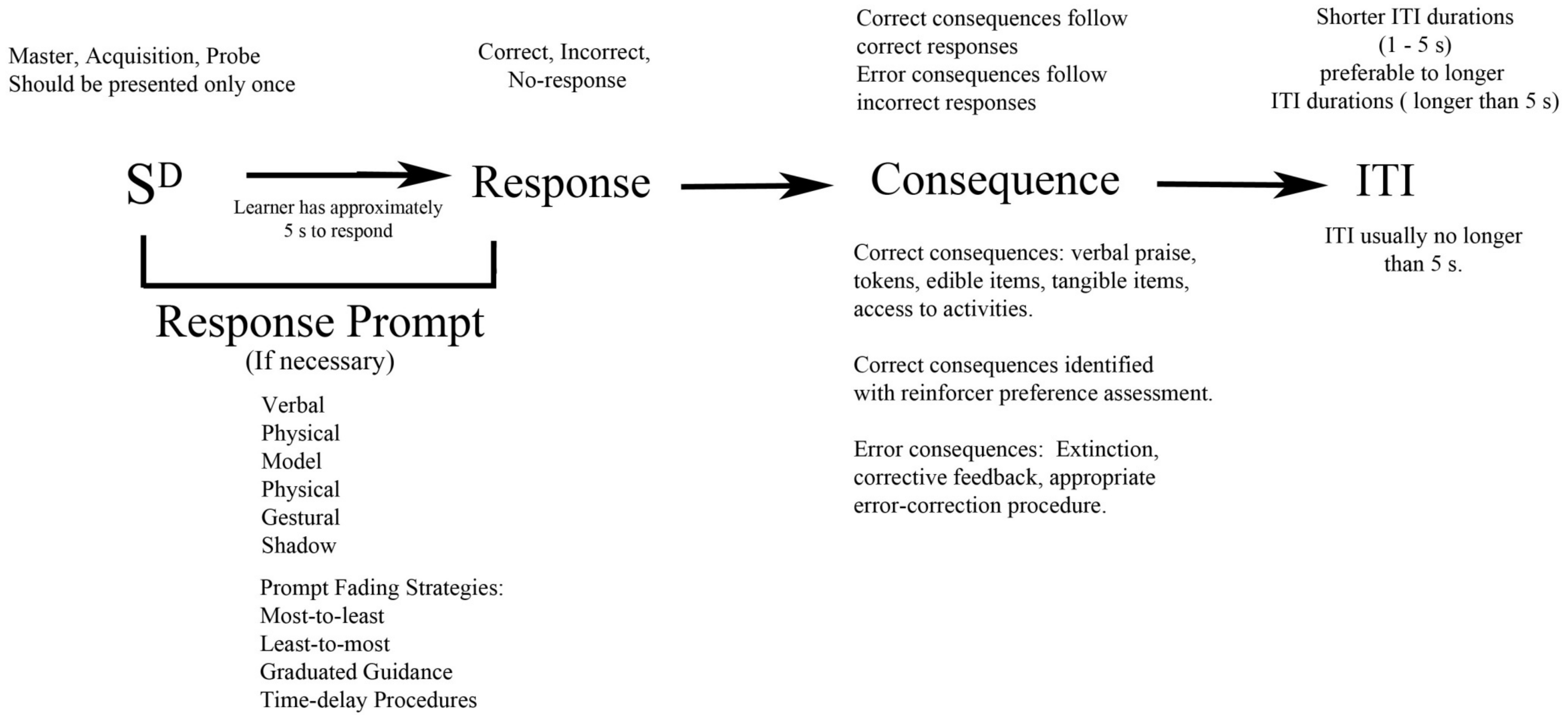

Figure 2.1: Summary of information presented in Chapter II.

Chapter III

Discrete-Trial Teaching: Data Recording, Analyses and Presentations

Traditionally, literature reviews and instructional guides on discrete-trial teaching (DTT) have focused on the different DTT components, effective applications of the procedure, and how DTT compares to other teaching methods such as incidental teaching or fluency training. Minimal attention has been given to how DTT data are collected, analysed and presented. This chapter will focus on recent research surrounding the recording of DTT data, and how DTT data have been analysed and presented in the past.

DTT Data Recording

Continuous versus Discontinuous Data Recording. One criticism of DTT has been that the procedure is time-consuming and can require a large number of trials for a learner to acquire and master a target skill (Smith, 2001). For example, one participant in the Carr, Kologinsky, and Leff-Simon (1987) study required over 11,500 trials for responding to generalize across all object-action test phrases. One way to try and decrease time spent collecting DTT data would be to collect data discontinuously, in other words, only to record data on certain trials. Traditionally, DTT data have been collected on a continuous basis. After every trial the therapist records the important events or outcomes as it relates to the trial that was just administered such as learner response and whether the response was prompted (Cummings & Carr, 2009).

Love, Carr, Almason, and Petursdottir (2009) conducted an online survey regarding a range of topics in relation to early intensive behavioural intervention (EIBI). DTT is an important teaching procedure used in EIBI (Lovaas, 1987). The survey required participants to report how often they recorded DTT data. The majority of the participants indicated that data are collected after every trial. In other words, continuous data recording was the most often used data collection method. Approximately 27% of participants indicated that data were collected on a discontinuous basis; either on the first trial of a lesson or on a subset of trials.

Lerman, Dittlinger, Fentress, and Lanagan (2011) stated that data collection methods that involved collecting data on a trial-by-trial basis will result in greater precision but may be less practical. Continuous data recording methods may also be a better indicator of changes in learner performance. On the other hand, discontinuous methods may result in less precise data but may be of practical importance. Recent research has begun to compare continuous and discontinuous data recording methods (Cummings & Carr, 2009; Najdowski et al., 2009; Lerman et al. 2011).

Cummings and Carr (2009) conducted a study in which they compared continuous and discontinuous data recording methods during DTT. During the continuous conditions data were recorded on a trial-to trial basis, and during the discontinuous conditions data were only collected on the first trial of a particular block of trials. Therefore, during the discontinuous conditions learner performance could only be scored as either 0% correct or 100% correct.

Cummings and Carr (2009) showed that when data were collected discontinuously the mastery criterion (two consecutive sessions 100% correct responding) was reached quicker than if data were recorded continuously. On the other hand, recording data continuously resulted in better skill maintenance. Continuous data recording may lead to better maintenance because the mastery criterion is more conservative than for discontinuous recording. Recording data discontinuously may lead to premature conclusions regarding the mastery of a target skill (Cummings & Carr, 2009).   Another important finding from the Cummings and Carr study was that collecting data continuously allowed for a within-session analysis of the data.

Najdowski et al. (2009) systematically replicated Cummings and Carr (2009). The biggest difference between the two studies was that Najdowski et al. used two different mastery criteria during the continuous and discontinuous data collection conditions. The mastery criterion during the continuous data recording condition was greater than 80% correct learner responses during three consecutive sessions. The mastery criterion during the discontinuous data collecting was 100% correct learner responses during three consecutive sessions. Obviously, the mastery criterion during discontinuous data collection conditions has to be 100% correct learner responses given that the learner can only score 0% correct or 100% correct. Cummings and Carr kept the mastery criterion the same for both data collection conditions.

Najdowski et al. (2009) found little difference between the two data collection methods in terms of percentage correct learner responses and number of sessions required to reach mastery criterion. However, Najdowski et al. stated that the results can be attributed to the differences in the mastery criterion between the two data collection methods. As a consequence the finding of no difference between the two methods should be interpreted with caution given the methodological confound.

The Lerman et al. (2011) study was an extension of the Cummings and Carr (2009) and Najdowski et al. (2009) studies. Target skills were taught to children with developmental disabilities or autism using DTT. Data were recorded either continuously (after every trial) or discontinuously. There were two variations to the discontinuous data recording; data were recorded either after every trial or on each of the first three trials. The mastery criterion for both continuous and discontinuous data recording was 88% or above unprompted correct learner responding and an unprompted correct learner response on the first trial.

Lerman et al. (2011) showed that the mastery criterion was reached in the fewest number of sessions when data were recorded after the first trial only. The mastery criterion took on average more sessions to reach the mastery criterion when data were recorded after each of the first three trials and after every trial. However, there was no difference between the two in terms of the number of sessions it took, on average, to reach the mastery criterion. These findings support the findings of Cummings and Carr (2009) that recording data after the first trial only may lead to premature conclusions regarding whether the learner has mastered the target skill. However, therapists would have reached similar conclusions regarding the mastery of a target skill if they had recorded data after the first three trials only or after every trial.

Lerman et al. (2011) provided several guidelines for best practice as it relates to data recording and DTT. Continuous data recording is advantageous because it is more precise and provides more information regarding learner progress. It is also associated with improved maintenance of a target skill because the mastery criterion is more stringent. Continuous data recording also allows for a within-session analysis of DTT. Continuous data recording may lead to longer sessions, delayed reinforcement and possibly less reliable data. It is recommended here that for the time being DTT data should be collected continuously given that it is more precise.

DTT Data Analysis and Presentations

Traditionally, DTT data have been analysed and reported using methods such percentage correct (Carr & Dores, 1981; Williams, Koegel & Egel, 1981; Schuster, Wolery & Griffen, 1992; Young, Krantz, McClannahan, & Poulson, 1994; Crockett, Fleming, Doepke, & Stevens, 2007; Downs, Downs, & Rau, 2008; Leaf, Sheldon, & Sherman, 2010), trials-to-criterion (Grindle & Remington, 2002; Cummings & Carr, 2009), number of target skills acquired (Summers & Szatmari, 2009), and average duration of procedure (Miranda-Linne & Melin, 1992; Wolery, Holcombe, Werts, & Cipolloni, 1993). These methods track changes in learner performance on a session-to-session basis. These aforementioned analysis methods can be presented as either a table or a graph. The main information usually required to complete these analyses is whether the learner made a correct response and whether the response was prompted or unprompted. Therefore, if a DTT programme is effective it is expected that a number of things will happen. It is expected that the percentage of correct responses will increase, the number of trials needed to reach the mastery criterion will be fewer relative to a less effective programme, the number of target skills acquired by the learner will increase, and the time needed to train target skills will decrease. These methods are good for displaying general patterns and trends.

The most common method of analysing DTT data is percentage correct. A percentage correct analysis involves adding up the number of correct learner responses and dividing the number by the total number of trials, and results are multiplied by 100. An effective DTT programme should result in a higher overall number of correct responses than a less effective DTT procedure. A percentage correct analysis is advantageous because percentages can be easily calculated and graphs are easy to construct and understand.

DTT data are also analysed using the trials-to-criterion method. A trials-to-criterion analysis investigates how many trials it takes for a learner to master the target skill in accordance with the mastery criterion. For example, if the mastery criterion is 80% correct learner responses for three consecutive sessions a trials-to-criterion analysis will record how many trials it took for the learner to reach that criterion. A more effective DTT programme will result in fewer trials to reach the criterion than a less effective programme.

Another method used to evaluate learner performance during DTT is the time it took (on average) for a learner to master a target skill (Wolery et al., 1992). Methods such as trials-to-criterion and procedure-duration are often used when comparing DTT to other teaching programmes (Miranda-Linne & Melin, 1992; McGee, Krantz, & McClannahan, 1985; Holding, Bray, & Kehle, 2011), or when researchers are comparing different DTT component manipulations (Wolery et al., 1993).

Within-Sessions Analyses. All of the DTT data analysis methods discussed require that the outcome of a trial and the level of prompting be recorded. These are summed across all trials from a lesson to produce one data point. For example, in a DTT lesson 10 trials are administered and the learner emits eight correct responses. If the data were analysed and plotted as percentage correct it would produce a singular data point of 80%. In order to investigate learner progress for that particular lesson the same analysis has to be conducted every time the lesson is administered and the percentage correct scores investigated to examine whether the learner is progressing. Analysis methods such as percentage correct do not provide information as to how learner performance changed on a trial-by-trial basis. In other words, they do not provide any information as to what is occurring either between trials during the course of a lesson or events occurring within each individual trial.

Sequential Analysis is the generic term for a set of within-session data analysis methods which are used to identify patterns and sequences within data. These methods, described in Chapter I, include conditional probability analysis (Lerman & Iwata, 1993), lag sequential analysis (Sackett, 1979), Yule's Q (Oliver, Woodcock, & Humphreys, 2009), contingency space analysis (Matthews, Shimoff, & Catania, 1987), and Markov probability chains (Gottman & Notarius, 1978). All these methods produce a statistic which indicates how likely the occurrence or non-occurrence of an event is based on the occurrence or non-occurrence of another event or events. In order to perform a sequential analysis with DTT data, the data must be collected continuously (Cummings & Carr, 2009).

Cummings and Carr (2009) stated that one advantage of collecting data on a continuous basis is that it allows data to be analysed on a within-session basis. Heckaman, Alber, Hooper, and Heward (1998), and Lerman et al. (2011) are two examples of when DTT data were analysed using one of the sequential analysis methods. Heckaman et al. calculated the conditional probability of disruptive learner behaviour following two different types of prompt-fading procedures. The prompt-fading procedures were least-to-most, and progressive time-delay. Heckaman et al. showed that across all learners and both prompting procedures disruptive learner behaviour was less likely to follow an effective controlling prompt. However, disruptive behaviour was more likely for two of the learners to follow less effective prompts. Therefore, the type of prompt administered within each of the prompt fading procedures (effective versus less effective prompt) did have predictive properties.

The Lerman et al. (2011) study analysed DTT data on a within-session basis using a conditional probability analysis. The study compared continuous and discontinuous data collection methods in terms of how well these methods can predict learner performance across all trials. Twenty four acquisition targets were taught across 11 learners. Learners received either eight or nine trials per session which consisted of both the acquisition target and previously mastered targets. However, all data collected from the trials with mastered stimuli were not used in the analyses. Analyses were conducted across all acquisition targets.

Lerman et al. (2011) recorded whether the learner responded correctly or incorrectly on the first trial. The analysis consisted of calculating the conditional probability that the learner emitted a correct response on greater than 50% of all trials based on how they responded on the first trial. The aim of the analysis was to investigate whether accurate predictions can be made about learner performance across all trials when data are recorded on the first trial only.

The Lerman et al. (2011) results showed that the conditional probability of the learner making a correct response on greater than 50% of all trials given that the first response was a correct response was .92 (range .66 – 1.00). When the learner responded incorrectly on the first trial the conditional probability was .26 (range .00 – .71). The results indicated that how the learner responded on the first trial was a good predictor of learner performance across the rest of the trials.

Summary

The research (Heckaman et al., 1998; Lerman et al., 2011) and ideas (Cummings and Carr, 2009; Najdowski et al., 2009) presented in this chapter strongly indicates that further research should be conducted on analyzing DTT data on a within-session basis as there is a substantial gap in the DTT literature as it relates to analysing DTT on a within-session basis. The research involving continuous data collection showed that collecting DTT data continuously results in greater precision and an increased ability to assess learner progress (Lerman et al., 2011; Cummings & Carr, 2009). As it relates to the purposes of this thesis collecting data continuously also allows DTT data to be analysed on a within-session basis. If every occurrence of all relevant DTT events is recorded then several interesting analyses can be conducted. One such analysis includes conducting within-trial analyses to assess treatment integrity and evaluate the accuracy with which the procedure is administered. Markov probability chains would be an appropriate sequential analysis method to conduct such analyses.

The research involving discontinuous data collection, along with the results of Lerman et al. (2011), suggests that it is possible to predict the outcome of future DTT trials (with a certain degree of accuracy and up to a point) based on the outcome of a current trial. In other words, the events occurring during discrete teaching trials have predictive value. One way to investigate further the predictive properties of DTT outcomes on a between-trial basis is to conduct a sequential analysis using Markov probability chains. For example, it would be of interest to investigate the likelihood of long sequence of correct learner responses following an incorrect response, or the likelihood of a correct learner response following a trial which was administered with a procedural error. Markov probability chains also appear to be an appropriate method of analysis for such research questions.

Chapter IV

Treatment Integrity

Applied behaviour analysis (ABA) involves studying functional relationships between dependent variables (DVs), and independent variables (IVs) (Baer, Wolf, & Risley, 1968). In order to draw accurate conclusions about functional relationships researchers must ensure that IV are implemented with a high degree of integrity (Peterson, Homer, & Wonderlich, 1982). Treatment integrity is the extent to which treatment procedures are implemented in the manner in which they are designed (Peterson et al. 1982). Administering procedures with a high degree of integrity helps to minimize the likelihood that the results from a study are affected by confounding variables. Confounding variables are introduced into a study if the procedures are not administered in the prescribed manner (Cook & Campbell, 1979). Gresham, Gansle, and Noell (1993) provided the example of a token economy. The token economy is implemented with the intention of increasing appropriate learner behaviour. If the token economy was programmed so that tokens were to be administered without social praise, and the therapist provides social praise whenever a token is presented, the social praise is now a confounding variable. The social praise is a confounding variable because it cannot be determined whether an increase in appropriate learner behaviour was the result of receiving a token, receiving social praise, or both.

Measuring Treatment Integrity

Treatment integrity can be measured using direct observation methods, measuring permanent products, self-reporting, and behavioural interviews (Sanetti & Kratochwill, 2008; Liaupsin, Ferro, & Umbreit, 2012). Direct observation is the method used most often to measure treatment integrity, and has many of the same requirements as direct observations of DVs for the purposes of interobserver agreement. Direct observation methods require that each treatment component be adequately defined, that data be recorded on both the occurrence and nonoccurrence of each treatment component, and involve calculating a percentage for the occurrence and nonoccurrence of each component (Sanetti & Kratochwill, 2008). Direct observation methods include response-by-response recording, whole-interval recording, and checklists (Liaupsin et al., 2012).

Response-by-response recording consists of recording the frequency and (if needed) the duration with which the therapist administers each treatment component. Although response-by-response recording is the most complete method of assessing treatment integrity it is labour-intensive and may not be practical in many cases (Liaupsin et al., 2012). Checklists involve recording the occurrence and nonoccurrence of defined items on a checklist, and calculating a percentage of items on the checklist that were correctly administered (Liaupsin et al., 2012). Whole-interval measurement consists of dividing sessions into time-intervals. Observers record whether the treatment components administered during a recording-interval were correctly administered. The results are displayed as percentage of intervals in which the treatment components were correctly administered (Liaupsin et al., 2012).

Permanent products are another way of measuring treatment integrity. Permanent products are defined as, “a change in the environment produced by a behaviour that lasts long enough for measurement to take place,” (Cooper, Heron, & Heward, 2007, p. 95). Permanent products are measured once behaviour has already occurred. An example of a permanent product can be found in Alber, Nelson, and Brennan (2002). In the Alber et al. study students were required to complete homework assignments using two different methods. The completed worksheets are an example of a permanent product. An advantage of permanent products is that they can reduce the likelihood of observer reactivity. Observer reactivity occurs when participants being recorded change their behaviours because they are aware that they are being recorded (Foster & Cone, 1986; Sanetti & Kratochwill, 2008).

Treatment integrity can also be measured with therapist self-reporting and behavioural interviews. Self-reporting requires the therapist to report at the end of a session on their own levels of treatment integrity. Self-reporting requires that an appropriate self-reporting framework be in operation (for example checklists, or Likert scales), and that these be completed on a regular basis. However, self-reporting is susceptible to therapists either over-or-under estimating the integrity with which they administered the procedure and it may be difficult to obtain a true treatment integrity measure using only self-reporting (Sanetti & Kratochwill, 2008). Behavioural interviews involve conducting interviews with the therapist where they report back about the integrity with which the procedure was administered. Behavioural interviews can be labour-intensive and susceptible to self-reporting bias (Foster & Cone, 1986).

Reporting Treatment Integrity Data in ABA Publications

Several review articles have been published in order to evaluate the frequency with which publications within recognized behaviour-analytic journals reported treatment integrity data, and provided operationally defined IVs (Armstrong, Ehrhardt, Cool, & Poling, 1997; Gresham et al., 1993; Monchar & Prinz, 1991; Peterson et al., 1982; Wheeler, Baggett, Fox, & Blevins., 2006). The Peterson et al. (1982) review investigated all articles published in the *Journal of Applied Behavior Analysis* (*JABA*) between 1968 and 1980, and was the first to review the frequency with which treatment integrity data and operationally defined IV were provided. In total 539 articles were reviewed. The results from the review showed that on average approximately 20% of articles reviewed provided treatment integrity data. Of the studies which did conduct treatment integrity checks only 16% provided operational definitions of IVs.

Gresham et al. (1993) reviewed all the articles published in JABA between 1980 and 1990 that consisted of experimental studies assessing treatment effectiveness with learners under the age of 19. A total of 158 articles met the inclusion criterion. The results showed that 52 studies (34.2%) provided operational definitions of IV, and 25 (15.8%) provided treatment integrity data. When treatment integrity data were provided the scores were high and averaged 93.8%.

Monchar and Prinz (1991) reviewed 359 treatment outcome studies sourced across several different psychology journals, and found that 55% of studies did not make any reference to treatment integrity. The results from the Wheeler et al. (2006) review were similar to the Monchar and Prinz review. Wheeler et al. included all articles across a range of behavioural journals, and studies included in the review had to be experimentally based interventions with learners under 18 years of age who were diagnosed with autism spectrum disorders. Sixty articles met the criterion. Out of the 60 articles 55 (92%) provided operational definitions of the IV, but 68% (41) did not provide any treatment integrity data. Eleven articles (18%) both defined IV and conducted treatment integrity checks.

Armstrong et al. (1997) were the first to review the frequency with which treatment integrity data were reported in the *Journal of Developmental and Physical Disabilities* between 1991 and 1995. The inclusion criterion for articles to be reviewed was that articles had to include descriptions of experimental procedures designed to improve behaviour for at least one participant. Thirty nine articles met the criterion. The results showed that only 23% of the articles conduced treatment integrity checks.

These reviews (Armstrong et al., 1997; Gresham et al., 1993; Monchar & Prinz, 1991; Peterson et al., 1982; Wheeler et al., 2006) found that less than 50% of all articles reviewed did not provide any treatment integrity data (range = 15.8% - 45%). The reviews (Gresham et al., 1993; Peterson et al., 1982; Wheeler et al., 2006) that investigated the frequency with which operational definitions of IV were provided found a larger range (16% to 92%)

compared to the articles reporting treatment integrity data. It was rare for articles to provide both treatment integrity data and operational definitions of IV (range = 16% to 18%).

Factors Affecting Treatment Integrity

There are several factors that contribute to treatment integrity. These are the amount of experience a therapist has with administrating the procedure (Fryling, Wallace, & Yassine, 2012), treatment drift (Peterson et al., 1982), inadequate operational definitions of IV (Gresham et al., 1993), and the complexity of the procedure (Cooper et al., 2007).

Much research has shown that treatment integrity levels increased beyond initial baseline levels once therapists received treatment-specific training (DiGennaro, Martens, & Kleinman, 2007; DiGennaro, Martens, & McIntyre, 2005; Koegel, Russo, & Rincover, 1977; Mueller et al., 2003; Sarokoff & Sturmey, 2004; Vladescu, Carroll, Paden, & Kodak, 2012). Training methods in these studies included video-modelling, rehearsal, corrective feedback, instruction manuals, and modelling. Generally, the studies have shown that therapists can be trained in a relatively short period of time, and that increases in treatment integrity scores were maintained.

All IVs of a study should be defined in a manner that is clear, unambiguous, and comprehensive (Gresham et al., 1993). One of the fundamental tenets of ABA is that it is technological. In other words, all procedures and variables within a study should be described in sufficient detail so that the procedure can be replicated (Baer et al., 1968).

Treatment drift occurs when the administration of a treatment procedure is different at the end of a study than at the beginning (Peterson et al., 1982). An example of treatment drift is if the treatment protocol requires a therapist to wait 5 s before administering the reinforcer following appropriate learner behaviour, but after a few sessions the therapist begins to provide the reinforcer after waiting only 3 s. Treatment drift can result from complex procedures which make it difficult for therapists to administer all aspects of the procedure with the same levels of integrity across the entire length of the study (Cooper et al., 2007). Cooper et al. stated, “Treatments that are simple, precise, and brief, and require relatively little effort, are most likely to be delivered with consistency than those that are not,” (p. 235). However, this does not imply that complex procedures should not be administered. When complex procedures are administered researchers need to be cognizant about potential problems relating to treatment integrity (Peterson et al., 1982).

Treatment Integrity and Validity

Validity refers to the extent that the outcome of a research study or intervention can be attributed to the procedure or intervention (Cook & Campbell, 1979; Johnston, & Pennypacker, 1980). Low levels of treatment integrity can compromise the validity of a study, and can lead to conclusions that an ineffective treatment is effective (false positives, or Type I errors), or that effective treatments are ineffective (false negatives, or Type II errors) (Gresham et al., 1993). Both false positives and false negatives can have detrimental effects on the development of effective interventions (Cooper et al., 2007). Low treatment integrity levels can also lead to inaccurate conclusions about the functional relationships between variables (Gresham et al., 1993). There are several different types of validity but the ones relevant to the current discussion are internal, external, and construct validity.

Internal Validity. Low levels of treatment integrity pose a major threat to the internal validity of a study. Internal validity is the degree to which changes in the DV can be attributed to manipulations of the IV (Kazdin, 1998). If the internal validity of a study is compromised it limits the ability with which accurate conclusions about functional relationships between variables can be made (Peterson et al., 1982). If the internal validity of a study is threatened it can lead to inaccurate conclusions regarding treatment effectiveness. For example, if the results of a study show that a treatment was ineffective, in the absence of any treatment integrity data, it cannot be determined whether the treatment was ineffective or whether the IV were incorrectly administered. In the case of an effective intervention it is not possible to determine whether the intervention was effective or whether the results were influenced by any confounding variables (Monchar & Prince, 1991).

External Validity. External validity is also affected by low levels of treatment integrity. External validity refers to the degree that findings from a study can be generalized across different participants, settings, and behaviours (Cooper et al., 2007). External validity is assessed by the process of replication (Baer et al., 1968). The external validity of a study is diminished if IV are not adequately defined making it no longer possible to fully replicate a procedure.

To illustrate the aforementioned points the Gresham et al. (1993) token economy example will be used again. The token economy is programmed so that the token is administered without social praise. However, the therapist administers the token with the additional social praise and this information is not included in the write-up of the study. The intervention results in an increase in appropriate learner behaviour and is deemed effective. Another researcher attempts to replicate the procedure and administers the token without providing the social praise and finds that the treatment is ineffective. Therefore because the procedure was not administered as programmed, and the description of the procedure did not include a description of the social praise accompanying the delivery of token, the procedure is not replicated as originally administered and the results not reproduced (Gresham et al., 1993).

Peterson et al. (1982) stated that researchers do have the ability to change aspects of the intervention as required as long as those changes are acknowledged. It should also be reported if at any time during the study the procedure was administered differently to what was prescribed.

Construct Validity. When treatment integrity is compromised construct validity is also affected (Monchar & Prinz, 1991). In other words, the ability to understand the causal mechanisms of a treatment procedure is compromised. If the treatment is not administered in the manner prescribed then the effects of the treatment are confounded with other extraneous variables and the causal mechanism for change in the target behaviour cannot be accurately determined (Monchar & Prinz, 1991).

Reporting Treatment Integrity Data

There are two ways in which treatment integrity scores can be reported. These are component integrity and session integrity (Gresham et al., 1993). Component integrity is used to measure the integrity with which each component of a procedure is administered. Session integrity reports the treatment integrity number across all components of a procedure (Gresham et al., 1993).

Treatment Integrity: Correlations and Systematic Manipulations

Research (Arkoosh et al., 2007; DiGennaro et al., 2005; DiGennaro et al., 2007; DiGennaro Reed et al., 2011; Wilder, Atwell, & Wine, 2006) has shown that behavioural treatments are most effective when treatment integrity levels are high (at least 80%). Studies investigating the relationship between treatment integrity and treatment effectiveness are either correlational, or involve measuring treatment effectiveness when treatment integrity levels are systematically manipulated (DiGennaro Reed et al., 2011). Correlation studies involve the use of various methods to increase the integrity of a procedure. Data are recorded on how learner behaviour changes with changes in treatment integrity. Studies involving systematic manipulations of treatment integrity investigate treatment effectiveness across varying levels of treatment integrity.

DiGennaro et al. (2007) is an example of a study where a correlational analysis was conducted to assess the relationship between treatment integrity and treatment effectiveness.

The therapists were teachers who were trained to administer an intervention designed to reduce problem behaviours of learners in a school setting. The main aim of the study was to investigate whether the integrity with which therapists administered the intervention was subject to contingencies of reinforcement. In particular the researchers aimed to assess the effects of negative reinforcement. Therapists were able to avoid consultant meetings if the intervention was administered with 100% integrity. During the meeting with the consultants all of the treatment integrity errors were reviewed and the therapist had to practice the correct application of the errors three times. The study used a multiple baseline across participants design, and training consisted of didactic teaching, corrective feedback, coaching and modelling. The results from the DiGennaro et al. study indicated that treatment integrity did increase to levels above baseline following training. The correlation analysis showed that there were significant correlations between increases in treatment integrity and decreases in problem behaviour.

The Wilder et al. (2006) study provides an example in which treatment effectiveness was monitored when treatment integrity levels were systematically manipulated. Wilder et al. (2006) conducted a study in which learner compliance was monitored when administering a three-step prompting procedure with two typically developing learners. Treatment integrity levels were systematically varied across three different conditions (0%, 50% and 100%). The 100% treatment integrity condition involved administering the prompting procedure following each instance of learner noncompliance, and only half of all noncompliance trials during the 50% condition. During the 0% treatment integrity condition the procedure was not administered following any instances of noncompliance.

The Wilder et al. (2006) results showed that learner compliance was highest during the 100% and 50% treatment integrity levels and lowest during the 0% condition. The levels of compliance for each condition were clearly differentiated. There were clear increases in learner compliance between baseline and both the 50% and 100% treatment integrity conditions. However, no increase in learner compliance was evident during the 0% treatment integrity condition. Therefore, the Wilder et al. results showed a clear functional relationship between treatment integrity levels and learner compliance.

Treatment Integrity: Commission and omission Errors

St. Peter Pipkin, Vollmer, and Sloman (2010) broadly categorized treatment integrity failings as errors of commission and omission. Commission errors occur when the therapist administers procedures not prescribed in the intervention protocol. An example of a commission error is if the therapist provides a response prompt if the procedure is programmed to be administered with no response prompts. Omission errors occur when the therapist does not implement prescribed parts of an intervention. An example of an omission error is if the therapist did not provide a token for correct learner responses when a token economy was in place (St. Peter Pipkin et al.).

Research, primarily involving differential reinforcement of alternative behaviour (DRA) schedules, has investigated how these two different types of treatment errors affect intervention outcomes. DRA schedules involve withholding reinforcers for inappropriate behaviours and providing reinforcers for defined alternative behaviour, and an omission error would include not reinforcing appropriate behaviours. A commission error occurs when reinforcers are provided for inappropriate behaviours.

St Peter Pipkin et al. (2010) stated that it cannot be assumed that commission and omission errors affect treatment effectiveness to the same extent. As it relates to DRA schedules, more research is needed to evaluate whether omission errors (not reinforcing appropriate learner behaviour), has the same effect on treatment effectiveness as reinforcing inappropriate learner behaviours (commission error). If treatment integrity is reported as a single measure (involving both commission and omission errors) it is not possible to outline how each one of these errors affect treatment effectiveness. It is possible that certain types of errors may lower the overall treatment integrity scores but not affect treatment effectiveness, whereas other errors may result in drastic decreases in treatment effectiveness. St. Peter Pipkin et al. referred to these as critical and noncritical treatment components. The issue of critical and noncritical treatment components is very relevant to other treatment procedures consisting of multiple components such as discrete-trial teaching (DTT), and any of the differential reinforcement schedules.

Treatment Integrity and Discrete-Trial Teaching

DTT is a teaching strategy used to teach functional skills to people with intellectual disabilities and consists of multiple components. These components are the S^D (therapist instructions to learner), a prompt (if necessary) to assist the learner with emitting a correct response, a learner response (correct, incorrect, no response), a consequence provided by the therapist based on the learner response, and an ITI. The integrity with which each of these components is administered contributes to the overall treatment integrity with which the procedure is administered. The following sections on DTT and treatment integrity will discuss the relationship between DTT-specific training and overall treatment integrity (Catania, Almeida, Liu-Constant, & Reed, 2009; LeBlanc, Ricciardi, & Luiselli, 2005; Sarokoff & Sturmey, 2004), the relationship between increases in DTT treatment integrity, and the corresponding effects on the number of correct learner responses (Downs, Downs, & Rau, 2008; Koegel et al., 1977; Sarokoff & Sturmey, 2008; Vladescu et al., 2012), the relationship between treatment integrity and disruptive learner behaviour during DTT (Dib & Sturmey, 2007), and research that has systematically manipulated specific types of component errors to evaluate how they affect correct learner responses (DiGennaro Reed et al., 2011).

Treatment Integrity and DTT-specific Training

Vismara and Rodgers (2010) stated that one concern as it relates to DTT and its administration is the level of expertise on the part of the therapist. Smith (2001) wrote that in order for therapists to administer DTT with high treatment integrity they must receive DTT-specific training. Studies by Catania et al. (2009), LeBlanc et al. (2005), and Sarokoff and Sturmey (2004) demonstrated that once therapists received DTT-specific training treatment integrity increased. The training could be completed relatively quickly and the increases in treatment integrity scores were maintained. All three studies trained therapists across 10 DTT components that included correct administration of S^D, prompting procedures, error-correction procedures, and ITI. Treatment integrity scores were reported as an aggregate across these components. For the LeBlanc et al. and Sarokoff and Sturmey studies therapists were trained to administer each DTT component with at least 90% accuracy.

Sarokoff and Sturmey (2004) used a behavioural skills training package (BST) to train special education teachers with various degrees of experience administering DTT procedures. The BST package consisted of written instructions, vocal feedback, rehearsal of procedures, and modeling the correct administration of the DTT procedure. Baseline treatment integrity ranged between 43% and 49%. Once the training was completed treatment integrity scores ranged between 97% and 99%.

The LeBlanc et al. (2005) study used a DTT training procedure consisting of abbreviated performance feedback to train teacher aides at a special needs school. Correct administration of DTT components was reinforced with verbal praise, and therapists were corrected if a component was administered incorrectly. Baseline treatment integrity scores ranged between 32% and 43%. Following training treatment integrity scores increased to above 90% accuracy. No more than five training sessions for each therapist was required.

The increases in treatment integrity scores were maintained for up to 11 weeks following training.

The Catania et al. (2009) study consisted of training staff employed at a special needs school to implement DTT procedures accurately using video-modelling. The video-modelling consisted of a voice-over script, and an explanation of the modeled skill. Two out of the three therapists had experience administering DTT procedures and the other therapist had no previous DTT experience. The mean treatment integrity baseline results for the two therapists who had experience administering DTT were 48% and 63%, and 21% for the therapist with no experience. Following training the mean treatment integrity scores for all three therapists ranged between 85% and 98%.

The Catania et al. (2009) study found that one therapist had a higher treatment integrity score during the baseline probe than during the training phase and investigated which error was contributing to this result. The results showed that it was the incorrect implementation of the error-correction procedure which affected treatment integrity the most. However, during the baseline probe the learner did not emit any incorrect responses but did emit six incorrect responses during the training sessions. Therefore, because the event never occurred it was not possible for that specific procedural error to occur. The results from the Catania et al. study highlight that in order to conduct a thorough and comprehensive treatment integrity analysis of specific DTT errors data must be recorded on a sufficient number of trials for treatment integrity scores to be truly representative of what is actually occurring within that DTT programme.

The combined results from the Catania et al. (2009), LeBlanc et al. (2005), and Sarokoff and Sturmey (2008) studies show that once therapists had received DTT-specific training treatment integrity scores increased above baseline levels regardless of whether therapists had previous experience administering DTT. Although these studies were able to demonstrate increases in treatment integrity following training they did not address some key issues. None of the studies investigated whether an increase in treatment integrity scores resulted in increases in correct learner responses. The LeBlanc et al. and Sarokoff and Sturmey studies did not provide a comprehensive analysis of treatment integrity scores for each DTT component assessed. Some errors may have occurred more frequently than others and would affect treatment integrity scores to a greater extent. For example, the therapist may be administering all of the S^Ds incorrectly, but administering the rest of the procedure with 100% integrity. In that case, it is the incorrect implementation of one specific type of error that is contributing to the low integrity score. The question then is whether it is necessary to retrain the therapist on all of the DTT components, or just specifically to administer S^Ds correctly. Therefore, although DTT-specific training was completed relatively quickly during the training phase for all the studies, the time needed to retrain therapists could potentially be reduced even further if analyses were conducted to identify which specific type of component errors were contributing the most to the low treatment integrity scores and retraining therapists to avoid those particular errors.

Treatment Integrity and Correct Learner Responses

Research (Downs, Downs, & Rau, 2008; Koegel et al., 1977; Sarokoff & Sturmey, 2008; Vladescu et al., 2012) has shown that correct learner responses increased once therapists had been trained to administer DTT procedures with high treatment integrity. The results from the studies indicate that increases in correct learner responses was observed once therapists had been trained to administer DTT trials with at least 90% treatment integrity. All of the studies involved DTT training across all DTT components.

Koegel et al. (1977) conducted a multiple baseline study and recorded treatment integrity scores and correct learner responses before and after therapists had been trained to correctly administer DTT trials. Therapist training involved reading training manuals and watching videos of correct and incorrect application of the different DTT components. The therapists also received periodic corrective feedback from instructors during training sessions.

Baseline data from the Koegel et al. (1977) study showed that a majority of the therapists administered DTT with than 60% treatment integrity. Only one therapist scored above 70% correct consistently. Once training was completed treatment integrity scores for all therapists were at least 90%. This increase in treatment integrity was accompanied by an increase in learner performance. Koegel et al. stated “generally, for any given session, systematic improvement in the child’s behaviour did not occur unless the teacher working in that session had been trained to use the techniques to a high criterion,” (p. 197).

Although the Koegel et al. (1977) study provided interesting results the study is limited in one regard. The study did not identify where exactly the procedural errors were occurring. Koegel et al. calculated a percentage correct score for the application of each defined DTT component and these percentages were averaged across the session. Therefore, it is not possible to tell which errors contributed the most to the low levels of treatment integrity.

Koegel et al. (1977) stated that other factors besides proficiency in implementing the different DTT components may contribute to learner performance. Downs et al. (2008) referred to these other factors as DTT support skills. These include deciding which stimuli to teach, choosing effective positive consequences for correct learner responding, and how problematic learner behaviour is managed. Therefore, Downs et al. conducted a study in which therapists were trained to implement DTT and also recorded data on DTT support skills. The support skills included whether the therapist was ready to begin the therapy session once the learner arrived and whether the therapist had read the clinical notes before the start of that particular therapy session. The Downs et al. (2008) study consisted of three phases; baseline, intervention and follow-up. Before the baseline phase therapists who had no experience with DTT were trained to administer DTT procedures using didactic teaching, corrective feedback and live modelling. Following the baseline session further training was provided. Maintenance data were recorded after the intervention phase.

Treatment integrity data from the Downs et al. (2008) study were analysed as percentage correct implementation of the DTT procedure and DTT support skills as a function for each session across the three phases. Learner responses were presented as percentage correct. The baseline treatment integrity results ranged between 63% and 80%. Following therapist training the results showed that treatment integrity levels increased between the baseline and intervention phases. Treatment integrity levels were above 90% during the intervention phase. The percentage of correct learner responses increased for all learners involved during the intervention phase. Furthermore, maintenance data showed that treatment integrity levels were above 95% for all therapists.

The finding that learner performance systematically increased once treatment integrity levels increased is consistent with the results from Koegel et al. (1977), and the finding that DTT specific training is required to increase treatment integrity levels was also consistent with other research (Catania et al., 2009; Koegel et al., 1977; LeBlanc et al., 2005; Sarokoff & Sturmey, 2004). Furthermore, the Downs et al. (2008) results suggests that regular retraining and supervision of DTT programmes are required to maintain high treatment integrity levels. The Downs et al. (2008) study did not provide an analysis of which part of the DTT training (DTT-specific or DTT supported skills training) contributed most to both treatment integrity and the increases in correct learner responses. It would have been interesting if an analysis could have been conducted to differentiate the relative contribution of both DTT-specific and DTT support skills training on treatment integrity levels and learner performance.

Sarokoff and Sturmey (2008) conducted a study in which BST was used to train therapists to correctly implement DTT procedures when teaching children diagnosed with autism. BST consisted of vocal feedback, written instructions and modelling. The study consisted of an initial baseline phase, an intervention phase, and a follow-up phase. Therapists were selected because they administered DTT with less than 50% proficiency.

Results from the Sarokoff and Sturmey (2008) study found that baseline treatment integrity levels ranged between 23% and 50%. Once BST had been completed treatment integrity scores for all therapists were above 90%. Treatment integrity gains were also maintained over time and were above 94% during the follow-up phase for all therapists. Sarokoff and Sturmey also found that high treatment integrity levels generalized when teaching new learners and new skills. In terms of learner performance, the results found that the percentage of correct learner responses increased for all learners once treatment integrity levels have increased above baseline.

Vladescu et al., (2012) conducted a study similar to Catania et al. (2009) in which video-modeling was used to train therapists to administer DTT. The Vladescu et al. study extended the research of Catania et al. by not only evaluating the effectiveness of DTT training using video-modeling but also correct learner responses. The study consisted of a baseline phase, the video-modeling phase, teaching novel tasks phase (expressive labels and matching-to-sample), and child training phase. Six probe trials were administered to evaluate therapist progress with both an adult confederate (adult probe), and a learner with a disability (child probe). The final phase of the study was the child training phase where trained therapists had to administer a new task (receptive identification).

The initial baseline results from the Vladescu et al. (2012) study showed low levels of treatment integrity. Treatment integrity results ranged between approximately 20% and 50%. During the video-modeling phase therapists were trained until they administered DTT with at least 90% accuracy for two consecutive sessions. During the teaching novel tasks phase results from the adult and child probe showed that treatment integrity remained above 80%. High treatment integrity was also found during the child training phase. Learners were also able to reach the mastery criterion (at least 90% correct responding for two consecutive sessions) for the task taught during the child training phase.

DTT Treatment Integrity and Disruptive Learner Behaviour

Dib and Sturmey (2007) conducted a study to investigate whether an increase in DTT treatment integrity levels coincided with changes in disruptive learner behaviour. The disruptive learner behaviour was stereotypy. Stereotypy was defined as repetitive movements (such as hand-flapping) and vocalizations (Dib & Sturmey, 2007). The study consisted of an initial baseline phase and a training phase. Therapists were selected on the basis that they administered DTT with low levels of procedural integrity.

The results from the Dib and Sturmey (2007) study showed that the initial baseline levels of DTT treatment integrity were very low with the range being between 0 and 4%. Once therapists had received DTT-specific training treatment integrity levels increased to 100% for all therapists. Levels of stereotypy also decreased once treatment integrity levels had been improved.

DTT Commission and Omission Errors

Not much research has been conducted to evaluate how specific types of commission and omission errors affects correct learner responding. Examples of omission errors in DTT includes failing to provide positive reinforcing consequences following correct learner responses, the absence of an ITI, and failing to provide a negative consequence following an incorrect learner response. Examples of DTT commission errors include providing positive reinforcing consequences following incorrect learner responses, excessive prompting, and multiple successive presentations of S^D.

The DiGennaro Reed et al. (2011) study consisted of systematically manipulating the frequency of a specific type of commission error to investigate how it affected learner performance. DiGennaro Reed et al. specifically examined the effects of commission errors during DTT with learners diagnosed with autism. The commission error for this study was reinforcing incorrect learner responses. The IV in the study was the frequency with which the commission errors were applied. No incorrect responses were reinforced during the 0% commission error condition, every second incorrect response was reinforced during the 50% commission error condition, and every incorrect response was reinforced during the 100% commission error condition. The DV was percentage correct learner responses for each condition.

The results from the DiGennaro Reed et al. (2011) study showed that percentage correct responding was highest in the 0% commission error condition. An interesting finding from the study is that two out of three learners' performance in terms of correct responding was undifferentiated between the 50% and 100% commission error conditions. This finding would suggest that reinforcing some incorrect learner responses may be just as detrimental to learner performance as reinforcing every incorrect learner response. An interesting extension of the DiGennaro Reed et al. study could include additional conditions such as 20%, 40%, 60%, and 80% commission error conditions.

The results from the Dib and Sturmey (2007), Downs et al. (2008), Koegel et al. (1977), and Sarokoff and Sturmey (2008) studies indicated that learner behaviours (correct responses and engaging in problem behaviours) can be affected by the level of integrity with which DTT is administered. The combined results from these studies indicate that changes in learner behaviour were evident once therapists had been trained to administer DTT with at least 90% accuracy. Collectively, the research presented in this chapter suggests that initial DTT training is not enough to ensure that the procedure is administered with a high degree of integrity. The majority of the studies indicated that treatment integrity levels only increased above baseline levels following additional DTT-specific training and ongoing supervision. Downs et al. (2008) stated that in order for DTT programmes to be administered effectively therapists require "ongoing consultation with a professional who has significant experience implementing DTT and supervising DTT programming with students with disabilities," (p. 244).

DTT, Treatment Integrity, and Markov Probability Chains

A vast majority of the DTT studies described in this chapter presented treatment integrity scores as overall session integrity scores which are calculated across all components of the procedure. The findings from the studies referenced in this chapter have found that DTT studies generally do not provide component integrity data, and has identified the need for DTT studies to investigate and report component integrity scores. One way of investigating component integrity with DTT is to conduct a local analysis with continuously collected DTT data. Markov probability chains can be used to conduct such an analysis. Such an approach has not been taken previously.

A local analysis of DTT component integrity using Markov probability chains can lead to findings that have clinical utility. The obtained probability values for certain sequential events from a Markov chains analysis can be compared to expected probability values from what is prescribed in the DTT literature, and from how the DTT procedure has been programmed by the programme consultant. For example, a consequence should be followed by an ITI with a conditional probability of 1.00. If a Markov chains analysis is conducted and the observed conditional probability value of a CONSEQUENCE – ITI sequence is .70, it can be concluded that the probability of a consequence followed by an ITI is less than expected. One recommendation that can be made to the programme consultant is that the therapist should be retrained to administer ITI as prescribed.

Introductory Chapters I - IV Summary

There are few studies in the DTT literature where DTT data have been analysed on an event-by-event, or trial-by-trial basis. DTT data are usually analysed using whole-sessions methods such as percentage correct. These whole-sessions methods involve collecting data across the duration of the study and aggregating them into single-session bins. These methods are effective in displaying general patterns and trends within the data. However, they could be obscuring potentially important sequences occurring on an event-by-event or trial-by-trial basis. These sequences may have clinical value and could be used to improve the quality of the therapy that learners with intellectual disabilities receive as part of their regular teaching programmes. For example, they could be used to identify and quantify specific types of within-trial treatment integrity errors so that therapists can be retrained to administer the procedure according to the prescribed protocol. Previous research has shown functional relationships between within-trial treatment integrity errors and impaired learning (e.g. DiGennaro Reed, Reed, Baez, & Maguire, 2011).

Sequential analysis can be used to conduct these types of analyses with DTT data on both a within and between-trial basis to assess whether clinically useful results can be obtained. DTT is a procedure that involves several transitions between the various within-trial components, and also involves a large number of transitions between trials. Markov transition matrices appear well-suited for analyzing DTT data because the matrices are exhaustive; they can simultaneously display the frequencies and probabilities of all possible transitions between all components of the procedure. This feature is useful because it does not involve individual calculations for each transition. However, the clinical utility of analyzing DTT data using MTM has not been established or evaluated. This thesis will provide such an evaluation.

The thesis had four broad aims. These were: (1) Introduce Markov transition matrices as a method of assessing DTT data on a within-trial basis to identify within-trial treatment integrity errors so therapists could be retrained; (2) Use various sequential analysis methods to analyse DTT data on a between-trials basis to assess parts of the procedure in a manner that has not been done previously; (3) Draw conclusions about the quality of therapy that learners were receiving as part of their regular teaching programmes based on the results from these aforementioned analyses; and (4) Comment on the appropriateness of analyzing DTT data using sequential analysis (specifically Markov transition matrices).

Chapter V

Study 1

Introduction

One way to measure the quality of an intervention is to record the treatment integrity with which the intervention is administered. It is important to have effective methods of evaluating the treatment integrity of a procedure so any procedural errors that may be adversely affecting the quality of the intervention can be corrected. Research has showed that learning can be adversely affected when treatments are not administered as prescribed (e.g., DiGennaro Reed, Reed, Baez, & Maguire, 2011; Downs, Downs, & Rau, 2008; Koegel, Russo, & Rincover, 1977; Sarokoff & Sturmey, 2008; Vladescu, Carroll, Paden, & Kodak, 2012; Wilder, Atwell, & Wine, 2006).

Discrete-trial teaching (DTT) is often a keycomponent of early intensive behavioural interventions. The efficacy of DTT in teaching skills to learners with intellectual disabilities has been demonstrated on multiple occasions (e.g., Anderson, Taras, & O'Malley, 1996; Lovaas, 1987; McEachin, Smith, & Lovaas, 1993; Vismara & Rogers, 2010). In efficacy studies the procedures are carried out by researchers and therapists that have been trained to administer DTT with a high degree of treatment integrity (Eikeseth et al., 2012; Stock et al., 2013). The effectiveness of DTT in teaching skills to learners with intellectual disabilities in the natural environment is an issue of great importance in applied behaviour analysis. Studies evaluating the effectiveness of DTT in the natural environment often find that there is a lesser degree of emphasis placed on constantly evaluating treatment integrity (Bibby, Eikeseth, Martin, Mudford, & Reeves, 2002; Eikeseth et al., 2012; Magiati, Charman, & Howlin, 2007; 2007; Stock et al., 2013).

Traditionally, DTT within-trial treatment integrity studies reported data as percentage correct (e.g., Downs et al., 2008; Koegel et al., 1977; Vladescu et al., 2012).

Percentage correct is a whole-sessions analysis in which data collected across the duration of the study are aggregated into single-session bins (Fahmie & Hanley, 2008). Examples of DTT data analyses on an event-by-event basis are sparse (e.g., Belfiore, Fritts & Herman, 2008; Koegel, Glahn & Nieminen, 1978). Given that DTT is a procedure that consists of several components that follow each other sequentially, sequential analyses can be conducted to evaluate the treatment integrity with which the within-trial components of the procedure are being administered. Markov transition matrices (MTM) appear to be an appropriate method given exhaustive nature of the matrices; it displays the probability of all transitions between all events.

Aims of Study

The aim of the study was to use MTM to assess the within-trial treatment integrity of DTT teaching programmes of learners with intellectual disabilities. Such analyses allowed for a component analysis to be conducted so that errors affecting treatment integrity on a within-trial basis could be identified. It was of interest to evaluate whether using MTM would be an effective method of identifying problematic sequential dependencies which adversely affected treatment integrity.

Study 1 was an observational study. Therapy sessions, consisting of DTT, were video-recorded at the time and location where they normally occurred. The study consisted of two phases; a baseline phase and an intervention phase. The baseline phase consisted of identifying errors which affected the treatment integrity of the DTT programmes. Once the errors were identified consultants and therapists were notified of the results, and further follow up recordings were made (if possible) to address if the frequency and probability of the identified errors decreased, and whether treatment integrity increased.

Method

Participants

Recruitment and Consent. Ethics approval for this study was obtained on May 5th 2011 from the University of Auckland Human Participant's Ethics Committee (reference number 2011/160). Agencies and psychologists that provide consultancy services concerning DTT, who could refer therapists and learners to participate in the study, were approached by the researcher. Agencies and psychologists to be approached had to be Board Certified Behaviour Analysts © (BCBA) and meet the requirements set out in the Consumer Guidelines document by Autism SIG (www.autismpppsig.org). The consultants and psychologists approached prospective participants (therapists, families, and schools), and obtained consent. Learners were selected on the basis that they were receiving DTT as part of their regular teaching programmes. Five learners, eight therapists, and four consultants were recruited.

Dyads. Eight therapist-learner dyads participated in the study. Dyad 1 consisted of Therapist 1 and Learner 1, Dyad 2 of Therapist 2 and Learner 2, Dyad 3 of Therapist 3 and Learner 3, Dyad 4 of Therapist 4 and Learner 1, Dyad 5 of Learner 4 and Therapist 5, Dyad 6 of Learner 5 and Therapist 6, Dyad 7 of Learner 5 and Therapist 7, and Dyad 8 of Learner 5 and Therapist 8.

Learners. Learner 1 was an 8 year old non-verbal male who was diagnosed with autism at 22 months by a paediatrician. Learner 1 had been with the current service provider for approximately 12 months and had received therapy from an organisation that did not meet the requirements prescribed by Autism SIG for at least four years prior. Learner 1 received therapy during the week for approximately 2.5 hours daily from Monday to Saturday (15.5 hours per week). Learner 1 attended a special needs school during the week (Monday to

Friday). Acquisition S^D for Learner 1 included identifying objects, identifying locations within the home from pictures, identifying body parts, and matching items.

Learner 1 was also part of Dyad 4. Approximately 10 months had elapsed between Learner 1 being part of Dyad 1 and being part of Dyad 4. By this time, Learner 1 was 9 years old and received therapy four days (10.5 hours per week) a week (Monday, Tuesday, Thursday and Saturday). Learner 1's family had also relocated to a new residential address. The acquisition S^D consisted of identifying actions from photos, identifying printed numbers, identifying photos of people, identifying safety signs, object labels and identifying body parts.

Learner 2 was a 9 year old male diagnosed with global developmental delay by a developmental paediatrician at age 18 months. At the time of the study Learner 2 had been receiving therapy for 7.5 years. Learner 2 started school at age 6 and attended a special needs unit located within a mainstream school from Monday to Friday. Learner 2 received therapy every school day for approximately 1 to 2 hours. Acquisition S^D for Learner 2 were identifying the days of the week from printed words, identifying a picture of an object from among a range of pictures of objects, and counting how many pictures of an object were on the table.

Learner 3 was a 9 year old non-verbal male who was diagnosed with autism at age 18 months by a paediatrician. Learner 3 also received a diagnosis of epilepsy at age 7 months, with seizure onsets at 4 months. Learner 3 was enrolled at a mainstream school and attended school from Monday to Friday. Learner 3 had been receiving therapy since 3 years of age, and was receiving 12.5 hours of therapy per week. Acquisition S^D consisted of identifying colours, sight reading, identifying people from photos and, word discrimination.

Learner 4 was a 5 year old male who was referred to their current service provider by a developmental paediatrician. Learner 4 was diagnosed with epilepsy by a neurologist at age 36 months. His epilepsy medications were Clonazepam (0.18 ml three times per day) and Lamictal (15 mg twice a day). Learner 5 received 10.5 hours of therapy per week and attended a mainstream school for 3 hours 45 minutes per day for 5 days per week (although he was frequently absent due to his epilepsy seizures). Acquisition S^D consisted of identifying actions and nonverbal imitation.

Learner 5 was a 3.5 year old verbal female. Learner 5 was diagnosed with autism spectrum disorder at age 18 months by two paediatricians. Learner 5 had been receiving therapy since 19 months of age from the same service provider, and received approximately 24 hours per week of therapy. Acquisition S^D were identifying the function of various objects, identifying colours, identifying actions, identifying pronouns, matching pictures with objects, drawing, identifying emotional states, identifying locations, identifying object labels, identifying genders and categories, labeling body parts, vocal imitation, and quantitative concepts.

Therapists and Consultants. Therapists 1, 3, 4, 5, 6, 7, and 8 were employed by an autism service provider, and Therapist 2 was employed as a full-time teacher aid at the school that Learner 2 was attending. Therapists 1, 3, 5, 6, 7, and 8 had Master's degrees in psychology. Therapist 2 did not have a tertiary level qualification. Therapist 4 had a Bachelor's degree in psychology.

Therapist 1 had been employed as a therapist for approximately 12 months, and had been the therapist for Learner 1 for approximately a year. Therapist 2 had worked with Learner 2 on a part-time basis for 5 months, before working with him on a full-time basis for 1.5 years. Therapist 2 had been trained to administer DTT procedures by a PhD-level BCBA, who was also the consultant for the programme. Therapist 3 had approximately 27 months experience as a therapist, and had been the therapist for Learner 3 for approximately 16 months. Therapist 4 had been working with Learner 1 for approximately 2 years, and had 9 years' experience as a therapist. Therapist 5 had been a therapist for 15 months, and had been working with Learner 4 for two months. Therapist 6 had been working with Learner 5 for 21 months, and as a therapist for 21 months also. Therapist 7 had 9.5 years' experience working as a therapist, and had been the therapist for Learner 5 for 14 months. Therapist 7 was also the consultant for Learner 5. Therapist 8 had been working as a therapist for 15 months, and has been working with Learner 5 for 10 months.

During the course of the study Therapists 1, 3, 4, and 5 were receiving regular monthly supervision. Therapist 2 received no supervision during the course of the study. Therapists 6, 7, and 8 received regular supervision every 2 weeks.

All of the programme consultants were BCBA. The consultants for Dyads 1, 3, 4, 5, 6, 7, and 8 had Master's degrees in psychology, and the consultant for Dyad 2 had a PhD in psychology.

Equipment

The study was an observational study, and therapy sessions were video-recorded at the time and place where they normally took place. The video-recording equipment consisted of three high-definition Flip Video ™ flip cameras, and three camera tripod stands on which the cameras could be mounted. The flip cameras had a recording capacity of approximately one hour. Once the recordings were made, the footage was downloaded onto a PC computer for analysis with ObsWin 3.4 ™ (Martin, Oliver, & Hall, 1999) software, and visual basic software using Microsoft Excel 2010 ©. Videos were edited (when necessary) using Windows Live Movie Maker ©.

Setting

For Dyad 1, therapy took place in a room at the learner's home. The learner received DTT at numerous locations within the room. Some trials were administered on the floor, and other trials at a table. The learner and therapist sat opposite each other. No other people entered the room during the course of the sessions. Three cameras were used to record the therapy sessions for Dyad 1. One camera recorded all trials administered at the table. The other two cameras were placed in opposite corners of the room to record all trials administered on the floor. The researcher was not present in the room when recordings were made.

DTT for Dyad 2 took place at the learner's school in a room located in the school's special needs unit. DTT was conducted at a table and the therapist sat next to the learner. Other students and staff members were able to enter and exit the room. Two cameras were used. One camera was placed on the edge of the table directly in front of the learner and therapist. The other camera was stationed on a tripod directly behind the learner and therapist, and overlooked the table where DTT took place. The researcher was present in the room but was out of sight, and sat approximately 5 m directly behind the learner and therapist.

Therapy for Dyad 3 took place at the learner's home in a separate room. Some trials were administered with the learner sitting at a table, and other trials were administered when the learner was sitting on the floor. In both cases the therapist sat directly across from the learner. No other people entered the room. For the first six sessions, two cameras were used. One camera was placed on a tripod in the doorway to record all trials which were administered on the floor. The second camera was placed on a tripod next to the table to record all trials administered at the table. From the seventh recording onwards no more trials were administered on the floor. The camera that was positioned next to the table was no longer used, and the camera that was positioned in the doorway was repositioned to record trials administered at the table. The researcher sat in an adjacent room while DTT was taking place.

Therapy sessions for Dyad 4 took place at the learner's home in a separate room. All trials were administered on the floor, and the learner and the therapist sat directly across from each other. No other people entered the room during DTT. One camera was used to make the recordings. The camera was placed on top of a camera stand and overlooked the area where DTT trials were administered. The researcher sat in an adjacent room, but could be seen by the learner and the therapist.

Therapy sessions for Dyad 5 took place in the learner's home in the living room area. DTT was administered on the floor with the learner and therapist sitting opposite each other. Other people in the house could enter the room. During DTT, the researcher sat by a table in an adjacent room, and was fully visible to both the therapist and learner. One camera was used.

For Dyads 6, 7, and 8, DTT took place at the learner's home in the living room and in the learner's bedroom. When trials were administered in the living room, they were administered on the floor and at a table. When trials were administered in the learner's bedroom, they were administered on the floor. The therapist always sat directly across from the learner. One camera was used to record the trials. The researcher could be seen by the learner and therapist, and repositioned the camera as required.

Event Definitions

Events to be coded from the video-recordings were decided on after a review of the DTT literature (Chapter II) and in consultation with programme consultants. Therapist instructions were coded as either *SDMASTER* or *SDACQUIRE*. *SDMASTER* involved the presentation of an instruction that the learned had already mastered, and *SDACQUIRE* involved instructions which had not reached the mastery criterion. The mastery criterion for all programmes was greater than 80% correct learner responses across three consecutive sessions. In order to distinguish *SDMASTER* from *SDACQUIRE*, observers were given a list containing the different stimuli used for each DTT programme, outlining which had been mastered and which were in acquisition.

Four different types of response prompts were coded. These were verbal, gestural, full physical guidance, and shadow prompts. *VERBAL PROMPTS* were defined as any verbal utterances provided by the therapist to assist the learner with emitting a correct response that differed from the S^D. *GESTURAL PROMPTS* were defined as the therapist pointing with an extended index finger towards the correct stimulus. *PHYSICAL GUIDANCE PROMPTS* were defined as the therapist making physical contact with the learner in a manner that assisted them in emitting a correct response. To code S*HADOW PROMPTS*, the videos in which shadow prompting occurred were edited so that a vertical line was inserted, which represented the point where the hand of the therapist had been extended far enough to shadow the movements of the learner. *SHADOW PROMPTS* were defined as the left hand of the therapist fully crossing the vertical line.

Learner responses were scored as either *CORRECT* or *INCORRECT*. *CORRECT* responses were defined as the learner emitting a response consistent with therapist instructions, and *INCORRECT* responses were defined as a learner emitting a response not consistent with therapist instructions. Incorrect responses included trials in which the learner made no responses.

Therapists also provided several different types of consequences. *ERROR CONSEQUENCES* were defined as the therapist making any kind of verbal or nonverbal indication to the learner that a response was incorrect, or removing all stimuli from workspace without providing a consequence once an incorrect response was made. *CORRECT CONSEQUENCES* included tokens, social praise, edibles, access to tangible items such as toys, access to a DVD player, and access to their picture exchange communication system folders (PECS; Bondy & Frost, 2002), from which they could requests items or activities. *TOKENS* were defined as tokens making contact with the token board. *SOCIAL PRAISE* was defined as the therapist providing any kind of verbal utterance following a learner response to indicate to the learner that the response was correct. These included high fives, and singing songs. *EDIBLE* was defined as the learner making physical contact with an edible item that the therapist had provided following a learner response. *PECS* was defined as the therapist presenting the learner with their PECS folder at the end of a DTT programme. *DVD PLAYER* was defined as either the learner or therapist switching on the DVD player at the end of a DTT programme. *TANGIBLE* was defined as the learner making physical contact with an item (i.e., toys and books) provided by the therapist following a response.

The variables *SOCIAL PRAISE, TOKEN*, *EDIBLE*, and *TANGIBLE* were combined to form the variable *CORRECT CONSEQUENCE*. *CORRECT CONSEQUENCE* was scored once per trial at the offset of the last consequence they received. For example, if they received a token and social praise, and the social praise occurred after the token, then *CORRECT CONSEQUENCE* was coded at the offset of the social praise.

The variable *ITI* was coded if the time between the offset of the last consequence, and the start of the next trial was at least 3 seconds for Dyads 1, 2, and 3, at least .5 seconds for Dyad 4, and at least 1 second for Dyads 5, 6, 7, and 8. The lengths of the *ITI* were specified by the programme consultants.

The variables *SDMASTER*, *SDACQUIRE*, *VERBAL PROMPT, GESTURAL PROMPT*, *PHYSICAL GUIDANCE PROMPT*, *SHADOW PROMPTS, CORRECT, INCORRECT*, *ERROR CONSEQUENCE, EDIBLE, TOKENS, PECS*, and *DVD PLAYER* were always defined as discrete events, and were scored as soon as they occurred.

SOCIAL PRAISE was scored as a discrete event for all dyads except Dyad 3, for which it was scored as an event with duration. When scored as a discrete event, it was scored as soon as the therapist had stopped providing social praise. If the social praise had not stopped by the time the next S^D was delivered, the offset of the event was coded at the onset of the next S^D. When scored as a duration measure, the onset was coded as soon as the therapist had made a verbal utterance following a learner response, and the offset was scored as soon as the therapist ceased providing social praise. If social praise had not stopped by the time the next S^D was delivered it was coded at the onset of the next S^D. *TANGIBLE* was also scored as an event with duration for Dyad 3. The onset was scored as soon as the learner made physical contact with the item, and the offset was scored as soon as the therapist had removed the item.

Measurement

From the video recordings, all components of discrete trials were coded and time-stamped using Obswin 3.4 software. The software allowed video-recordings to be stopped, replayed, slowed down, and sped up. Each event to be coded was designated a key on a computer keyboard which the observer had to press in order to score the event.

For Dyad 1, nine sessions were video-recorded and used in the analyses. Recordings were made over the course of five weeks. For Dyad 2, 22 recordings were made, and 19 were suitable for coding and analyses. The first 16 recordings were made over a 10 week period. A further three sessions were recorded six months following the 16th recording, and these were recorded over a period of five days. For Dyad 3, 18 recordings were made and analysed. Learner 3 was in hospital for approximately three weeks between the third and fourth recordings. The first 15 recordings were made over eight weeks. Three more recordings were made six months later, and these were recorded over a period of eight days.

For Dyad 4, six sessions were recorded and used in the analyses. The recordings were made over four weeks. For Dyad 5, two sessions were recorded and analysed. Learner 5 was withdrawn from the study following the second recording session due to health complications. For Dyad 6, 13 sessions were video-recorded and used in the analyses. Video-recording took approximately three and a half months to complete. For Dyad 7, nine sessions were video-recorded, and used in the analyses. It took approximately 6 months to record all of the sessions for Dyad 7. For Dyad 8, 13 sessions were video-recorded, and used in the analyses, and it took approximately five months to complete the recordings.

Interobserver Agreement

Interobserver agreement (IOA) scores were calculated using ObsWin 3.4 software. IOA scores were calculated for 22% of total sessions for Dyad 1, 16% of sessions for Dyad 2, 17% of sessions for Dyad 3, 33% of sessions for Dyad 4, 50% of sessions for Dyad 5, 15% of sessions for Dyad 6, 22% of sessions for Dyad 7, and 15% of sessions for Dyad 8. IOA scores were calculated using time-window analysis with a tolerance of +/- 2 s (MacLean, Tapp, & Johnson, 1985), and reported as a percentage. An agreement was scored if the same event was coded within 2 s. A disagreement was scored if two different events were scored within 2 s, or when the same event was scored, but was scored more than 2 s apart. Percentage agreement was calculated by dividing the number of agreements by the total number of agreements and disagreements. The resulting ratio was multiplied by 100% (Mudford, Martin, Hui, & Taylor, 2009).

Some videos were excluded from the IOA process because they were used to train observers. Videos to be coded for IOA were randomly selected from all available recordings. The remaining video-recordings were allocated a number, and the Microsoft Excel © function *RANDBETWEEN* was used to randomly select which videos were to be used in the IOA process. The researcher was not present in the room while the observers coded the videos and was located in an adjacent room. Observers could contact the researcher via mobile telephone during the course of a coding session if they needed assistance (e.g., observers accidently pressed the F9 key which immediately stopped the video that was playing. The researcher had to restart the video at the point which it had stopped so that the IOA process could resume). Observers also had a pen and paper with which they could take notes.

Observers. The criteria for selecting observers to code the data for Dyads 1, 3, 4, 5, 6, 7, and 8 were that they had no prior experience administrating DTT procedures, and that they were not currently employed by any autism service provider (as requested by consultants for these dyads). No such restrictions were in place when recruiting observers to code events for Dyad 2.

Eleven observers were recruited. Four of the observers were recruited from Stage 2 and 3 courses in the experimental analysis of behaviour from the University of Auckland. Three PhD students from the University of Auckland's experimental analysis of behaviour research unit, and one PhD student (who was also a BCBA) from the University of Auckland's applied behaviour analysis programme were recruited. The two other observers had not previously been enrolled in any behaviour analytic courses, but had been enrolled in a Stage 1 Psychology paper at the University of Auckland. One other observer was recruited from the University of Auckland, and had not attended any Psychology courses. None of the observers, except the BCBA PhD student, had any prior experience coding events.

Observer Training. Observer training consisted of familiarising observers with written definitions of events, and showing sections from the video-recordings which contained these events. Instances of events to be coded were pointed out to observers by stopping the video once events of interest had occurred. Observers had the opportunity to ask questions throughout the initial training phase. Once observers were trained with the definitions, they were trained to use the software. Once the functions of the programme were explained, a training video was loaded, and observers got the opportunity to use the programme.

Once observers had been familiarised with the events to be coded and the software used to code events, they were required to code other training videos for which IOA scores were calculated. If the IOA scores met the training criterion observers were considered trained and proceeded to the IOA proper. The training criterion was a score of at least 80% agreement for the occurrence of an event for at least one training video. During the observer training phase the researcher was present in the room where the coding took place.

Observer Retraining. In the event that the IOA scores fell below 80% agreement, the data were studied to determine the cause of the disagreements. Initially, observers were given the opportunity to recode the relevant sections where disagreements occurred following brief retraining. The data from these recodings were used in the final IOA calculations. If IOA remained below 80% agreement following retraining the following remedial measure was taken; data were studied to determine if the low levels of agreement resulted from observer errors or definitional errors. Observational errors resulted if the observer coded events in a manner not consistent with the definition for a given event. In that case the observers were retrained using the observer training methods described earlier. Once retraining had occurred the videos were recoded. Definitional errors resulted if the description of an event was not sufficiently accurate. In the case of a definitional error the definition for that event was re-written, recordings re-coded with the new definition, and the observer retrained with the new definition.

Video-Editing

Once the recordings had been made, they were edited (if needed) using Windows Live Movie Maker ©. Videos for Dyads 2 and 3 were edited so that recordings began as soon as the first trial from the first lesson was administered, and finished as soon as the last trial from the last lesson had been administered. The videos from Dyad 1 were edited so that different sections of footage from each camera angle were combined to produce one video containing footage from the best angle from which to code the events. The sequential order of the trials was maintained.

Further editing for Dyad 2 consisted of editing in the vertical line needed to code shadow prompts. The videos from Dyad 4, 5, and 6 did not require editing. One recording was edited for Dyad 7 (the ninth recording). The flip camera was switched off at the end of a DTT programme and switched on at the start of the next DTT programme (at the request of the therapist), resulting in nine segments of video. Video-editing consisted of combining these nine segments into one video while maintaining the sequential order of the trials. During one recording session for Dyad 8 (the fifth recording), a new therapist took part in the session as part of their training. The new therapist had not consented to be video-recorded, and only trials that were administered by Therapist 8 were recorded. Video-editing consisted of combining all video-segments into one video while maintaining the sequential order of the trials.

Study Design

The study consisted of two phases: a baseline phase and a follow-up phase. All dyads participated in the baseline phase. Once the baseline data had been analysed, and treatment integrity errors identified, the results were reported to the therapists and consultants in a meeting. The researcher reported the errors to the therapist and consultant, and the consultant decided if further retraining was required. Therapists and consultants were also emailed a summary of the meeting outlining the treatment integrity errors that were discussed at the meeting. None of the consultants reported that any additional therapist retraining took place following the meeting and the emails.

Dyads 4, 6, 7, and 8 participated in a follow-up phase. The purpose of the follow-up phase was to further evaluate the treatment integrity errors that were identified in the baseline phase and investigate whether new errors could be identified. Follow-up phase recordings started between one and four weeks following the meeting with the therapists and consultants. No follow-up phase recordings were made for Dyads 1 and 3 because the therapists were no longer employed with the service-provider. No follow-up phase recordings were made for Dyads 2 and 5 because of learner health complications.

Social Validity

Therapists and consultants completed a social validity questionnaire once all video-recordings and data analyses for all dyads had finished. Five therapists and two consultants completed the questionnaires; one version of the questionnaire was for the therapists, and the other version was for the consultants. Social validity questionnaires for the consultants can be seen in Appendix A, and the questionnaire for the therapists in Appendix B. The questionnaires contained nine items for the therapists, and seven for the consultants. The consultant questionnaire differed from the therapist one, in that some questions were only relative to the therapists, and the wording for some of the questions on the consultant version was different. The questionnaires contained six ratings ranging from *strongly disagree* to *strongly agree*.

Analyses

All analyses were conducted using Obswin 3.4 software and Microsoft Excel 2010 ©. Markov transition matrices were used as the main method of analysis to identify within-trial treatment integrity errors. Whole-sessions methods such as percentage correct and blocks-of-trials were used to obtain treatment integrity scores for all mastered trials, all acquisition trials, and for the overall treatment integrity scores (mastered and acquisition trials). The conditional probabilities of selected errors were calculated for all dyads, and plotted on a blocks-of-trials basis.

Results

Interobserver Agreement

The IOA scores for all dyads can be seen in Table 5.1. Table 5.1 shows the mean percentage agreement and range of percentage agreement scores of each variable across all dyads. For all dyads the mean percentage agreement scores were above 80%. The mean percentage agreement score across all variables and all dyads was 93.77% with a range of 81.12% to 100%.

Treatment Integrity Error Analyses

Treatment integrity errors were identified by constructing Markov transition matrices. The transition matrices (Tables 5.2 – 5.13) show the probabilities of the sequences, and the number of times the sequences were observed (in brackets). Probability values were calculated by dividing the number of times a transition occurred by the total number of times the first event in the sequence occurred. Due to the large amount of data output a $p = .05$ (5%) criterion (henceforth to be referred to as the .05 criterion) was established by which treatment integrity errors were identified. The .05 criterion required that error sequences had to occur with a probability of at least .05 for them to be identified. In Tables 5.2 – 5.13 treatment integrity error sequences that reached the .05 criterion were bold-faced.

The probabilities of certain error sequences were investigated by plotting their probabilities per blocks of 15 events. For example, if the error sequence of interest was INCORRECT – CORRECT the probability of this error sequence was calculated every time an incorrect learner response had occurred 15 times. Plotting the probabilities like this allows for an investigation of how the probability of these error sequences changed over time while keeping the number of times the first event in the sequence occurred constant. This allows for a fair comparison of the probabilities between blocks as the error sequence has an equal

Table 5.1

Interobserver agreement results for all dyads.

Dyad	Event	Mean	Range
1	SDMASTER	97.13	97.37 – 96.88
	SDACQUIRE	100	100 – 100
	GESTURAL	82.86	85.71 – 80.00
	PHYSICAL	93.75	100 – 87.50
	CORRECT	96.04	97.44 – 94.64
	INCORRECT	96.43	100 – 92.86
	ERROR CONSEQUENCE	94.45	100 – 88.89
	SOCIAL	88.14	90.57 – 85.71
	TOKEN	100	100 - 100
2	SDMASTER	94.91	97.30 – 92.31
	SDACQUIRE	93.21	96.97 – 90.12
	SHADOW	86.11	90.32 – 80.00
	VERBAL	90.54	93.33 – 85.42
	PHYSICAL	89.87	95.00 – 84.62
	CORRECT	96.37	100 – 93.42
	INCORRECT	86.83	96.23 – 81.25
	SOCIAL	97.82	100 – 96.15
	TOKEN	99.11	100 – 98.75
	DVD	100	100 – 100
3	SDMASTER	98.15	100 – 94.44
	SDACQUIRE	99.17	100 – 97.50
	GESTURAL	96.08	100 – 88.24
	CORRECT	100	100 – 100
	INCORRECT	95.24	100 – 85.71
	ERROR CONSEQUENCE	90.11	100 – 84.62
	EDIBLE	100	100 – 100
	TANGIBLE	94.46	98.78 – 86.92
	SOCIAL	86.12	89.44 – 83.40
	PECS	100	100
4	SDMASTER	97.43	95.83 – 99.03
	SDACQUIRE	96.48	92.96 – 100
	PHYSICAL	100	100 – 100
	GESTURAL	93.64	87.27 – 100
	CORRECT	97.64	96.06 – 99.21
	INCORRECT	91.92	91.84 -92.00
	ERROR CONSEQUENCE	93.75	87.50 –100
	SOCIAL	95.99	94.31 – 97.66

Table 5.1 (continued)

Dyad	Event	Mean	Range
5	SDMASTER	99.44	-
	SDACQUIRE	95.35	-
	GESTURAL	100	-
	CORRECT	96.15	-
	INCORRECT	93.75	-
	ERROR CONSEQUENCE	92.31	-
	SOCIAL	86.54	-
	TOKEN	100	-
6	SDMASTER	82.50	77.50 – 87.50
	SDACQUIRE	94.41	91.67 – 97.14
	GESTURAL	85.71	85.71
	VERBAL	92.86	85.71 – 100
	CORRECT	94.69	94.55 – 94.83
	INCORRECT	85.71	85.71
	ERROR CONSEQUENCE	100	100
	SOCIAL	88.46	80.77- 96.15
7	SDMASTER	89.21	87.50 – 90.91
	SDACQUIRE	95.48	95.12 – 95.83
	GESTURAL	100	100
	PHYSICAL	100	100
	VERBAL	96.00	96.00
	CORRECT	96.47	92.94 – 100
	INCORRECT	89.19	84.62 – 93.75
	ERROR CONSEQUENCE	87.50	75.00 – 100
	SOCIAL	81.12	79.78 – 82.46
8	SDMASTER	87.16	80.77 – 93.55
	SDACQUIRE	93.40	86.75 – 100
	MODEL	96.43	92.86 – 100
	PHYSICAL	90.00	80.00 – 100
	VERBAL	90.00	80.00 – 100
	GESTURAL	100	100
	CORRECT	89.81	84.75 – 94.87
	INCORRECT	88.89	77.78 – 100
	ERROR CONSEQUENCE	100	100
	SOCIAL	83.40	80.00 – 86.79

number of opportunities to occur between blocks. Therefore, when calculating the conditional probability of an error, the denominator remains constant across blocks (i.e., 15).

Baseline Phase

Table 5.2 contains the baseline results for Dyad 1, and eight two-event error sequences were identified. The eight error sequences can be separated into three groups; errors where the therapist did not administer the prescribed error consequence following incorrect learner responses, the ITI not lasting for its prescribed minimum duration of 3 s, and learner self-corrections.

The error sequences involving the therapist not administering the prescribed error consequence following incorrect learner responses will be described first. The prescribed error consequence required the therapist to provide any kind of verbal utterance to indicate to the learner that the response was incorrect (e.g., *no* or *try again*). As seen in Table 5.2 the probability of incorrect learner responses being followed by the prescribed error consequence was .28 (*n* = 25) indicating that incorrect learner responses were more likely to be followed by events that were not part of the prescribed error correction procedure. The two most likely error sequences were full physical guidance prompts (INCORRECT – PHYSICAL, *n* = 21, p = .24), and gestural prompts (INCORRECT – GESTURAL, *n* = 12, p = .14). These sequences indicate that the therapist was likely to administer response prompts instead of the prescribed error consequence following incorrect learner responses.

The second group of error sequences involved the ITI not lasting for its prescribed minimum duration. These sequences were CORRECT CONSEQUENCE – SDMASTER (*n* = 80, p = .20), ERROR CONSEQUENCE – SDMASTER (*n* = 11, p = .24), CORRECT CONSEQUENCE – SDACQUIRE (*n* = 64, p = .24), and ERROR CONSEQUENCE – SDACQUIRE (*n* = 9, p = .20). As seen in Table 5.2 these error sequences were more likely to occur when the S^D for the next trial was a mastered S^D than if the next S^D was an acquisition S^D. This finding was consistent across both correct and error consequences.

Table 5.2

The Dyad 1 baseline phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All the cells that are bold-faced are the treatment integrity errors that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	PHYSICAL	CORRECT
SDMASTER	.00 (1)	.00 (0)	.01 (2)	.01 (2)	.84 (221)
SDACQUIRE	.01 (3)	.00 (01)	.06 (13)	.03 (7)	.68 (147)
GESTURAL	.16 (13)	.60 (49)	.00 (0)	.01 (1)	.20 (16)
PHYSICAL	.18 (40)	.37 (21)	.02 (1)	.02 (1)	.40 (23)
CORRECT	.02 (10)	.01 (4)	.00 (2)	.01 (3)	.00 (0)
INCORRECT	**.06 (5)**	.03 (3)	**.14 (12)**	**.24 (21)**	**.19 (17)**
ERROR CONSEQUENCE	**.24 (11)**	**.20 (9)**	.09 (4)	.13 (6)	.00 (0)
CORRECT CONSEQUENCE	**.20 (80)**	**.16 (64)**	.06 (26)	.02 (8)	.01 (3)
ITI	.58 (131)	.29 (65)	.10 (22)	.04 (8)	.00 (0)

Table 5.2 (continued)

	INCORRECT	ERROR CONSEQUENCE	CORRECT CONSEQUENCE	ITI	SUM
SDMASTER	.14 (36)	.00 (0)	.01 (2)	.00 (0)	1.00 (263)
SDACQUIRE	.21 (45)	.00 (0)	.00 (1)	.00 (0)	1.00 (217)
GESTURAL	.01 (1)	.01 (1)	.01 (1)	.00 (0)	1.00 (82)
PHYSICAL	.00 (0)	.00 (0)	.02 (1)	.00 (0)	1.00 (57)
CORRECT	.01 (3)	.02 (9)	.93 (396)	.00 (0)	1.00 (427)
INCORRECT	.03 (3)	.28 (25)	.02 (2)	.00 (0)	1.00 (88)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.02 (1)	.31 (14)	1.00 (45)
CORRECT CONSEQUENCE	.00 (0)	.02 (10)	.00 (0)	.53 (213)	1.00 (404)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (226)
SUM					1809

The final two-event error sequence that reached the .05 criterion was INCORRECT – CORRECT indicating that the learner self-corrected following incorrect responses. Learner self-corrections were observed 17 times and the probability was .19.

Figure 5.1 contains the probabilities per blocks of 15 events for the six errors that were observed most frequently. As seen in Figure 5.1 the probability of the INCORRECT – CORRECT sequence increased from .08 to .27 across blocks. It was the only error of the six

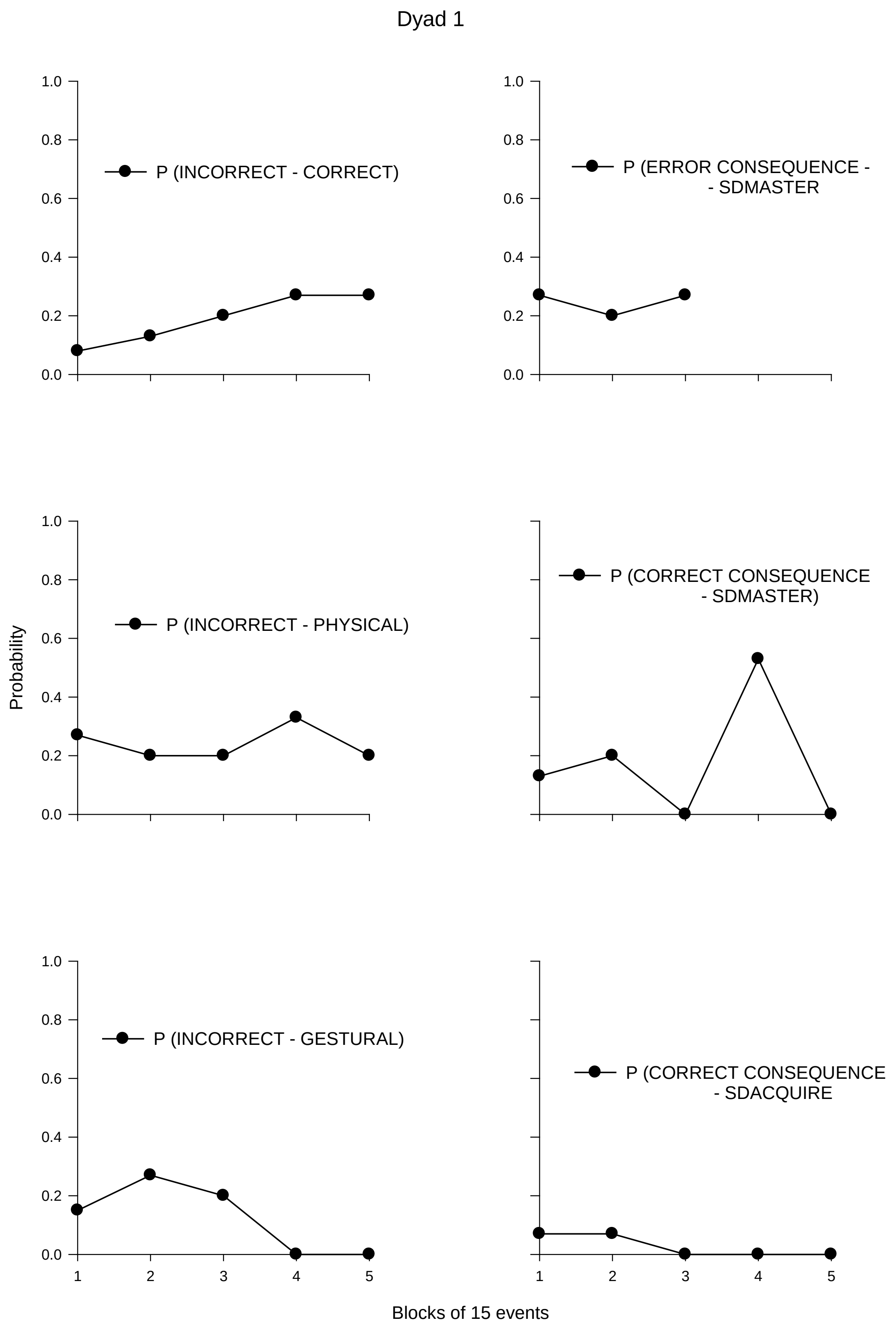

Figure 5.1: The probability of selected treatment integrity errors for Dyad 1 for the last 5 blocks of 15 events across both the baseline and follow-up phases which showed this increasing trend. The sequences INCORRECT – GESTURAL and CORRECT CONSEQUENCE – SDACQUIRE showed a decreasing trend. The probabilities of the sequence INCORRECT – PHYSICAL and ERROR CONSEQUENCE – SDMASTER appeared stable whereas the probability of the CORRECT CONSEQUENCE – SDMASTER sequence varied between .00 and .53. The errors that were reported to the consultant were the error sequences involving no ITI, and the prescribed error consequence not following incorrect learner responses.

Table 5.3 shows the baseline results for Dyad 2. Fourteen two-event error sequences reached the .05 criterion. Six of the errors involved response prompts, four involved not administering the prescribed error consequence following incorrect learner responses, two errors consisted of multiple S^D presentations, and two errors involved the ITI not lasting for its minimum prescribed duration of 3 s.

The first group of errors to be described are the sequences consisting of response prompt errors. The sequences SHADOW – SDMASTER (*n* = 237, p = .46), and SHADOW – SDACQUIRE (*n* = 230, p = .45) contains multiple errors. Shadow prompts were not a prescribed part of this learner's DTT programme yet 513 shadow prompts were administered. Furthermore, shadow prompts were likely to precede the S^D. Prompts were prescribed to be administered simultaneously with the S^D or in the period following the S^D and preceding learner responses.

Four sequences consisting of multiple successive response prompt presentations were identified. These sequences were PHYSICAL – PHYSICAL (*n* = 13, p = .05), PHYSICAL – VERBAL (*n* = 88, p = .37), VERBAL – PHYSICAL (*n* = 54, p = .12), and VERBAL – VERBAL (*n* = 105, p = .23). These sequences indicate integrity errors because prompts were

Table 5.3

The baseline phase results for Dyad 2. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All the cells that are bold-faced are the treatment integrity errors that reached the .05 criterion.

	SDMASTER	SDACQUIRE	SHADOW	PHYSICAL	VERBAL
SDMASTER	**.10 (118)**	.00 (1)	25 (.02)	.05 (56)	.05 (57)
SDACQUIRE	.01 (6)	**.05 (25)**	.04 (19)	.11 (57)	.08 (41)
SHADOW	**.46 (237)**	**.45 (230)**	.00 (2)	.02 (8)	.00 (2)
PHYSICAL	.04 (9)	.06 (15	.01 (2)	**.05 (13)**	**.37 (88)**
VERBAL	.04 (18)	.02 (11)	.03 (13)	**.12 (54)**	**.23 (105)**
CORRECT	.05 (72)	.01 (14)	.01 (20)	.00 (1)	.00 (3)
INCORRECT	**.18 (71)**	**.05 (20)**	.03 (11)	**.12 (48)**	**.38 (152)**
CORRECT CONSEQUENCE	**.19 (256)**	.03 (37)	**.12 (159)**	.00 (2)	.00 (0)
ITI	.49 (434)	.21 (181)	.30 (262)	.00 (2)	.00 (0)

Table 5.3 (continued)

	CORRECT	INCORRECT	CORRECT CONSEQUENCE	ITI	SUM
SDMASTER	.65 (789)	.11 (140)	.03 (36)	.00 (0)	1.00 (1222)
SDACQUIRE	.36 (193)	.34 (182)	.02 (11)	.00 (0)	1.00 (534)
SHADOW	.04 (21)	.03 (13)	.00 (0)	.00 (0)	1.00 (513)
PHYSICAL	.39 (94)	.08 (20)	.00 (0)	.00 (0)	1.00 (241)
VERBAL	.43 (192)	.09 (39)	.04 (16)	.00 (0)	1.00 (448)
CORRECT	.00 (1)	.00 (3)	.92 (1288)	.00 (0)	1.00 (1402)
INCORRECT	**.22 (87)**	.01 (5)	.02 (9)	.00 (0)	1.00 (403)
CORRECT CONSEQUENCE	.02 (28)	.00 (1)	.00 (0)	.64 (877)	1.00 (1360)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (879)
SUM					7002

to be administered in a least-to-most fashion across trials. The results indicate that prompts were administered unsystematically within-trials.

The second group of errors involved events following incorrect learner responses. The prescribed error consequence was that the therapist should remove all stimuli from the table following incorrect learner responses. As seen in Table 5.3 the prescribed error consequence was never administered. The most likely events following incorrect learner responses were verbal prompts (INCORRECT – VERBAL, *n* = 152, p = .38), learner self-corrections (INCORRECT – CORRECT, *n* = 87, p = .22), and the presentation of a mastered S^D (INCORRECT – SDMASTER, *n* = 71, p = .18).

The next group of error sequences consisted of multiple S^D presentations, as indicated by the sequences SDMASTER – SDMASTER (*n* = 118, p = .10), and SDACQUIRE – SDACQUIRE (*n* = 25, p = .05). The last group of error sequences consisted of the ITI not lasting for its prescribed minimum duration. Similar to Dyad 1 the probability of correct consequences not being followed by an ITI of minimum prescribed duration was higher when the following S^D was a mastered S^D (CORRECT CONSEQUENCE – SDMASTER, *n* = 256, p = .19). The other error sequence involved correct consequences being followed by shadow prompts (CORRECT – SHADOW, *n* = 159, p = .12). In this case shadow prompts indicated the start of the next trial.

Figure 5.2 contains the probabilities of the four most frequently observed error sequences (in terms of observed frequency) for the last 5 blocks of 15 events. Figure 5.2 the only error sequence that shows a decreasing trend is INCORRECT – VERBAL decreasing from .73 to .13. The probability of the SHADOW – SDMASTER sequences appeared stable and the probabilities for the CORRECT CONSEQUENCE – SDMASTER and CORRECT CONSEQUENCE – SHADOW sequences varied.

The errors that were reported to the consultant were the sequences involving the response prompt errors. It was suggested that the therapist be retrained on how to administer prompts in a manner consistent to what had been prescribed.

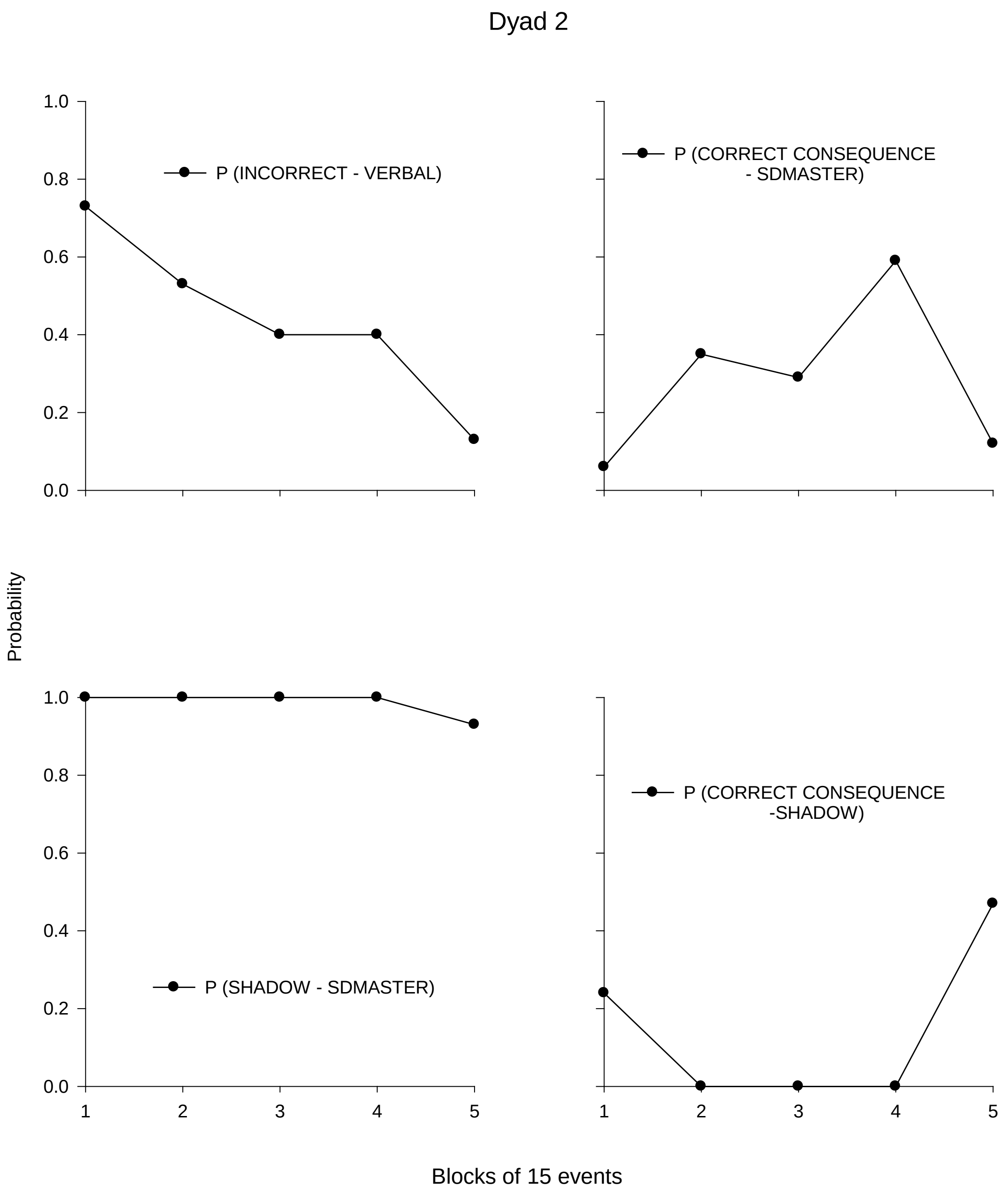

Figure 5.2: The probability of selected treatment integrity errors for Dyad 2 for the last 5 blocks of 15 events across both the baseline and follow-up phases.

Baseline results for Dyad 3 can be seen in Table 5.4, and four error sequences reached the .05 criterion. The first error sequence consists of two successive within-trial presentations of the gestural prompt (GESTURAL – GESTURAL, n = 11, p = .05). Gestural

Table 5.4

The baseline phase results for Dyad 3. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All the cells that are bold-faced are the treatment integrity errors that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	CORRECT	INCORRECT
SDMASTER	.00 (2)	.00 (0)	.04 (23)	.68 (375)	.22 (121)
SDACQUIRE	.00 (0)	.00 (2)	.21 (132)	.52 (320)	.22 (135)
GESTURAL	.06 (15)	.21 (50)	**.05 (11)**	.63 (154)	.02 (5)
CORRECT	.02 (15)	.01 (11)	.00 (1)	.00 (0)	.00 (3)
INCORRECT	.03 (8)	.00 (1)	.04 (11)	**.06 (16)**	.01 (4)
ERROR CONSEQUENCE	**.05 (13)**	.00 (0)	.00 (1)	.00 (0)	.00 (0)
CORRECT CONSEQUENCE	**.06 (50)**	.00 (4)	.00 (2)	.03 (27)	.00 (3)
ITI	.42 (448)	.52 (546)	.06 (62)	.00 (0)	.00 (0)

Table 5.4 (continued)

	ERROR CONSEQUENCE	CORRECT CONSEQUENCE	ITI	SUM
SDMASTER	.03 (19)	.02 (11)	.00 (0)	1.00 (551)
SDACQUIRE	.02 (10)	.03 (16)	.00 (0)	1.00 (615)
GESTURAL	.02 (5)	.01 (3)	.00 (0)	1.00 (243)
CORRECT	.00 (3)	.96 (859)	.00 (0)	1.00 (892)
INCORRECT	.85 (230)	.00 (0)	.00 (1)	1.00 (271)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.95 (253)	1.00 (267)
CORRECT CONSEQUENCE	.00 (0)	.00 (0)	.90 (803)	1.00 (889)
ITI	.00 (0)	.00 (1)	.00 (0)	1.00 (1057)
SUM				4785

prompts were prescribed to be administered once per trial. Learner self-corrections were observed 16 times with a probability of .06.

The error sequences ERROR CONSEQUENCE – SDMASTER (*n* = 13, p = .05), and CORRECT CONSEQUENCE – SDMASTER (*n* = 50, p = .06) indicated that during some trials the ITI was not administered for its prescribed minimum duration of 3 s. As was the case with Dyads 1 and 2 the probability of the ITI not lasting for its minimum prescribed duration was greater when the next S^D was a mastered S^D.

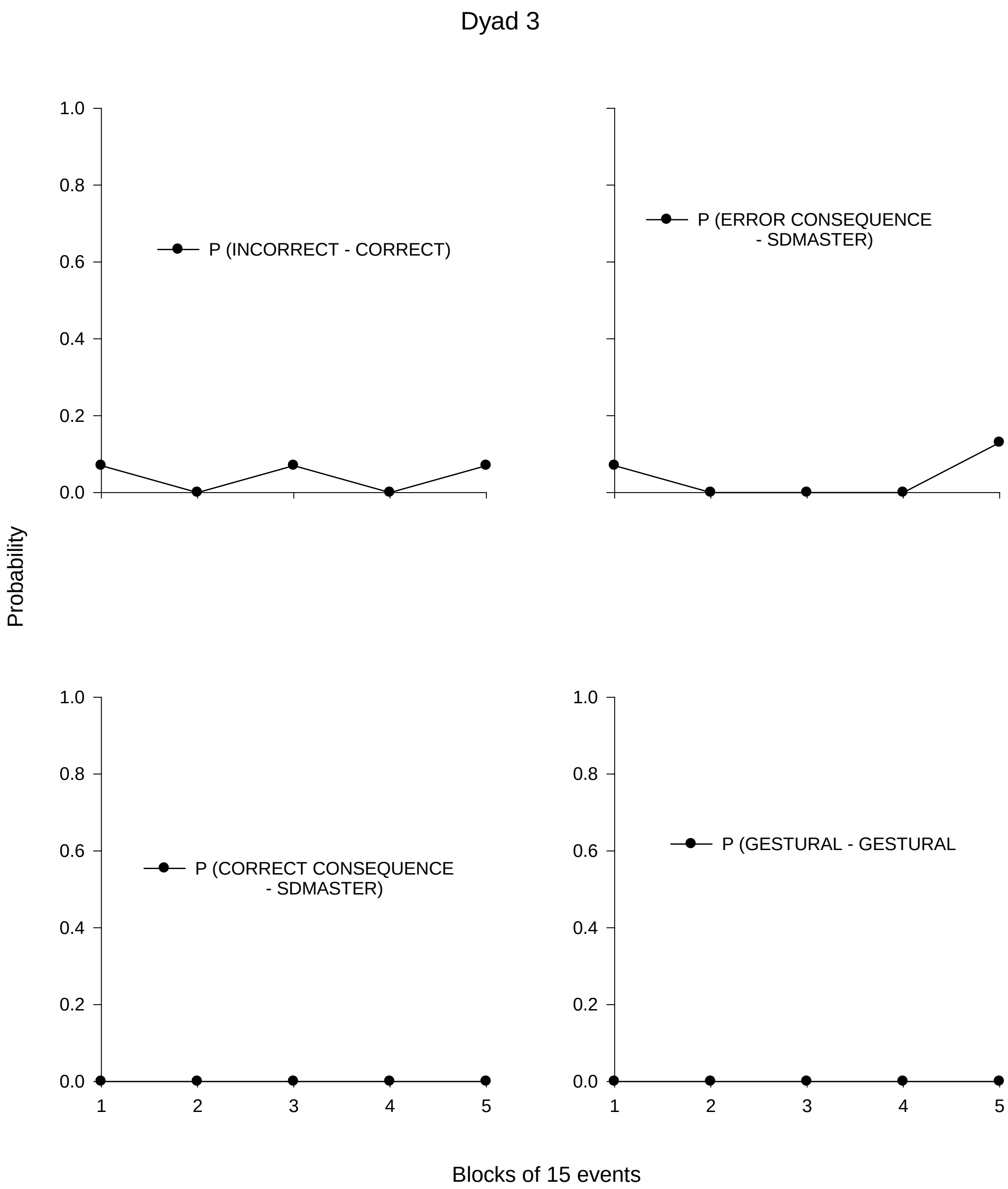

Figure 5.3: The probability of selected treatment integrity errors for Dyad 3 for the last 5 blocks of 15 events across both the baseline and follow-up phases.

The probabilities of the error sequences can be seen in Figure 5.3. None of the graphs showed any increasing or decreasing trends. The probabilities of the CORRECT CONSEQUENCE – SDMASTER and GESTURAL – GESTURAL sequences remained constant at .00 for the last five blocks of 15 events. The probability of learner self-corrections varied between .00 and .07, and between .00 and .13 for the ERROR CONSEQUENCE – SDMASTER sequence.

The results from Table 5.4 and Figure 5.3 show that treatment integrity was affected by errors that occurred infrequently. The meeting with the programme consultant and therapist consisted of reporting the learner self-correction error.

Table 5.5 displays the treatment integrity errors that were identified during the baseline phase for Dyad 4. Three two-event error sequences were identified and involved incorrect learner responses and error consequences. Learner self-corrections (INCORRECT – CORRECT) occurred 8 times with a probability of .10, and gestural prompts followed incorrect learner responses 9 times with a probability of .11 (INCORRECT – GESTURAL). The prescribed event following incorrect learner responses was the error consequence and consisted of any verbal utterance to indicate to the learner that the response was incorrect. The error consequence had to be followed by an ITI of 0.5 s minimum duration. The error sequence ERROR CONSEQUENCE – GESTURAL (n = 20, p = .30) reached the .05 criterion. Given that the minimum ITI duration was 0.5 s the gestural prompt and the error consequence were administered simultaneously. All three error sequences were reported to the programme consultant and therapist.

As seen in Table 5.6 four two-event error sequences reached the .05 criterion for Dyad 5. Two of the error sequences consisted of the therapist not administering the prescribed error consequence following incorrect learner responses. These sequences were INCORRECT – SDMASTER (n = 2, p = .06), and INCORRECT – SDACQUIRE (n = 2, p = .06). The prescribed error consequence for Dyad 5 was the same as for Dyad 1. The probability of the prescribed error consequence following incorrect learner responses was .64.

Table 5.5

The Dyad 4 baseline phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	CORRECT
SDMASTER	.00 (0)	.00 (0)	.01 (2)	.86 (164)
SDACQUIRE	.00 (1)	.00 (1)	.08 (19)	.67 (154)
GESTURAL	.11 (12)	.54 (61)	.00 (0)	.32 (37)
CORRECT	.00 (1)	.02 (7)	.00 (0)	.00 (0)
INCORRECT	.00 (0)	.00 (0)	**.11 (9)**	**.10 (8)**
ERROR CONSEQUENCE	.00 (0)	.00 (0)	**.30 (20)**	.00 (0)
CORRECT CONSEQUENCE	.01 (5)	.01 (5)	.01 (2)	.00 (0)
ITI	.44 (170)	.40 (155)	.16 (62)	.00 (0)

Table 5.5 (continued)

	INCORRECT	ERROR CONSEQUENCE	CORRECT CONSEQUENCE	ITI	SUM
SDMASTER	.12 (23)	.00 (0)	.01 (1)	.00 (0)	1.00 (190)
SDACQUIRE	.23 (53)	.00 (1)	.00 (0)	.00 (0)	1.00 (229)
GESTURAL	.02 (2)	.01 (1)	.01 (1)	.00 (0)	1.00 (114)
CORRECT	.01 (4)	.00 1)	.96 (349)	.00 (0)	1.00 (364)
INCORRECT	.02 (2)	.74 (62)	.04 (3)	00 (0)	1.00 (84)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.70 (46)	1.00 (66)
CORRECT CONSEQUENCE	.00 (0)	.00 (1)	.00 (0)	.96 (342)	1.00 (355)
ITI	.00 (0)	.00 (0)	.00 (1)	.00 (0)	1.00 (388)
SUM					1788

Table 5.6

The Dyad 5 baseline phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All the cells that are bold-faced are the treatment integrity errors that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	CORRECT	INCORRECT
SDMASTER	.00 (0)	.00 (0)	.00 (0)	.71 (24)	.29 (10)
SDACQUIRE	.00 (0)	.00 (0)	.04 (3)	.67 (52)	.26 (20)
GESTURAL	.00 (0)	.90 (37)	.00 (0)	.05 (2)	.02 (1)
CORRECT	.00 (0)	.01 (1)	.01 (1)	.00 (0)	.02 (2)
INCORRECT	**.06 (2)**	**.06 (2)**	.00 (0)	**.24 (8)**	.00 (0)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.04 (1)	**.29 (7)**	.00 (0)
CORRECT CONSEQUENCE	.01 (1)	.00 (0)	.01 (1)	.01 (1)	.00 (0)
ITI	.30 (31)	.37 (38)	.33 (34)	.00 (0)	.00 (0)

Table 5.6 (continued)

	ERROR CONSEQUENCE	CORRECT CONSEQUENCE	ITI	SUM
SDMASTER	.00 (0)	.00 (0)	.00 (0)	1.00 (34)
SDACQUIRE	.01 (1)	.03 (2)	.00 (0)	1.00 (78)
GESTURAL	.00 (0)	.00 (0)	.02 (1)	1.00 (41)
CORRECT	.01 (1)	.95 (89)	.00 (0)	1.00 (94)
INCORRECT	.64 (21)	.00 (0)	.00 (0)	1.00 (33)
ERROR CONSEQUENCE	.00 (0)	.04 (1)	.63 (15)	1.00 (24)
CORRECT CONSEQUENCE	.01 (1)	.00 (0)	.96 (88)	1.00 (92)
ITI	.00 (0)	.00 (0)	.00 (0)	1.00 (103)
SUM				1.00 (499)

Furthermore, results from Table 5.6 identified the ERROR CONSEQUENCE – CORRECT sequence (*n* = 7, p = .29) indicating that there was a .29 probability that the learner would emit a correct response following the application of the error consequence. Error consequences should be followed by a 1s ITI. Learner self-corrections (*n* = 8, p = .24) also reached the .05 criterion and was the only error sequence that reached the criterion for all baseline dyads.

No figure of the treatment integrity errors for Dyad 5 was constructed due to the limited number of data for this dyad. The errors regarding events following incorrect learner responses were reported to the programme consultant.

Baseline results for Dyad 6 can be seen in Table 5.7. Eight two-event error sequences were identified. Four of the error sequences involved events following incorrect learner responses. The prescribed event following incorrect responses was the error consequence. The two most likely treatment integrity error events following incorrect learner responses were verbal prompts (INCORRECT – VERBAL, *n* = 17, p = .17), and a correct learner response (INCORRECT – CORRECT, *n* = 15, p = 0.15). The learner was also likely to make another incorrect response (INCORRECT – INCORRECT, *n* = 8, p = .08). The correct DTT protocol sequence of INCORRECT – ERROR CONSEQUENCE sequence had a probability of .51 indicating that approximately half of all incorrect responses were followed by the prescribed error consequence. The two most likely treatment integrity error events to follow error consequences were verbal prompts (ERROR CONSEQUENCE, *n* = 7, p = .12), and correct learner responses (*n* = 5, p = .09). The prescribed sequence of ERROR CONSEQUENCE – ITI was observed 44 times with a probability of .76.

The sequences CORRECT – SDMASTER (*n* = .06, p = 27), and INCORRECT – SDMASTER (*n* = 5, p = .05) shows instances where learner responses were not followed by the prescribed consequence. The CORRECT – SDMASTER error sequences was the treatment integrity error that occurred with the highest frequency and probability of all the baseline phase errors. The final error sequence was CORRECT – ERROR CONSEQUENCE (*n* = 5, p = .15) indicating instances where correct learner responses were followed by error consequences. The errors that were reported to the consultant and therapist were the errors involving error consequences and incorrect learner responses.

Table 5.7

The Dyad 6 baseline phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	VERBAL	CORRECT
SDMASTER	.02 (8)	.00 (0)	.01 (4)	.06 (20)	.73 (234)
SDACQUIRE	.00 (0)	.01 (2)	.04 (8)	.13 (25)	.70 (136)
GESTURAL	.00 (0)	.41 (9)	.05 (1)	.00 (0)	.50 (11)
VERBAL	.01 (1)	.09 (8)	.00 (0)	.03 (3)	.72 (63)
CORRECT	**.06 (27)**	.02 (9)	.00 (0)	.02 (9)	.00 (1)
INCORRECT	**.05 (5)**	.01 (1)	.01 (1)	**.17 (17)**	**.15 (15)**
CORRECT CONSEQUENCE	.01 (6)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
ERROR CONSEQUENCE	.03 (2)	.00 (0)	.00 (0)	**.12 (7)**	**.09 (5)**
ITI	.60 (271)	.37 (165)	.02 (8)	.01 (6)	.00 (0)

Table 5.7 (continued)

	INCORRECT	CORRECT CONSEQUENCE	ERROR CONSEQUENCE	ITI	SUM
SDMASTER	.17 (55)	.00 (0)	.00 (0)	.00 (0)	1.00 (321)
SDACQUIRE	.12 (23)	.00 (0)	.00 (0)	.00 (0)	1.00 (194)
GESTURAL	.05 (1)	.00 (0)	.00 (0)	.00 (0)	1.00 (22)
VERBAL	.13 (11)	.00 (0)	.01 (1)	.00 (0)	1.00 (87)
CORRECT	.01 (4)	.88 (408)	**.15 (5)**	.00 (2)	1.00 (465)
INCORRECT	**.08 (8)**	.03 (3)	.51 (52)	.00 (0)	1.00 (102)
CORRECT CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.99 (405)	1.00 (411)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.76 (44)	1.00 (58)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (450)
SUM					2110

Table 5.8 contains the baseline data for Dyad 7 and shows that five two-event sequences reached the .05 criterion. Two of the errors involved the ITI not lasting for the minimum prescribed duration of 1 s. These sequences were CORRECT CONSEQUENCE – SDMASTER (*n* = 24, p = .06), and ERROR CONSEQUENCE – SDMASTER (*n* = 3, p =.05). Again it is interesting to note that incorrect application of the ITI was more likely when the next S^D was a mastered S^D.

The other three error sequences involved events following incorrect learner responses. The prescribed event following incorrect learner responses was the error consequence which was the same as for Dyad 6. As seen in Table 5.8 the probability of the error consequence following incorrect learner responses was .64 (*n* = 49). The most likely error sequence involving incorrect learner responses was INCORRECT – SDACQUIRE (*n* = 11, p = .14). The other two sequences were INCORRECT – VERBAL (*n* = 4, p = .05) indicating that verbal prompts were likely to follow incorrect learner responses, and INCORRECT – INCORRECT (*n* = 4, p = .05) indicating successive incorrect learner responses. All five error sequences were reported to the therapist (who was also the consultant) during the meeting.

The baseline data for Dyad 8 can be seen in Table 5.9. Nine two-event error sequences reached the .05 criterion. Five of the error sequences involved response prompt errors. The prescribed application of response prompts for this dyad was that the level of the prompt should increase across trials (least-to-most). The sequences GESTURAL – VERBAL (*n* = 17, p = .41), VERBAL – GESTURAL (*n* = 9, p = .09), VERBAL – VERBAL (*n* = 16, p = .17), VERBAL – PHYSICAL (*n* = 5, p = .05), PHYSICAL – VERBAL (*n* = 1, p = .14) shows that prompts were applied in an unsystematic manner within trials much like the results from Dyad 2.

Table 5.8

The Dyad 7 baseline phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	PHYSICAL	VERBAL
SDMASTER	.01 (2)	.00 (0)	.00 (0)	.00 (0)	.03 (7)
SDACQUIRE	.00 (0)	.00 (0)	.03 (7)	.02 (5)	.23 (48)
GESTURAL	.00 (0)	.13 (1)	.00 (0)	.00 (0)	.00 (0)
PHYSICAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
VERBAL	.02 (1)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
CORRECT	.02 (7)	.00 (1)	.00 (0)	.00 (1)	.00 (0)
INCORRECT	.03 (2)	**.14 (11)**	.00 (0)	.00 (0)	**.05 (4)**
CORRECT CONSEQUENCE	**.06 (24)**	.02 (7)	.00 (0)	.00 (0)	.00 (0)
ERROR CONSEQUENCE	**.05 (3)**	.04 (2)	.00 (0)	.00 (0)	.04 (2)
ITI	.51 (200)	.48 (188)	.00 (1)	.00 (0)	.00 (0)

Table 5.8 (continued)

	CORRECT	INCORRECT	CORRECT CONSEQUENCE	ERROR CONSEQUENCE	ITI	SUM
SDMASTER	.80 (191)	.15 (37)	.00 (0)	.01 (3)	.00 (0)	1.00 (240)
SDACQUIRE	.57 (119)	.14 (30)	.00 (0)	.00 (1)	.00 (0)	1.00 (210)
GESTURAL	.88 (7)	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (8)
PHYSICAL	.67 (4)	.00 (0)	.17 (1)	.17 (1)	.00 (0)	1.00 (6)
VERBAL	.93 (57)	.05 (3)	.00 (0)	.00 (0)	.00 (0)	1.00 (61)
CORRECT	.00 (1)	.00 (1)	.97 (372)	.00 (1)	.00 (0)	1.00 (384)
INCORRECT	.04 (3)	**.05 (4)**	.04 (3)	.64 (49)	.00 (0)	1.00 (76)
CORRECT CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.92 (345)	1.00 (376)
ERROR CONSEQUENCE	.04 (2)	.00 (0)	.00 (0)	.00 (0)	.84 (46)	1.00 (55)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (390)
SUM						1806

Table 5.9

The Dyad 8 baseline phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	VERBAL
SDMASTER	.02 (4)	.00 (0)	.00 (0)	.04 (9)
SDACQUIRE	.00 (0)	.01 (3)	.08 (21)	.07 (19)
GESTURAL	.00 (0)	.02 (1)	.02 (1)	**.41 (17)**
VERBAL	.00 (0)	.12 (11)	**.09 (9)**	**.17 (16)**
PHYSICAL	.00 (0)	.00 (0)	.00 (0)	**.14 (1)**
MODEL	.00 (0)	.93 (13)	.00 (0)	.00 (0)
CORRECT	.01 (3)	.03 (13)	.01 (3)	.00 (1)
INCORRECT	.03 (4)	.04 (6)	.03 (4)	**.12 (17)**
CORRECT CONSEQUENCE	.01 (4)	.01 (2)	.00 (0)	.00 (0)
ERROR CONSEQUENCE	.00 (2)	.02 (2)	.03 (3)	.04 (4)
ITI	.45 (196)	.51 (222)	.00 (0)	.02 (10)

Table 5.9 (continued)

	PHYSICAL	MODEL	CORRECT	INCORRECT
SDMASTER	.00 (0)	.00 (0)	.71 (150)	.20 (42)
SDACQUIRE	.00 (1)	.00 (1)	.53 (146)	.29 (80)
GESTURAL	.02 (1)	.00 (0)	.49 (20)	.00 (0)
VERBAL	**.05 (5)**	.00 (0)	.45 (43)	.04 (4)
PHYSICAL	.00 (0)	.00 (0)	.71 (5)	.00 (0)
MODEL	.00 (0)	.00 (0)	.00 (0)	.00 (0)
CORRECT	.00 (0)	.00 (1)	.01 (2)	.02 (7)
INCORRECT	.00 (0)	.00 (0)	**.07 (10)**	**.06 (8)**
CORRECT CONSEQUENCE	.00 (0)	.01 (2)	.00 (0)	.00 (0)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	**.05 (6)**	.00 (0)
ITI	.00 (0)	.02 (10)	.00 (0)	.00 (0)

Table 5.9 (continued)

	CORRECT CONSEQUENCE	ERROR CONSEQUENCE	ITI	SUM
SDMASTER	.00 (0)	.03 (6)	.00 (0)	1.00 (211)
SDACQUIRE	.00 (0)	.01 (3)	.00 (0)	1.00 (274)
GESTURAL	.02 (1)	.00 (0)	.00 (0)	1.00 (41)
VERBAL	.01 (1)	.06 (6)	.00 (0)	1.00 (95)
PHYSICAL	.00 (0)	.14 (1)	.00 (0)	1.00 (7)
MODEL	.07 (1)	.00 (0)	.00 (0)	1.00 (14)
CORRECT	.92 (350)	.01 (2)	.00 (0)	1.00 (382)
INCORRECT	.01 (1)	.65 (91)	.00 (0)	1.00 (141)
CORRECT CONSEQUENCE	.00 (0)	.00 (1)	.97 (344)	1.00 (354)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.86 (95)	1.00 (110)
ITI	.00 (0)	.00 (0)	.00 (0)	1.00 (438)
SUM				2067

Three of the error sequences consisted of events following incorrect learner responses. Theses sequences were similar to the sequences identified for Dyads 6 and 7. The prescribed error consequence was the same as for Dyads 6 and 7 and was observed 91 times with a probability of .65. The sequences INCORRECT – CORRECT (*n* = 10, p = .07), and INCORRECT – INCORRECT (*n* = 8, p = .06) indicate multiple learner responses. The sequence INCORRECT – VERBAL (*n* = 17, p = .12) shows that the therapist administered a verbal prompt instead of the prescribed error consequence.

The final error sequence was ERROR CONSEQUENCE – CORRECT (*n* = 6, p = .05) showing that that were occasions when the therapist allowed the learner to emit a correct response following the error consequence instead of allowing an ITI of at least 1 s. The multiple response prompt errors and errors involving incorrect learner responses were reported to the therapist and consultant.

Follow-up Phase

The follow-up results for Dyad 4 can be seen in Table 5.10, and shows that one two-event error sequence reached the .05 criterion. That error was learner self-corrections (*n* = 11, p = .16). Learner self-corrections increased in its frequency and probability between phases. The frequency and probability of the other two error sequences that were identified during the baseline phase did not occur at all during the follow-up phase.

Figure 5.4 contains a comparison of the error sequences across the baseline and follow-up phases. The probabilities of the INCORRECT – GESTURAL and ERROR CONSEQUENCE – GESTURAL sequences were reduced to .00 for all blocks during the follow-up phase. The probability of the INCORRECT – CORRECT sequence increased between phases. The probability of the INCORRECT – CORRECT sequence was .00 across the final 3 blocks in baseline before increasing in the follow-up phase.

Table 5.10

The Dyad 4 follow-up phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	PHYSICAL	CORRECT
SDMASTER	.00 (0)	.00 (0)	.09 (30)	.01 (3)	.77 (259)
SDACQUIRE	.00 (0)	.01 (1)	.29 (33)	.09 (10)	.50 (57)
GESTURAL	.03 (2)	.05 (3)	.03 (2)	.02 (1)	.80 (53)
PHYSICAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (14)
CORRECT	.01 (2)	.00 (0)	.00 (0)	.00 (0)	.00 (1)
INCORRECT	.00 (0)	.00 (0)	.00 (0)	.00 (0)	**.16 (11)**
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
CORRECT CONSEQUENCE	.01 (5)	.00 (1)	.00 (0)	.00 (0)	.00 (1)
ITI	.75 (328)	.25 (110)	.00 (0)	.00 (0)	.00 (0)

Table 5.10 (continued)

	INCORRECT	ERROR CONSEQUENCE	CORRECT CONSEQUENCE	ITI	SUM
SDMASTER	.13 (45)	.00 (1)	.00 (0)	.00 (0)	1.00 (338)
SDACQUIRE	.12 (14)	.00 (0)	.00 (0)	.00 (0)	1.00 (115)
GESTURAL	.08 (5)	.00 (0)	.00 (0)	.00 (0)	1.00 (66)
PHYSICAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (14)
CORRECT	.01 (4)	.01 (2)	.98 (387)	.00 (0)	1.00 (396)
INCORRECT	.01 (1)	.83 (57)	.00 (0)	.00 (0)	1.00 (69)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	1.00 (60)	1.00 (60)
CORRECT CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.98 (380)	1.00 (387)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (439)
SUM					1884

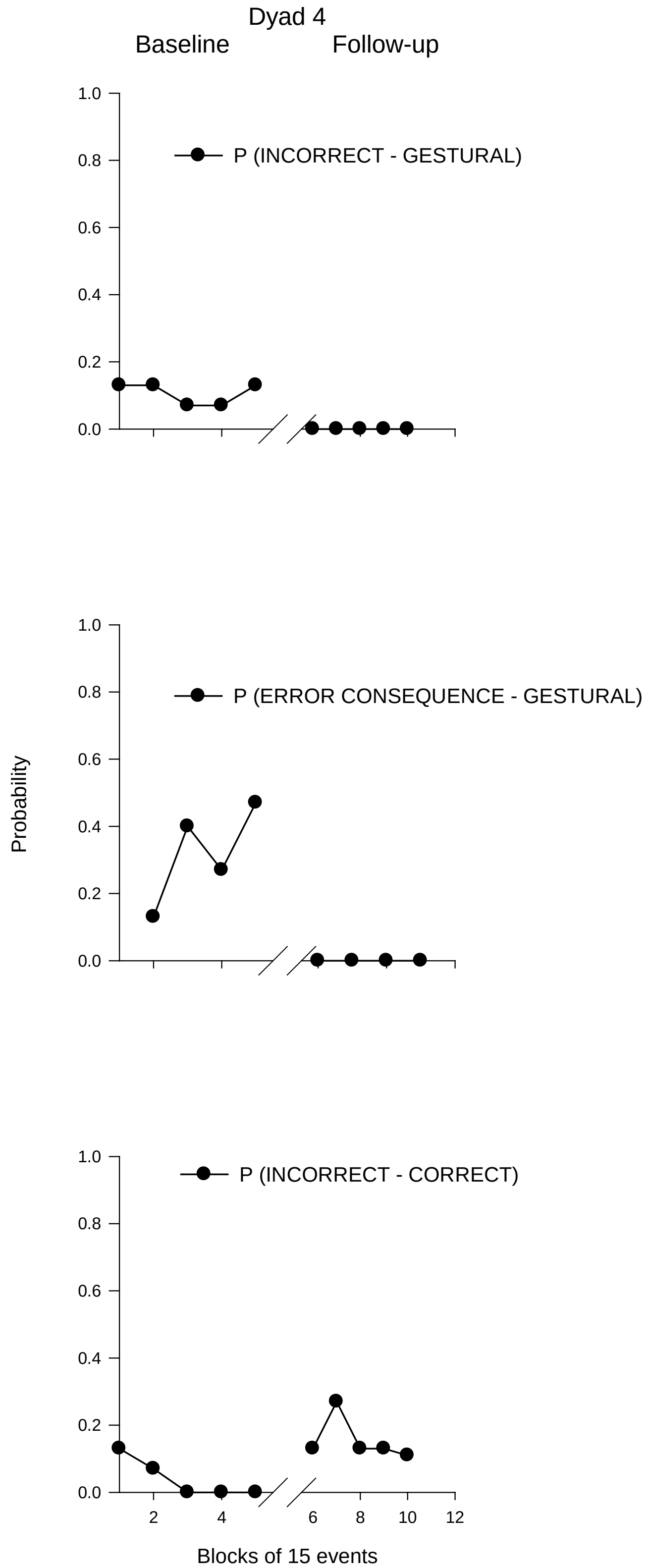

Figure 5.4: Probability of the treatment integrity errors for Dyad 4 during blocks of 15 events across both the baseline and follow-up phases.

The results show that in the follow-up phase two of three errors that were identified in the baseline phase were completely reduced. Furthermore, no new errors were identified in the follow-up phase.

The follow-up results for Dyad 6 can be seen in Table 5.11. Seven error sequences were identified; five of the sequences were the same as the ones identified in the baseline phase, and two new sequences were identified. The two new sequences were ERROR CONSEQUENCE – SDACQUIRE (*n* = 3, p = .05) indicating that error consequences were not followed by ITI of its minimum prescribed duration of 1 s, and INCORRECT – CORRECT CONSEQUENCE (*n* = .4, p = .07) indicating that the therapist provided correct response consequences when the learner had made incorrect responses. The two baseline phase error sequences of CORRECT – SDMASTER (*n* = 9, p = .02) and ERROR CONSEQUENCE – CORRECT (*n* = 2, p = .03) had lower frequencies and probabilities during the follow-up phase.

Figure 5.5 contains the probabilities of the four error sequences that were observed most frequently during the baseline phase. The baseline phase probabilities for all of the error sequences are varied and do not show any trends. In the follow-up phase only the ERROR CONSEQUENCE – CORRECT sequence was reduced to .00 for first two blocks of 15 events. The probabilities of the INCORRECT – VERBAL and ERROR CONSEQUENCE – VERBAL sequences remained varied during the follow-up phase. The INCORRECT – CORRECT sequence remained stable over the first 3 blocks of 15 events before being reduced to .00.

Follow-up results for Dyad 7 can be seen in Table 5.12. Four error sequences were identified and they were all related to events following incorrect learner responses. All of the sequences were low frequency low probability sequences. Despite the increase in the number

Table 5.11

The Dyad 6 follow-up results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	VERBAL	CORRECT
SDMASTER	.02 (5)	.00 (0)	.04 (9)	.00 (1)	.72 (156)
SDACQUIRE	.00 (0)	.02 (5)	.10 (23)	.15 (33)	.60 (131)
GESTURAL	.03 (1)	.00 (0)	.00 (0)	.00 (0)	.88 (30)
VERBAL	.06 (4)	.04 (3)	.00 (0)	.01 (1)	.85 (57)
CORRECT	.02 (9)	.04 (17)	.00 (0)	.03 (13)	.00 (1)
INCORRECT	**.08 (5)**	.03 (2)	.02 (1)	**.10 (6)**	**.05 (3)**
CORRECT CONSEQUENCE	.01 (5)	.01 (4)	.00 (1)	.00 (0)	.00 (0)
ERROR CONSEQUENCE	.00 (0)	**.05 (3)**	.00 (0)	**.22 (13)**	.03 (2)
ITI	.51 (188)	.49 (185)	.00 (0)	.00 (0)	.00 (0)

Table 5.11 (continued)

	INCORRECT	CORRECT CONSEQUENCE	ERROR CONSEQUENCE	ITI	SUM
SDMASTER	.14 (30)	.00 (0)	.07 (16)	.00 (0)	1.00 (217)
SDACQUIRE	.10 (23)	.01 (2)	.01 (3)	.00 (0)	1.00 (220)
GESTURAL	.09 (3)	.00 (0)	.00 (0)	.00 (0)	1.00 (34)
VERBAL	.03 (2)	.00 (0)	.00 (0)	.00 (0)	1.00 (67)
CORRECT	.00 (0)	.89 (338)	.01 (2)	.00 (0)	1.00 (380)
INCORRECT	**.05 (3)**	**.07 (4)**	.61 (37)	.00 (0)	1.00 (61)
CORRECT CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.97 (334)	1.00 (344)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.69 (40)	1.00 (58)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (373)
SUM					1755

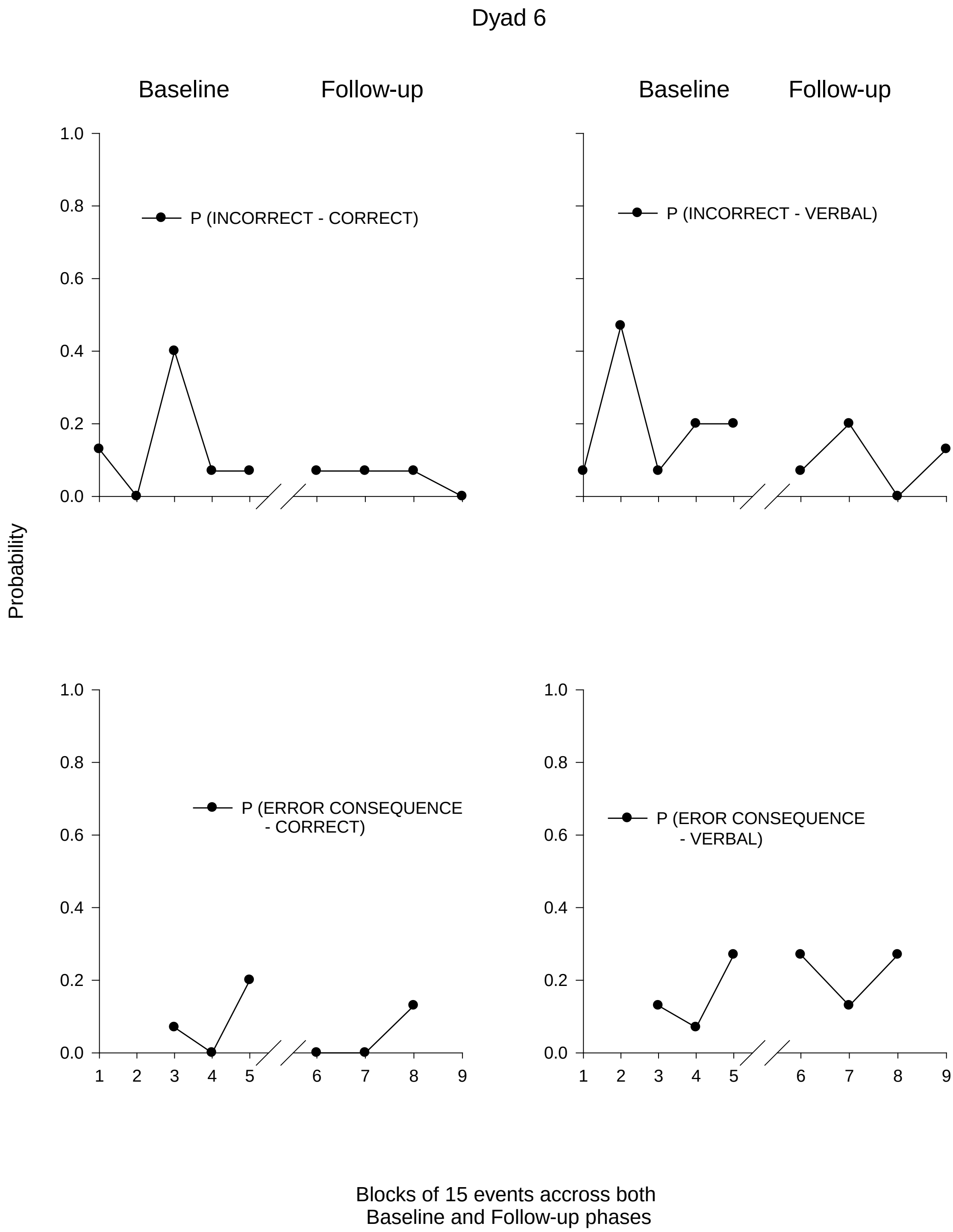

Figure 5.5: Probability of selected treatment integrity errors for Dyad 6 during blocks of 15 events across both the baseline and follow-up phases

of error sequences following incorrect learner responses between phases, the probability of the prescribed error consequence following incorrect learner response increased to .69 (n = 38). Two error sequences that were not identified in the baseline phase reached the .05 criterion during the follow-up phase. These sequences were INCORRECT – CORRECT

CONSEQUENCES (*n* = 3, p = .05), and INCORRECT – SDMASTER (*n* = 3, p = .05). The INCORRECT –CORRECT CONSEQUENCE shows that some incorrect learner responses were followed by the consequence for emitting correct response which for this learner was social praise.

Figure 5.6 contains of a comparison of the probabilities across phases for the two error sequence that was observed with the highest frequency in the baseline phase (INCORRECT – SDACQUIRE, and CORRECT CONSEQUENCE – SDMASTER). For the CORRECT CONSEQUENCE – SDMASTER sequence only the first 15 blocks of 15 events were plotted for the follow-up data. As seen in Figure 5.6, initially the probabilities for both sequences were immediately reduced to .00. However, the probability of the INCORRECT – SDACQUIRE sequence increased to .07. For the CORRECT CONSEQUENCE – SDMASTER the probability also increased to .20 before returning to .00.

Therefore, the probabilities of three out of the five error sequences were reduced between phases. The probability of one of the sequences remained constant and one increased. In addition two new error sequences were identified.

The follow-up results for Dyad 8 can be seen in Table 5.13, and one error sequence was identified. The sequence was INCORRECT – VERBAL (*n* = 18, p = .19) which increased in frequency and probability between phases. All other errors identified during the baseline phase had a probability of less than .05 during the follow-up phase and no new errors were identified. The probabilities and frequencies of the multiple response prompt errors were also reduced.

The probabilities of the error sequences involving incorrect learner responses and error consequences can be seen in Figure 7. The probabilities of the sequences INCORRECT

Table 5.12

The Dyad 7 follow-up phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	PHYSICAL	VERBAL
SDMASTER	.04 (6)	.00 (0)	.01 (1)	.00 (0)	.03 (5)
SDACQUIRE	.00 (0)	.01 (1)	.04 (5)	.10 (13)	.30 (40)
GESTURAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
PHYSICAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
VERBAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
CORRECT	.00 (0)	.01 (2)	.00 (0)	.00 (0)	.00 (0)
INCORRECT	**.05 (3)**	.02 (1)	.02 (1)	.00 (0)	**.05 (3)**
CORRECT CONSEQUENCE	.02 (4)	.01 (3)	.00 (0)	.00 (0)	.00 (0)
ERROR CONSEQUENCE	.02 (1)	.00 (0)	.00 (0)	.00 (0)	.00 (0)
ITI	.55 (155)	.45 (125)	.00 (0)	.00 (0)	.00 (0)

Table 5.12 (continued)

	CORRECT	INCORRECT	CORRECT CONSEQUENCE	ERROR CONSEQUENCE	ITI	SUM
SDMASTER	.76 (128)	.15 (26)	.00 (0)	.02 (3)	.00 (0)	1.00 (169)
SDACQUIRE	.39 (52)	.15 (20)	.00 (0)	.02 (2)	.00 (0)	1.00 (133)
GESTURAL	1.00 (7)	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (7)
PHYSICAL	.92 (12)	.08 (1)	.00 (0)	.00 (0)	.00 (0)	1.00 (13)
VERBAL	.90 (43)	.08 (4)	.00 (0)	.02 (1)	.00 (0)	1.00 (48)
CORRECT	.00 (1)	.00 (0)	.98 (242)	.00 (1)	.00 (0)	1.00 (246)
INCORRECT	.04 (2)	**.07 (4)**	**.05 (3)**	.69 (38)	.00 (0)	1.00 (55)
CORRECT CONSEQUENCE	.01 (2)	.00 (0)	.00	.00 (0)	.96 (237)	1.00 (246)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.02 (1)	.00 (0)	.96 (43)	1.00 (45)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (280)
SUM						1242

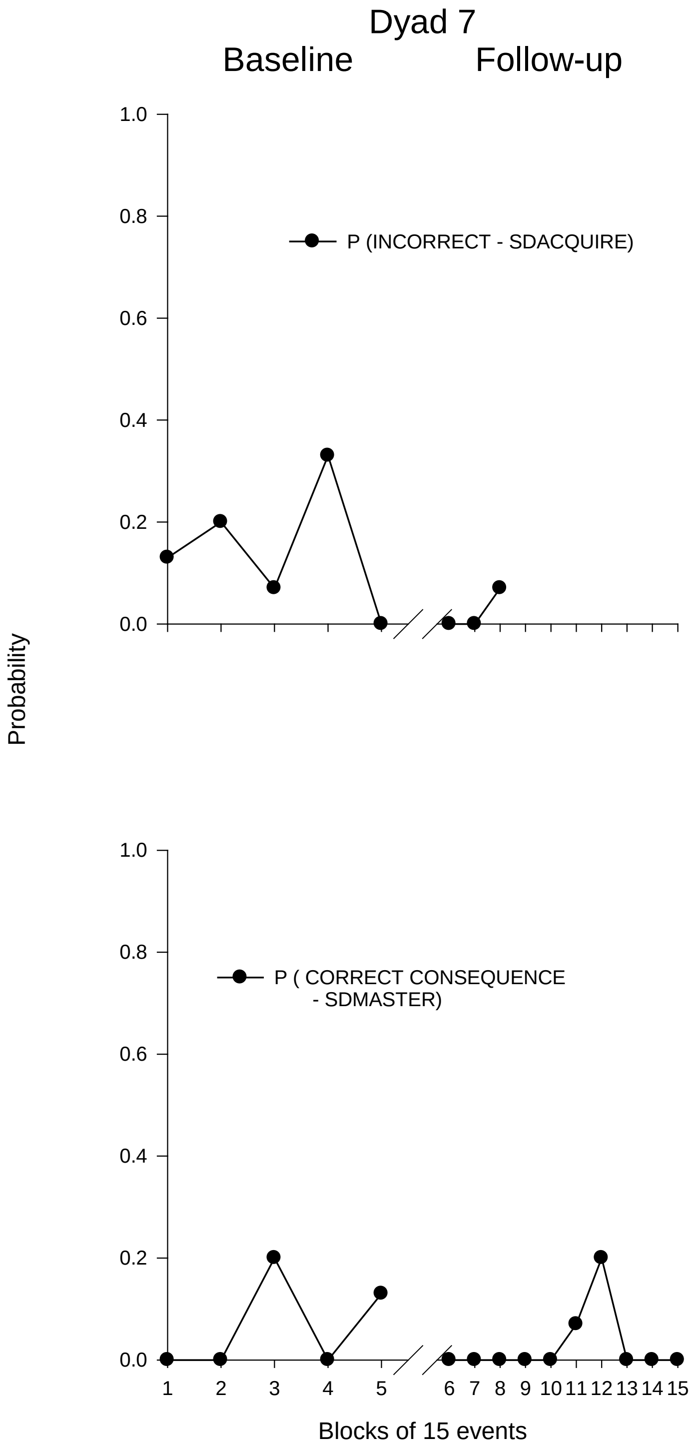

Figure 5.6: Probability of selected treatment integrity errors for Dyad 7 during blocks of 15 events across both the baseline and follow-up phases

– VERBAL, INCORRECT – CORRECT, and ERROR CONSEQUENCE - CORRECT showed variability during the baseline phase, and the INCORRECT – CORRECT sequence appear stable.

As seen in Figure 5.7 there was no change in trend for the INCORRECT – VERBAL sequence between phases. An interesting feature from Figure 5.7 is what happens to the probabilities of the INCORRECT – CORRECT, ERROR CONSEQUENCE – CORRECT and INCORRECT – INCORRECT sequences. The probabilities were reduced to .00 for the first few blocks of 15 events immediately following the meeting with the therapist and consultant. However, over time the probabilities returned back to their baseline levels.

Treatment Integrity Summary

A summary of the treatment integrity scores for all dyads can be seen in Table 5.14. Treatment integrity scores were calculated by dividing the number of trials that were administered following the correct DTT protocol by the total number of trials administered and multiplying the ratio by 100.

The baseline phase data will be described first. The number of baseline phase trials ranged between 110 (Dyad 5), and 1531 (Dyad 2). Table 5.14 shows that Dyads 1 and 2 had overall treatment integrity scores of less than 45% (42% and 27% respectively). The overall treatment integrity scores were at least 80% for Dyads 3, 4, 6, 7 and 8.

Treatment integrity scores for mastered trials ranged between 32% (Dyad 2) and 89% (Dyad 4). Acquisition trial treatment integrity scores had a larger range with the scores ranging between 19% and 87%. Five dyads had higher treatment integrity scores for mastered trials than acquisition trials.

Table 5.13

The Dyad 8 follow-up phase results. The table contains the probabilities of transitions between all events, and the frequency of the transitions (in brackets). All cells that are bold-faced are the treatment integrity error sequences that reached the .05 criterion.

	SDMASTER	SDACQUIRE	GESTURAL	VERBAL	CORRECT
SDMASTER	.02 (3)	.00 (0)	.01 (1)	.05 (9)	.74 (125)
SDACQUIRE	.00 (0)	.01 (4)	.07 (19)	.23 (64)	.49 (135)
GESTURAL	.00 (0)	.05 (1)	.00 (0)	.00 (0)	.95 (20)
VERBAL	.00 (0)	.01 (1)	.00 (0)	.02 (2)	.84 (90)
CORRECT	.01 (2)	.02 (9)	.00 (0)	.03 (11)	.00 (1)
INCORRECT	.02 (2)	.03 (3)	.00 (0)	**.19 (18)**	.03 (3)
CORRECT CONSEQUENCE	.01 (2)	.03 (9)	.00 (0)	.01 (3)	.01 (2)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.00 (0)	.01 (1)
ITI	.39 (158)	.61 (248)	.00 (1)	.00 (0)	.00 (0)

Table 5.13 (continued)

	INCORRECT	CORRECT CONSEQUENCE	ERROR CONSEQUENCE	ITI	SUM
SDMASTER	.15 (26)	.00 (0)	.02 (40)	.00 (0)	1.00 (168)
SDACQUIRE	.18 (49)	.00 (1)	.01 (3)	.00 (0)	1.00 (275)
GESTURAL	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (21)
VERBAL	.13 (14)	.00 (0)	.00 (0)	.00 (0)	1.00 (107)
CORRECT	.00 (1)	.94 (353)	.00 (0)	.00 (0)	1.00 (377)
INCORRECT	.03 (3)	.01 (1)	.68 (63)	.00 (0)	1.00 (93)
CORRECT CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.95 (339)	1.00 (355)
ERROR CONSEQUENCE	.00 (0)	.00 (0)	.00 (0)	.99 (69)	1.00 (70)
ITI	.00 (0)	.00 (0)	.00 (0)	.00 (0)	1.00 (407)
SUM					1873

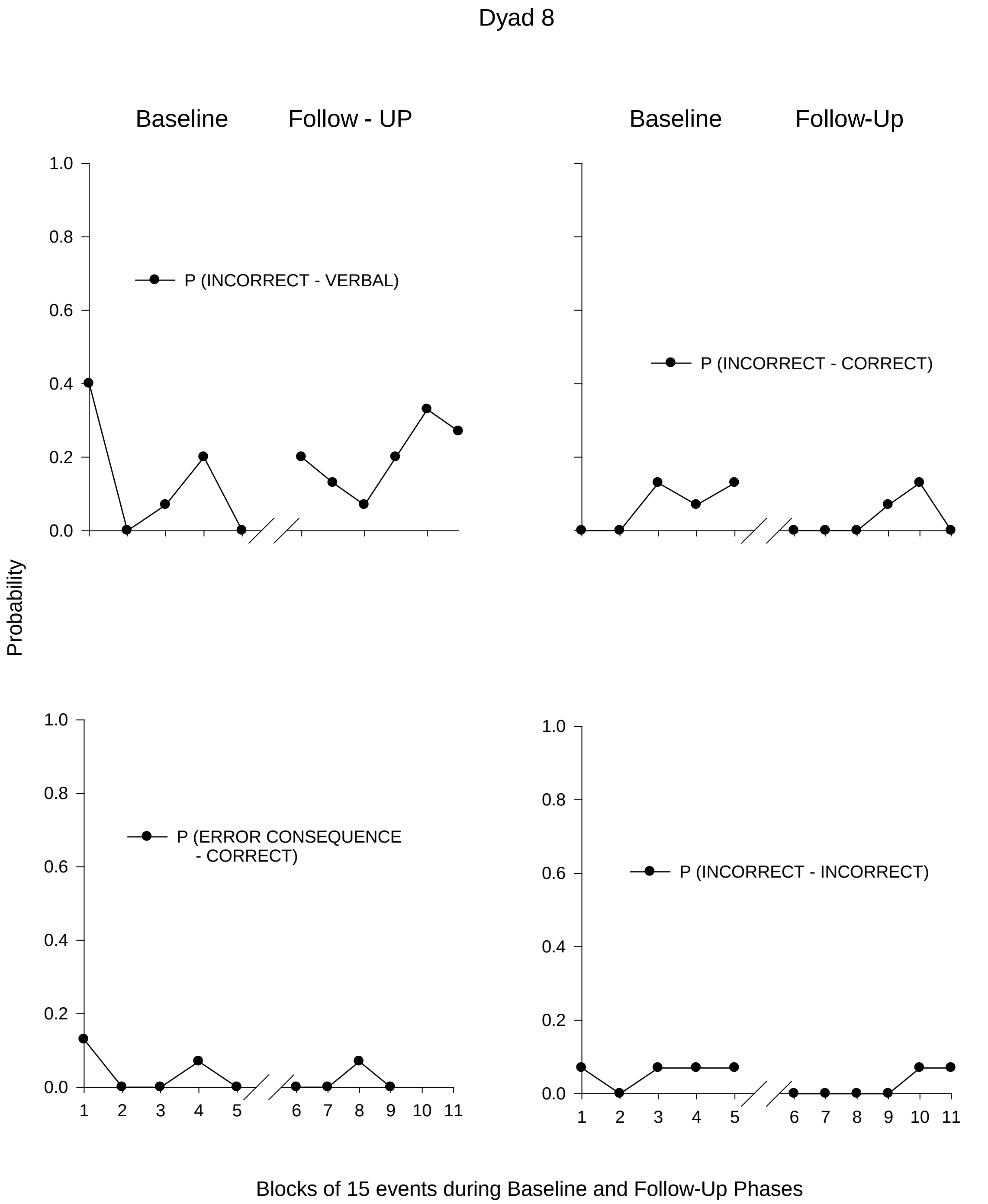

Figure 5.7: Probability of selected treatment integrity errors for Dyad 8 during blocks of 15 events across both the baseline and follow-up phases

Between 295 and 447 trials were administered in the follow-up phase. The overall treatment integrity scores ranged between 82% and 94%. As seen in Table 5.14 overall treatment integrity increased for 3 dyads (Dyads 4, 7 and 8) and remained constant for Dyad 6. Treatment integrity scores for mastered trials increased for all dyads and ranged between 84% and 94%. Acquisition trial treatment integrity scores ranged between 80% and 93%. Scores increased for 3 dyads. Dyad 8 had the largest increase in scores between phases, increasing from 76% to 85%. Treatment integrity decreased for Dyad 6 from 85% to 80%. Three dyads had higher treatment integrity scores for mastered trials.

Figures 5.8 and 5.9 were constructed to investigate the variability and range of treatment integrity scores for blocks of 30 trials. The figures also contain a solid vertical line which represents the 90% criterion for treatment integrity. The treatment integrity score for each block of trials was calculated by the dividing the number of trials that were administered following the correct DTT protocol by the number of trials in each block (30). These analyses were conducted for the overall (Figure 8) and acquisition trials (Figure 5.9) treatment integrity scores. Dyad 5 data were not included in these analyses due to insufficient number of trials. In Figure 5.9 five trial blocks were plotted in the baseline phase for Dyads 4, 6, 7 and 8, seven for Dyad 1, and 10 for Dyads 2 and 3.

Figure 5.8 shows that the overall treatment integrity scores varied in the baseline phase, especially for Dyads 1, 2, 3, 6 and 8. Treatment integrity scores ranged between 30% and 67% for Dyad 1, between 0% and 50% for Dyad 2, between 57% and 100% for Dyad 3, between 70% and 93% for Dyad 6, and between 53% and 93% for Dyad 8. The baseline data appeared stable for Dyads 4 and 7.

As seen in Figure 5.8 the follow-up phase results for Dyads 4, 7 and 6 displayed the same patterns in the follow-up phase than in the baseline phase. The data for Dyad 8 were

Table 5.14

Treatment integrity summary for all dyads.

		Baseline			Follow-up		
Dyad	Trial Type	Number of Trials	Correct DTT protocol	Treatment Integrity	Number of Trials	Correct DTT protocol	Treatment Integrity
1	Overall	473	200	42%	-	-	-
	Mastered	258	127	49%	-	-	-
	Acquisition	215	73	34%	-	-	-
2	Overall	1531	428	27%	-	-	-
	Mastered	1065	338	32%	-	-	-
	Acquisition	466	90	19%	-	-	-
3	Overall	1160	988	85%	-	-	-
	Mastered	548	449	82%	-	-	-
	Acquisition	612	537	88%	-	-	-
4	Overall	413	361	87%	447	420	94%
	Mastered	187	167	89%	336	317	94%
	Acquisition	226	194	86%	111	103	93%
5	Overall	110	85	77%	-	-	-
	Mastered	76	59	78%	-	-	-
	Acquisition	34	26	76%	-	-	-
6	Overall	503	411	82%	425	347	82%
	Mastered	313	250	80%	209	175	84%
	Acquisition	190	161	85%	216	172	80%
7	Overall	450	375	83%	295	261	88%
	Mastered	240	193	80%	163	142	87%
	Acquisition	210	182	87%	132	119	90%
8	Overall	476	383	80%	435	372	86%
	Mastered	207	179	86%	165	143	87%
	Acquisition	269	204	76%	270	229	85%

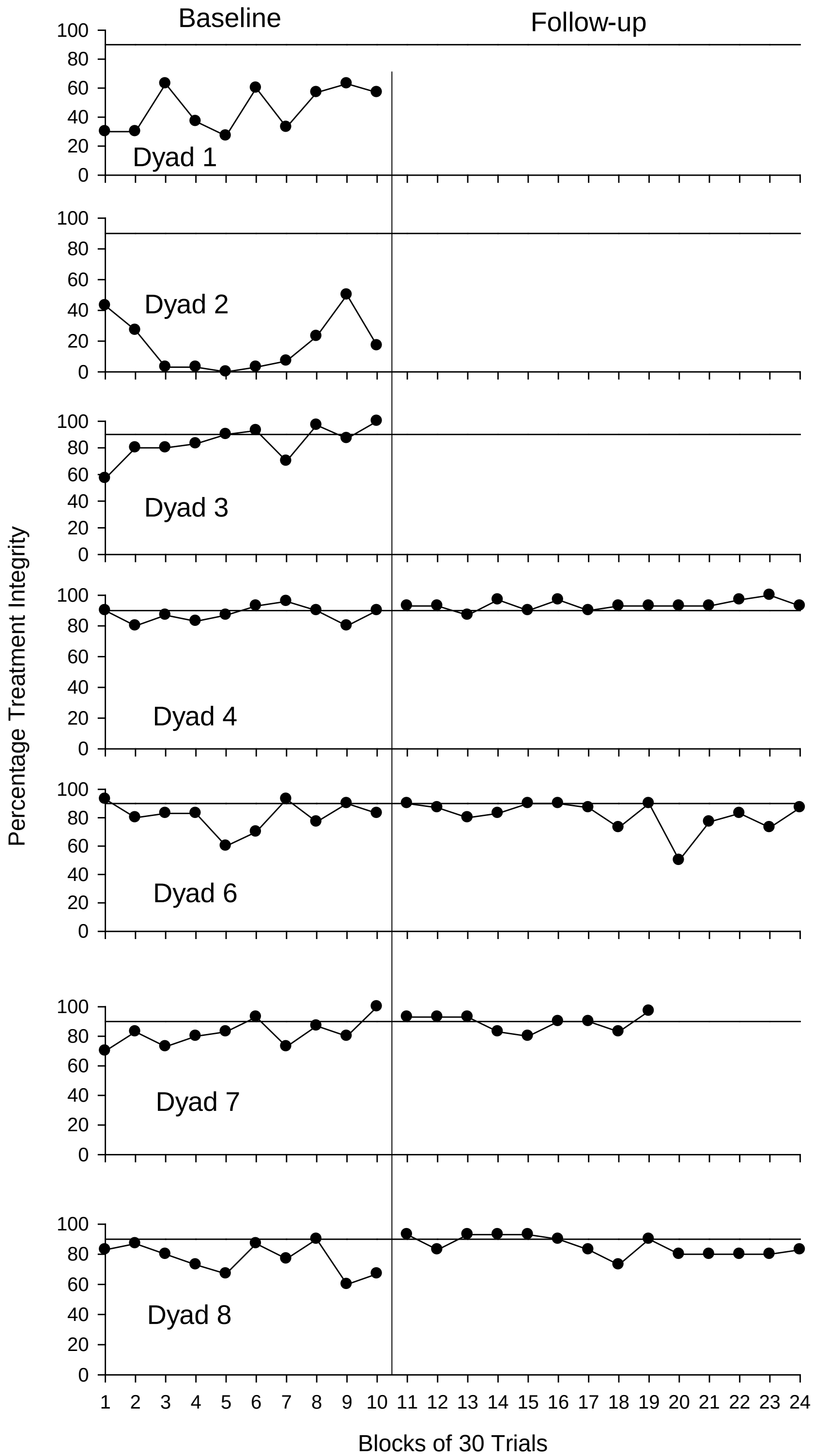

Figure 5.8: Overall treatment integrity scores for blocks of 30 trials for all dyads across both the baseline and follow-up phases.

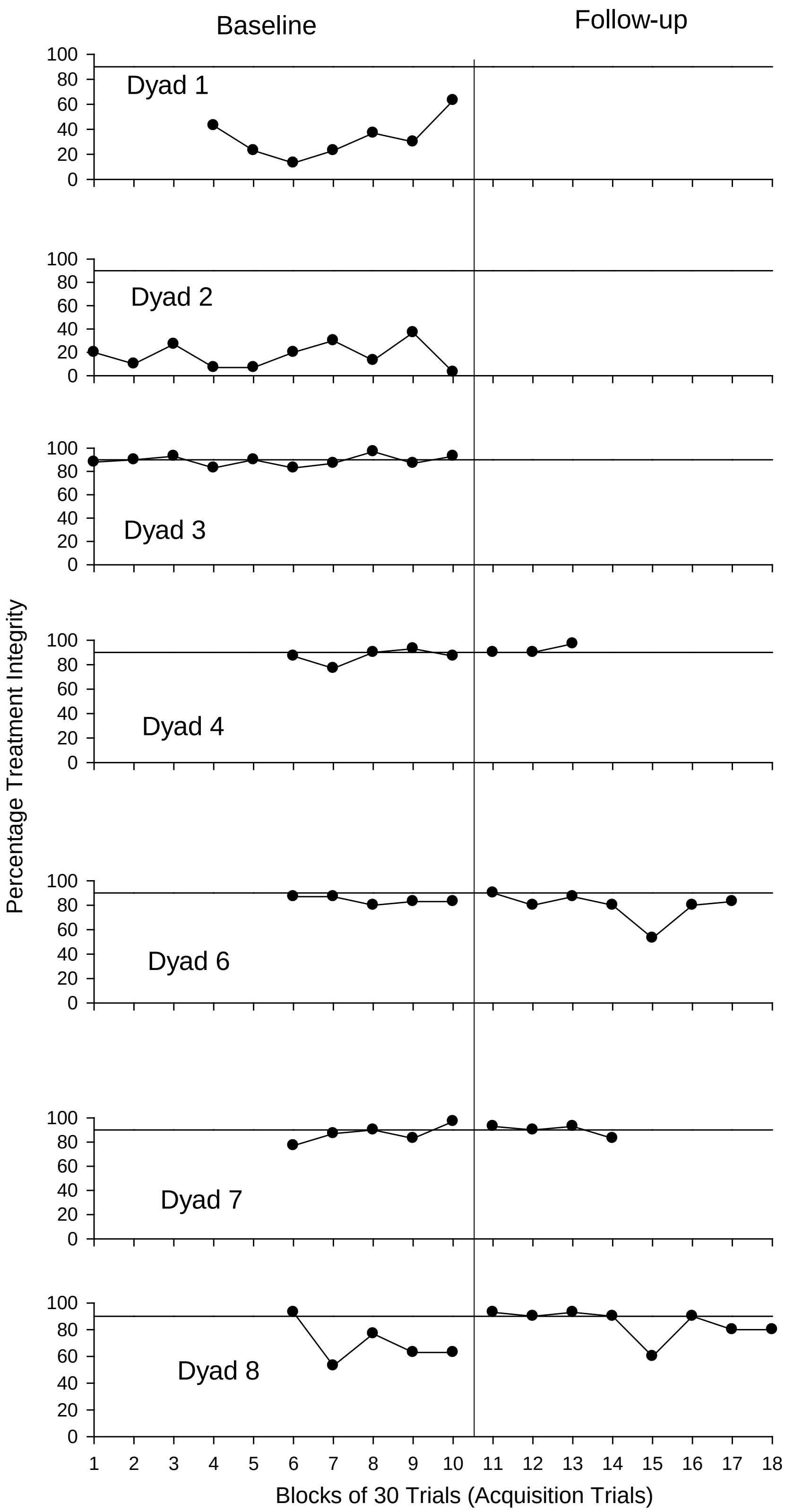

Figure 5.9: Acquisition trial treatment integrity scores for blocks of 30 trials for all dyads across both the baseline and follow-up phases.

less variable in the follow-up phase (range = 73% to 93%) and treatment integrity increased from 63% to 93% in the first block of trials in the follow-up phase. No changes were detected for Dyad 6. Although increases in treatment integrity were observed for 3 dyads in the follow-up phase, these increases are not apparent in Figure 8 because of the high baseline phase treatment integrity scores.

Figure 5.9 shows that the baseline phase acquisition trials treatment integrity scores were stable for Dyads 2, 3, 4, 6 and 7. The scores ranged between 13% and 63% for Dyad 1 and between 53% and 93% for Dyad 8. The follow-up scores for Dyad 6 were more varied in the follow-up phase whereas the follow-up data for Dyad 8 were more stable. No changes were detected for Dyads 4 and 7. This was the case for Figure 5.8, although integrity scores for acquisition trials increased for 3 dyads between phases, the increases were not apparent due to the high baseline phase scores.

Social Validity

Seven social validity questionnaires were distributed to five therapists and two consultants. All questionnaires were returned and a summary of the results can be seen in Table 5.15. Table 5.15 consists of the number of therapists and consultants who answered each question and the distribution of the scores. Each statement was converted into a number between 1 and 6 (Don’t know = 1, Strongly Disagree = 2, Disagree = 3, Neutral = 4, Agree = 5, Strongly Agree = 6). The column labeled *Distribution* shows the distribution of the answers therapists and consultants scored for each item. The 9 items can be categorized as assessing the social significance of the goals of the study (Items 1 and 2), the social significance of the appropriateness of the procedure (Items 5 and 6), and the social appropriateness of the results (Items 3, 4, 7 and 8) (Wolfe, 1978). Item 9 measured the overall satisfaction with the methods and procedures used to carry out the study and report the results.

Table 5.15 shows that the issue of treatment integrity and correcting DTT errors were of importance to therapists and consultants as they all scored Items 1 and 2 as *agree* or *strongly agree*. All the therapists scored *agree* that they were comfortable with the being video-recorded as part of the data collection process and only 1 therapist scored *agree* that that the learner increased in their levels of off-task and non-compliant behaviour. One consultant stated that the procedure used in the study (video-recording) was less intrusive than anticipated. However, one therapist did provide written feedback that the data collection process was slow and that it would have been helpful if feedback from video-recording could be provided more than once.

All therapists and consultants scored *agree* and *strongly agree* that the method used to display the data and convey the important findings were clear and easy to understand, and a majority of the therapists and consultants scored *agree* and *strongly agree* that the results were useful in helping to correct treatment integrity errors. All of the therapists and consultants scored *agree* or *strongly agree* that they would recommend the methods and procedures used in this study as an effective way of evaluating DTT integrity errors (Item 9).

Table 5.15

Social validity ratings. The table consists of the number of participants that answered a question, and the distribution of the answers.

Items	N	Distribution 1-2-3-4-5-6
1. It is important to administer DTT with a high degree of integrity.	7	0-0-0-0-4-3
2. It is important for DTT treatment integrity errors to be identified and corrected.	7	0-0-0-0-3-4
3. The results from video-recordings were presented to me in a clear and easy-to-understand manner.	7	0-0-0-0-5-2
4. I was surprised by the treatment integrity errors that were identified as part of the DTT that I administer/for which I am the consultant.	7	0-0-3-2-2-0
5. On days when video-recordings took place the learner exhibited more off-task or non-compliant behaviour than on days for which video-recordings were not made.	5	0-0-3-1-1-0
6. I was comfortable being video-recorded.	5	0-0-0-0-5-0
7. I found the results useful in helping me to correct treatment integrity errors.	7	1-0-0-0-5-1
8. The results from the study were beneficial to the learner.	7	1-0-0-1-3-2
9. I would recommend this method as an effective way to evaluate treatment integrity errors in DTT.	7	0-0-0-0-5-2

Note. Rating 1 = Don't know, Rating 2 = Strongly Disagree, Rating 3 = Disagree, Rating 4 = Neutral, Rating 5 = Agree, Rating 6 = Strongly Agree.

Discussion

General Findings

This study introduced Markov transition matrices (MTM) as a method of identifying within-trial treatment integrity errors in DTT. The results show that the matrices, in conjunction with the .05 decision rule, were effective in identifying one-step within-trial treatment integrity errors. The error sequences which were consistent across all dyads involved allowing learner self-corrections, not providing ITI of prescribed duration (especially when next S^D was a mastered S^D), and incorrect application of the error-correction procedures.

Using a probability based method such as MTM is a novel approach to studying within-trial treatment integrity during DTT. Traditional methods of analysing DTT treatment integrity errors calculate the percentage of trials which were administered following correct DTT protocol on a session-by-session or blocks-of-trials basis (Catania, Almeida, Liu-Constant & Reed, 2009; Koegel, Russo & Rincover, 1977; Leblanc, Ricciardi, & Luiselli, 2005, Sarokoff & Sturmey, 2004, Vladescu, Carroll, Paden, & Kodak, 2012) and do not always specify which errors are affecting treatment integrity and to what extent. MTM provided specific information regarding the treatment integrity of within-trial events during DTT taking into account all possible transitions between all components of a discrete trial. Therefore, instead of only presenting overall treatment integrity scores (such as Figures 5.8 & 5.9) across all DTT components, the use of MTM provided detailed and specific information regarding where the errors were taking place taking into account all components of the procedure.

Analysing and presenting the data using MTM provided practically useful results which enabled targeted therapist retraining to occur. Once therapists had received retraining increases in the within-trial treatment integrity of DTT programmes for three dyads were observed. That is, the results from the analyses allowed for detailed and specific feedback to consultants and therapists about where therapist retraining had to occur. Whether this type of specific feedback results in significant decreases in time spent retraining therapists can be an area targeted for future research. Belfiore, Fritts, and Herman (2008) conducted a within-trial treatment integrity analysis and found that the ITI was the most problematic part of the procedure (5% of trials were being administered with ITI as prescribed). As part of the intervention staff received additional re-training on how to apply the ITI as prescribed, and during the intervention phase the ITI was administered as prescribed on 81% of all trials. Belfiore et al. stated "using intratrial data in this manner created a more efficient staff training programme" (p. 101). The results from Study 1 offer further evidence that providing component-specific therapist re-training can be an efficient way of re-training therapists.

Data output from the analyses were kept to a minimum. This finding was desirable because one of the main criticisms of MTM in sequential analysis is that they can produce a large amount of data output which can make data presentation and interpretation difficult (Sackett, 1979). By using data from the one-step analyses only along with the .05 decision rule data output were largely reduced without losing any critical information as it pertains to identifying treatment integrity errors. The social validity data indicates that consultants and therapists all agreed or strongly agreed that the results from the analyses were presented in a clear and easy-to-understand manner. Therefore, some of the concerns regarding MTM were mitigated by the methods and procedures used in this study. In addition the results from the social validity questionnaires indicate that the consultants and therapists who participated in the study were generally satisfied with the methods and procedures used to address an issue which they deemed important.

The overall treatment integrity scores for the baseline phase were at least 80% for six dyads, and the treatment integrity scores for acquisition trials were at least 75% for six dyads.

These baseline scores are higher than what has been traditionally reported in studies evaluating treatment integrity and DTT. For example, Leblanc, Ricciardi and Luiselli (2005) reported overall baseline treatment integrity ranging between 32% and 43%, Sarokoff and Sturmey (2004) between 43% and 49%, and Dib and Sturmey (2007) ranging between 0% and 4%.   However, previous research (e.g., Koegel, Russo, & Rincover, 1977) indicated that treatment integrity had to be at least 90% to promote optimal learning. Therefore, even though the baseline treatment integrity scores were high for the dyads for which follow-up data were recorded they still had not reached the at least 90% criterion which meant that it was worthwhile intervening. Only two dyads from the current study (Dyads 1 and 2) scored overall treatment integrity scores of less than 50%, and those same dyads scored less than 40% treatment integrity for acquisition trials.

The high baseline treatment integrity results could have been due to observation reactivity (Kazdin, 1977). In the context of Study 1, observation reactivity means that the therapists would have shown higher levels of adherence to procedural protocol because they were aware that they were being video-recorded. The therapists were also aware that the data were recorded and analysed with the intention of identifying within-trial treatment integrity errors. Having awareness of being video-recorded and knowing the purpose behind collecting and analysing the data, could have led to higher levels of adherence to procedural protocol compared to times when therapists were not being video-recorded.

The high treatment integrity scores support the statement made by Downs, Downs and Rau (2008) that in order for DTT to be administered with a high level of treatment integrity regular supervision of DTT programmes is required. The dyads with high treatment integrity scores did receive regular supervision either every two weeks or once per month. The treatment integrity scores from Dyad 1 were the exception as the scores were less than 50% even though the therapist received supervision once per month. Dyad 2 (the lowest treatment integrity scores) received irregular and infrequent supervision. However, it should not be assumed that higher overall treatment integrity scores will result in high treatment integrity on every DTT component as shown in the MTM.

Follow-up data were recorded for four dyads and increases in treatment integrity scores were observed for three dyads once errors identified during the baseline phase had been reported to consultants and therapists retrained. However, given the high levels treatment integrity during the baseline phase a ceiling effect was observed much like in the Volkert, Lerman, Trosclair, Addison, and Kodak (2008) study. Experiment 1 of the Volkert et al. study compared the rates at which acquisition skills were acquired using different DTT task sequencing procedures. One condition was a constant condition that consisted of serial training (i.e., training one acquisition stimulus at a time until mastered). The other two procedures were variations of interspersed training (i.e., interspersing mastered trials with acquisition trials). The results showed that skill acquisition occurred rapidly during the constant condition. Skills acquisition was so rapid in the constant condition that the relative effectiveness of the interspersed training procedures could not be demonstrated. A similar ceiling effect was encountered in this study. Although correcting the errors identified during the baseline phase increased treatment integrity scores in the follow-up phase, the increases were small (as seen in Figure 5.8 and 5.9) and the benefits of correcting these treatment integrity errors were somewhat obscured. Unfortunately, it was not possible to obtain follow-up results for Dyads 1 and 2 who had baseline treatment integrity scores similar to those of Leblanc et al. (2005), Sarokoff and Sturmey (2004), and Dib and Sturmey (2007). Despite the aforementioned problems the graphs (Figures 5.1 – 5.7) of individual treatment integrity error sequences still produced meaningful and practically useful results as several error sequence reached the .05 criterion.

Treatment drift was observed during the follow-up phase. Treatment drift occurs when the therapist alters the treatment protocol of a procedure which in turn affects the treatment integrity of the procedure (Gresham, Gansle, & Noell, 1993; Peterson, Homer, & Wonderlich, 1982). For all follow-up dyads the probability of some treatment integrity error sequences were reduced to .00 immediately following the meeting with the therapists and consultants. Over time some of the probabilities returned to their baseline phase levels. The Dyad 8 INCORRECT – INCORRECT sequence is an example of treatment drift. The probability of the sequence was reduced to .00 in the first four blocks of 15 events (Figure 5.7) before increasing to its baseline probability in the next two blocks.

The finding of treatment drift suggests that the methods used to retrain therapists were only sufficient to produce changes in the short-term but not for maintenance. Consultants did not provide any additional retraining following the meeting and the email summary of the meeting. It may also be the case that, over time, a feedback system has been created where the therapist changed the treatment protocol based on their experience with the learner. For example, if a S^D was prescribed to be delivered once but the therapist reliably presents multiple S^D then it may the case that the therapist has learned that the learner is more likely to respond following the second S^D than the first and alters the treatment protocol to accommodate that behaviour. Perhaps correct learner responses reinforce this type of therapist drift and a dynamic is created where the learner is training the therapist rather than what was intended. Unplanned alterations of the treatment protocol will reduce the integrity with which the procedure is administered.

The study provided several interesting findings. One of these was that in most cases ITI were more likely not to last for their prescribed duration when the next S^D was a mastered S^D rather than an acquisition S^D. This finding suggests that certain types of errors may be trial-type specific. That is, certain errors may be more likely to occur during mastered trials than acquisition trials and vice versa. This question is an empirical one and awaits further research.

The finding of the ITI duration error supports the findings of Belfiore, Fritts, and Herman (2008), and Koegel, Glahn, and Nieminen (1978) who also found treatment integrity errors in regards to correctly administering the ITI. Belfiore et al. video-recorded three staff-members administering DTT to learners diagnosed with autism, and assessed the treatment integrity of the within-trial components. Specifically, they recorded whether the S^D was presented in a clear manner, whether therapists waited between 3 – 5 s for a learner response, how learners responded (i.e., correct or incorrect), whether the therapists provided the appropriate consequence, and whether the ITI lasted for the prescribed duration (between 2 and 5 s). Baseline results showed that, on average, the ITI was administered as prescribed on 5% of trials across all three therapists. The average for other parts of the trial ranged between 85% and 99%.

The intervention consisted of the programme consultants explaining to the therapists how each part of the procedure should be administered. Training also consisted of staff self-monitoring; staff watched samples of their video-recording and scored whether they presented the component as prescribed. As part of the intervention staff received additional re-training on how to apply the ITI as prescribed. Following intervention the treatment integrity for the correct application of the ITI increased to 81% on average. However, it was still lower than the average for the other components which ranged between 98% and 100%. Maintenance data were recorded four weeks following the intervention. On average, the ITI was administered as prescribed on 80% of all trials across the three therapists. The average for all other components was 100%. The authors did not provide any possible explanation for these results.

Noell, Gresham, and Gansle (2002) stated that learner performance may be less affected by decreased levels of treatment integrity when learners have been exposed to a procedure for a prolonged period of time. Learner 1 had been receiving DTT from that particular service-provider and therapist for 12 months when data collection began so the statement by Noell et al. may be applicable here. However, any statements regarding learner performance should be made with caution given that no formal analyses of learner performance were conducted. It could also be the case that the majority of correct responses were made during mastered trials.

Error Sequence Analyses

One of the main aims of the study was to assess DTT programmes so that treatment integrity errors could be identified and corrected. Administering DTT with a high degree of integrity will increase the quality of teaching that the learner receives and allows for increased accuracy when making clinical decisions regarding whether skills have been mastered. The following section will provide a discussion on the researched adverse effects of specific DTT treatment integrity errors to put into context why these errors should be corrected (even if they are occurring with low frequencies and probabilities). Research from both the experimental analysis of behaviour and applied research will be discussed.

One error that occurred for some dyads was providing consequences prescribed for correct responding following incorrect responses (henceforth to be referred to as error reinforcement, indicated in the results tables as INCORRECT – CORRECT CONSEQUENCE). This error sequence reached the .05 criterion for Dyads 6 and 7 during the follow-up phase. Error reinforcement has been shown to have detrimental effects on discrimination learning (Davison & McCarthy, 1980). Davison and McCarthy (1980) employed a signal detection procedure where pigeons were required to peck the left key when the sample keylight had been illuminated for 5 s and the right key when the keylight had been illuminated for 10 s. The normative consequence for incorrect responding (left key peck for 10 s keylight and right key peck for 5 s keylight) was a 3 s blackout. Across conditions the probability of error reinforcement was varied between .10 and .90. The results showed that the percentage of correct discriminations decreased as the probability of error reinforcement increased. The Davison and McCarthy results illustrated the detrimental effects of error reinforcement on discrimination learning.

DiGennaro Reed, Reed, Baez and Maguire (2011) found similar detrimental effects of error reinforcement when systematically manipulating the frequency with which incorrect responses were reinforced during DTT. Errors were reinforced during 0% of trials, 50% of trials and 100% of trials. The results showed that learner performance (measured as percentage correct) decreased as the rate of error reinforcement (50% and 100%) increased. DiGennaro Reed et al. (2011) found that for two learners reinforcing errors following half of all incorrect responses were just as detrimental as reinforcing every incorrect response.

For some dyads errors in relation to the response prompt were identified. For Dyads 2 and 8 the main problem surrounding the response prompts was that the therapist administered multiple response prompts unsystematically within a trial as oppose to administering them in least-to-most fashion across trials. Other frequently observe response prompt errors involved therapists providing multiple successive prompts (e.g., Dyad 3) and administering response prompts following incorrect learner responses instead of the prescribed error consequence (e.g., Dyads 1, 2, 4, 6, 7 & 8).

Two studies (Holcombe, Wolery, & Snyder, 1994; Noell et al., 2002) have been conducted examining the effects of sequential errors involving response prompts during DTT. Holcombe et al. manipulated the integrity with which controlling prompts were delivered. Prompts were prescribed to be delivered either simultaneously with the S^D (0 s delay), or 4 s after the presentation of the S^D. The treatment integrity error was that the prompt was not administered at its prescribed time. The error can be conceptualized as a sequential error because the prompt is not following the S^D as prescribed. During the high integrity condition the prompt was administered as prescribed, and during the low-integrity condition the prompt was administered on approximately half of the trials. The results showed that the high-integrity condition resulted in fewer instructional sessions to mastery, fewer learner errors, and fewer instances of non-responding.

The Noell et al. (2002) study consisted of not administering the response prompt when it was prescribed to be administered. Conditions consisted of 100% treatment integrity, 67% treatment integrity (one-third of trials prompt not administered correctly), and 33% treatment integrity (two-thirds of trials prompt not administered correctly). The Noell et al. results showed that learner performance improved above baseline (no prompts) levels for all participants once treatment integrity levels increased, including the 33% treatment integrity condition. It was interesting to note that for some participants the 67% treatment integrity condition yielded similar results in terms of learner performance than the 100% treatment integrity condition. On average, the 67% and 100% conditions were always more effective than the 33% condition. More research is needed to evaluate the effects of response prompt errors like the ones identified in the current study.

The methods, procedures and results from the current study closely reflect those of Carroll, Kodak and Fisher (2013). Carroll et al. conducted a series of experiments in which they evaluated DTT integrity errors. Carroll et al. observed DTT in the natural environment and took treatment integrity measures on nine different therapist behaviours. These behaviours were securing learner attention, providing clear S^D, making sure learner was ready before delivering the S^D, presenting the S^D twice, providing correct consequences (praise and access to tangible items), administering the controlling prompt, and not responding to problem behaviour. They found that the three most common treatment integrity errors were presenting the S^D more than once, not providing consequences following correct learner responses, and not administering the controlling prompt. The first two errors were also identified in the current study for several dyads.

Experiment 3 of the Carroll et al. (2013) study consisted of a component analysis to evaluate the effect that each one of these errors had on learner performance which was measured as percentage correct. The S^D presentation error consisted of presenting the S^D twice with the second S^D consisting of different wording from the first. Programmed errors were administered during 67% of all trials. All other DTT components were administered with high levels of treatment integrity (above 90%).

The Carroll et al. (2013) results showed that percentage correct responses for two learners were lowest when the S^D was presented twice and the controlling prompt not provided. Percentage correct responses were lowest for the third learner when correct consequences were not provided following correct responses. The finding that different errors had different effects on learner performance is interesting and warrants further research.

Learner self-corrections were identified for seven out of the eight dyads during the baseline phase. The self-correction error occurs when the therapist allows the learner to emit another response, usually the correct response, following an initial incorrect learner response. Lovaas (2003) characterised learner self-corrections as the learner engaging in a lose-shift strategy. Lovaas further stated that learner self-corrections should be avoided as learners can learn wrong associations between the S^D and the response. Both Lovaas and Leaf and McEachin (1999) recommended that learner self-corrections should be followed by the therapist ending the trial with the prescribed consequence for incorrect responding and repeating it so that it can be determined whether the learner can emit the required response for that particular S^D. Anecdotally, the video-recordings showed that therapists often provided correct consequences following learner self-corrections. The prescribed protocol following learner self-corrections were as described by Lovaas and Leaf and McEachin for all dyads.

More research is needed to evaluate and identify DTT treatment integrity errors. Carroll et al. (2013) mentioned two limitations of the current studies examining the effect of DTT treatment integrity errors. The first limitation is that the ecological validity of the studies is compromised because researchers select treatment integrity errors to be evaluated arbitrarily and these errors may not be an accurate reflection of frequently occurring treatment integrity errors in the natural environment. Studies such as the current one are needed to further identify commonly occurring DTT errors to increase the ecological validity of controlled studies.

The second limiting factor is that studies usually evaluate only one type of treatment integrity error when it is possible that several errors can occur within the same trial. For example providing a correct consequence following an occurrence of learner self-correction is trial which contains two treatment integrity errors. Therefore, research evaluating individual versus multiple within-trial errors is needed. The results from the current study have shown that DTT programmes can consist of multiple treatment integrity errors.

The .05 Criterion

Given the large amount of data output, even when using a simple one-step Markov model, a criterion was needed to provide a threshold beyond which error sequences had to occur before intervention. For example, during the baseline phase for Dyad 8 11 variables were coded which resulted in 121 different one-step outcomes. In this study the .05 criterion was selected arbitrarily to identify error sequences. The criterion was useful in identifying treatment integrity error sequences which the therapist and consultants also recognized as problematic. The criterion was also useful in reducing the output needed by highlighting only certain sequences.

However, the .05 criterion is not without fault. If an event occurred with relatively low frequency compared to other events then it was more likely that an error associated with that event would reach the .05 criterion. For example, if 100 correct responses occurred then errors associated with correct responses had to occur five times before reaching the criterion. On the other hand, if 1000 ITI were recorded then errors involving ITI had to occur 50 times before reaching criterion. Therefore, how likely errors were to be identified using this criterion was dependent on relative frequency of the first event in the sequence. Based on this problem it is strongly recommended that the frequency of a sequence also be considered when evaluating error sequences.

Therefore, more research is needed to establish a criterion for selecting which errors to intervene with when using MTM to identify treatment integrity errors. It may the case then that if Error A is occurring with a high-probability and Event B with a low-probability that the occurrence of Error B is actually more detrimental to learner performance than Event A. Under those circumstances it be would be better for the learner to intervene on Error B. As it relates to DTT it is possible that not providing an ITI of prescribed duration is not as detrimental as error reinforcement even if the ITI error is occurring at a higher frequency and probability. Intervention should occur on the error reinforcement error instead.   Research is needed to determine at which level of frequency DTT errors can occur before learner performance is affected. For example, Noell et al. (2002) found that 67% treatment integrity in regards to the correct application of response prompts was in some cases just as effective as 100% treatment integrity of response prompt application. Davison and McCarthy found severe decreases in percentage correct responses once the error reinforcement occurred with a probability .40. Research of this nature will be useful for researchers and consultants when evaluating treatment integrity errors, not just for DTT but also for other interventions consisting of multiple components (Fryling, Wallace, & Yassine, 2012).

Limitations

A procedural limitation of the current study was that for most dyads video-recordings could not be made on consecutive days and in order to collect enough data to produce meaningful and interpretable results video-recordings were made across a number of weeks for most dyads. The original design of study was to record for approximately a week. However, given the several issues experienced (learner and therapist health, and therapist availability) the data collection process was prolonged. One therapist stated in the social validity questionnaire that it would have been helpful if results could have been reported more often. This can be achieved if results are obtained quickly.

Another procedural limitation was that for all dyads there was a delay of several weeks between video-recording the final baseline session and reporting the results to the therapists and consultants. During this time therapists were still administering DTT. Therefore, it was possible that levels of treatment integrity may have changed without reporting the baseline results to the therapists and consultants. It is possible that other extraneous variables may have led to the increases in treatment integrity. For example, the errors that were identified in the baseline data could have been detected and corrected during the regular supervision sessions which the therapists received.

One way to strengthen the design of the study would be to provide feedback according to a nonconcurrent multiple baseline (Baer, Wolf, & Risley, 1968). Using a nonconcurrent multiple baseline strengthens the design of the study because it decreases the likelihood of the effects of history, which in turn strengthens the internal validity of the results (Watson & Workman, 1981). Using a nonconcurrent multiple baseline makes it unlikely that therapist behaviour would be influenced by other extraneous variables. If it is not possible to provide feedback according to a nonconcurrent multiple baseline design then it is suggested that the researcher records and analyse data on the day before the results are to be reported to ensure that the results from the initial analyses are still relevant.

A limitation of the study as it relates to the analyses is that no analyses could be conducted in regards to learner performance. One reason relates to the difficulties discussed above in terms of irregular recording. Making irregular recordings meant that learner progress on acquisition stimuli could not be accurately tracked, especially given that learners received the same DTT programmes from therapists who had not provided consent between video-recordings. Using traditional methods learner performance measures such as percentage correct and trials-to-criterion would not have provided meaningful results.

Another limitation of the study as it that the researcher had no further input regarding therapist retraining after presenting the baseline data to consultants and therapists. Additional therapist retraining was not reported by any of the consultants following presentation of the baseline results. Future studies should provide specific types of therapist retraining once the baseline results have been obtained. However, the follow-up phase results did provide some interesting findings, and as stated earlier the results suggest that vocal feedback alone produces short-term treatment integrity increases that could not be maintained in the long-term.

Chapter VI

Study 2

Introduction

During discrete-trial teaching (DTT), numerous trials are administered resulting in several transitions between trials. Given these numerous between-trials transitions the data can be analysed using various sequential analysis methods. It seems appropriate to develop and conduct analyses to investigate whether events occurring during one particular trial have any correlation with events occurring on subsequent trials. That is, can clinically significant results be produced when analysing DTT data on a between-trials basis.

The DTT literature contains a few examples of when DTT data have been analysed on a between-trials basis (Barbetta, Heron, & Heward, 1993; Heckaman, Alber, Hooper, & Heward, 1998; Lerman, Dittlinger, Fentress, & Lanagan, 2011). All of the studies used conditional probability analyses, and collectively the results suggest that events occurring during the course of a trial can have predictive value concerning events and outcomes on subsequent trials. Heckaman, Alber, Hooper, and Heward (1998) investigated the probability of disruptive learner behaviour following two different types of prompt-fading procedures (least-to-most and progressive time-delay). The results showed that for some learners disruptive behaviour was less likely to follow an effective prompt compared to a less effective prompt.

Lerman et al. (2011) compared continuous and discontinuous data collection methods in terms of how well these methods could predict learner performance across all trials. Data were recorded on the first trial and the analyses conducted to calculate the probability that the learner would respond correctly on more than 50% of all subsequent trials. If the learner responded correctly on the first trial the probability of responding correctly on more than

50% of subsequent trials was .92, and if the learner responded incorrectly on the first trial the probability was .26.

Barbetta et al. (1993) compared the effectiveness of two error-correction procedures. The analyses consisted of calculating the probability that the learner would make a correct response the next time the stimulus for which they responded incorrectly on a previous trial was presented. The two procedures were active student response and no response. The active response procedure consisted of the therapist providing vocal feedback that the response was incorrect, modelling the correct response, and repeating the S^D. The no response procedure consisted of vocal feedback only. Five of the six learners made more correct responses during the active student response procedure, and for one learner no difference between procedures was detected.

Prompt-fading procedures, data collection methods, and error-correction procedures are three examples of where sequential analyses can be used to conduct between-trials analyses. The areas examined in this study are the local effects of within-trial error-correction treatment integrity errors, between-trials error-correction procedure treatment integrity, and DTT task sequencing.

Within-Trial Error-Correction Treatment Integrity Errors

The results from Study 1 showed that all therapists made treatment integrity errors following incorrect learner responses. These errors were learner self-corrections, administering other events (such as prompts and S^D) following incorrect responses instead of the prescribed error consequence, not providing ITI of prescribed duration, and error-reinforcement. It was of interest to investigate whether these errors affected learner performance during acquisition trials.

Based on previous research, it is expected that learners would be less likely to respond correctly following these within-trial error-correction errors. The combined findings of

DiGennaro Reed et al. (2011), and Davison and McCarthy (1980) demonstrated that reinforcing incorrect responses can have detrimental effects on response accuracy during discrimination learning. Although not directly associated with incorrect learner responses, research by Holcombe, Wolery, and Snyder (1994), Noell, Gresham, and Gansle (2002), and Grow et al. (2009) found that presenting the prompt at a time that it was not prescribed resulted in a greater number of incorrect responses. All of the aforementioned studies presented their results using traditional whole-sessions methods. None of the studies conducted any within-sessions analyses to evaluate learner performance on acquisition trials immediately following acquisition trials during which the therapist made these treatment integrity errors. One of the aims of this study is to provide such analyses.

Between-Trial Error-Correction Treatment Integrity

Error-correction procedures follow incorrect learner responses to increase the likelihood of correct responses during future trials. Several different types of error-correction procedures exist (McGhan & Lerman, 2013). These include providing verbal-feedback that the response was incorrect (e.g., Smith, Mruzek, Wheat, & Hughes, 2006), modeling the correct response (e.g., Barbetta et al., 1993), and requiring the learner to repeat the response for a certain number of correction-trials (e.g., Rogers & Iwata, 1991; Worsdell et al., 2005).

Effective error-correction procedures are those which increase the probability or frequency of correct learner responses on future trials following incorrect responses (McGhan & Lerman, 2013). Research suggests that error-correction procedures requiring more active learner participation appear to be more effective than procedures that require less active learner participation although results have not always been conclusive (Barbetta et al., 1993; McGhan & Lerman, 2013; Rogers & Iwata, 1991; Worsdell et al., 2005).

The Rogers and Iwata (1993) study compared three error-correction procedures. The first procedure consisted of re-presenting the S^D from the previous error-trial until the learner made a correct response. The second procedure involved presenting stimuli from a different DTT programme. The third procedure was providing verbal-feedback only. The data were analysed as the cumulative number of correct responses across blocks of 20 trials. Of the seven learners, five made more correct responses during the first and second procedures compared to the verbal-feedback only procedure.

McGhan and Lerman (2013) compared the effectiveness of four error-correction procedures: 1. Verbal-feedback only (all stimuli were removed and no error-correction trials were administered); 2. Model (the therapist models the correct response on the next trial, and no error-correction trials were administered); 3. Active student response (the therapist provides a gestural prompt and vocal feedback about what the correct response was and re-presents the S^D); and 4. Direct rehearsal (the therapist provides a gestural prompt and vocal feedback about what the correct response was, and re-presented the S^D from the previous trial. Following the prompted trial, if the learner responded correctly, the therapist re-presented the S^D until the learner made three consecutive unprompted correct responses). The procedures were compared in terms of trials-to-criterion and percentage correct learner responses.

The results from the McGhan and Lerman (2013) study showed that the model error-correction procedure resulted in the fewest trials-to-criterion for four learners, and the active student response procedure for one learner. The vocal feedback procedure did not result in mastery of any target stimuli for one learner and two target stimuli for another. The verbal-feedback only procedure was the only error-correction procedure in which some stimuli were not mastered.

Part 1 of the Worsdell et al. (2005) study compared error-correction procedures consisting of single-response repetitions, and multiple-response repetitions. The single-response repetition procedure consisted of one re-presentation of the S^D with a model prompt.

The multiple-response repetition procedure consisted of five prompted re-presentations of the S^D. The data were analysed using the total number of skills mastered, and the mean number of correct responses per session. The results showed that all learners performed better during the multiple-response procedure.

Some of the error-correction procedures used in the aforementioned studies also included a between-trials component. That is, certain types of trials had to follow incorrect learner responses. The multiple-response procedure required that the trial following the incorrect response be prompted and the same S^D be presented for five consecutive trials. The active student response procedure also required the re-presentation of the S^D, whereas vocal feedback required the presentation of a different S^D.

Some studies that investigated error-correction procedures do provide treatment integrity data concerning the application of the entire procedure. Barbetta et al. (1993) provided treatment integrity data of whether the therapist correctly administered the active student response and no response error-correction procedures. Treatment integrity was 100% for both procedures. Worsdall et al. (2005) reported treatment integrity scores of 95%, Rogers and Iwata (1991) stated that "no procedural variations were reported," (p. 778), and. Smith et al. (2006) reported 95% treatment integrity. It is important to point out that treatment integrity was always presented as percentage correct and that these studies were conducted in highly controlled environments.

Little is known about the degree of treatment integrity with which error-correction procedures are implemented in the natural environment. It seems plausible that a therapist can administer DTT with a high degree of within-trial treatment integrity while neglecting the correct application of the error-correction procedure, specifically when it has a between-trials component. For example, if the error-correction procedure requires the therapist to re-present the S^D in the trial following incorrect responses, and they do not, then the treatment integrity of the error-correction procedure has been compromised even though the within-trial components may have been administered as prescribed. The opposite also seems plausible as the therapist can administer the within-trial components with low levels of treatment integrity, but the error-correction with a high degree of treatment integrity.

The prescribed error-correction procedures used in the current study all involved a between-trial component. That is, there had to be either a re-presentation of the S^D (prompted or unprompted), or the presentation of a different S^D. Therefore, if the prescribed error-correction procedure is to re-present the S^D with a prompt and the therapist presents a different S^D without a prompt, a treatment integrity error of the error-correction procedure has occurred. This type of error is a sequential error, and it seems appropriate to conduct a sequential analysis investigating the transitions between trials following incorrect responses in order to evaluate the integrity of the error-correction procedure in a manner similar to Study 1. Given that the within-trial treatment integrity scores for several dyads were above 80% (overall and for acquisition trials) it can be assessed whether high within-trial treatment integrity scores resulted in high levels of between-trial error-correction treatment integrity.

Task Sequencing

Another factor that influences learner performance during acquisition trials is task sequencing. Task sequencing refers to the different procedures used to order S^D to increase acquisition and maintain mastery (Lovaas, 2003). These procedures are serial training, concurrent training, combined training and interspersed training. The task sequencing procedures relevant to this study are concurrent and interspersed training. Concurrent training consists of training several acquisition stimuli concurrently. Each acquisition stimulus is trained in the presence of other acquisition stimuli (Cuvo, Klevans, Borakove, Van Landuyt, & Lutzker, 1980; Lovaas, 2003). The order in which the different acquisition stimuli are presented is determined randomly, a process known as random rotation (Lovaas, 2003). Concurrent training is also referred to as simultaneous training (Cuvo et al., 1980).

Interspersed training consists of interspersing acquisition stimuli with stimuli which the learner has already mastered (e.g., Dunlap, 1984; Koegel & Koegel, 1986; Neef, Iwata, & Page, 1980; Weber & Thorpe, 1992). For example, a DTT programme using interspersed training consisting of 20 trials may have 10 acquisition trials and 10 mastered trials presented in a random order.

Generally, studies comparing different task sequencing procedures have shown interspersed training to be superior to other task sequencing procedures (e.g., Dunlap & Koegel, 1980; Koegel & Koegel, 1986). Interspersed training has also been associated with greater positive learner affect (Dunlap, 1984). Few studies have been conducted to directly compare the relative effectiveness of interspersed and concurrent training. Two such studies were Dunlap (1984), and Rowan and Pear (1985).

Dunlap (1984) compared serial and concurrent training to interspersed training when teaching multiple acquisition skills to learners with intellectual disabilities. Serial training consists of repeated presentations of the same acquisition S^D until it has been mastered. Once the skill had been mastered the next acquisition S^D is introduced and taught until mastered. All skills are taught in this manner until all skills have been acquired. Serial training is also referred to as mass trials (Lovaas, 2003) and successive training (Cuvo et al., 1980). Prompts are administered as prescribed during initial training, and faded over time. Prompt fading is an essential part of the procedure (Lovaas, 2003).

The interspersed mastered activities in the Dunlap (1984) study were selected following parent interviews and if learners could perform tasks with at least 80% accuracy during unprompted probe trials. The dependent variable in this study was trials-to-criterion. The data were presented as the average number of trials it took the learner to acquire each of the acquisition tasks across all of the sessions for each condition. The results showed that mastery was obtained from fewer trials with interspersed training.

Rowan and Pear (1985) compared concurrent and interspersed training when teaching three learners with intellectual disabilities to name pictures. The data were analysed in terms of the cumulative number of stimuli mastered, percentage accuracy during unprompted probe trials, mean number of trials needed to reach criterion, and the total number of trials administered.   Results were also presented for generalisation and maintenance. Interspersed training was superior in terms of the number of stimuli mastered, but the generalisation and maintenance data were inconclusive.

Other experiments have also been conducted comparing the relative effectiveness of various other task sequencing procedures (e.g., Cuvo et al., 1980; Doyle, Wolery, Ault, Gast, & Wiley, 1989; Dunlap & Koegel, 1980; Koegel & Koegel, 1986; Panyan & Hall, 1978; Schroeder & Baer, 1972; Waldo, Guess, & Flannagan, 1982; Weber & Thorpe, 1992). The main methods of analyses and data presentation were changes in response accuracy between successive probe sessions, trials-to-criterion for each procedure, percentage correct unprompted acquisition trial responses, number of skills mastered, and total instructional time. None of the aforementioned studies conducted any tests, probes, or analyses to specifically investigate whether correct unprompted acquisition trial responses were more likely following mastered trials than acquisition trials. Such a conclusion appears to be inferred from the between-session analyses which indicate quicker skill acquisition rates, and fewer trials-to-criterion when using interspersed training compared to serial and concurrent training.

Consider a scenario in which a researcher is comparing concurrent and interspersed training in terms of correct unprompted acquisition trial responses per block of 20 trials. The learner makes 16 correct acquisition trial during interspersed training, and 10 during concurrent training (a typical whole-sessions analysis). The conclusion is that interspersed training is more effective because the learner makes more correct acquisition trial responses. The assumption is that the procedure was more effective because of interspersing the mastered trial in with the acquisition ones. This is a reasonable conclusion given that interspersing the mastered trials is the main difference between the procedures.

However, analysing the results on a trial-by-trial basis (i.e. sequential analysis) may reveal features within the data that are obscured by the whole-sessions analysis. For example, during interspersed training it is possible for either a mastered or an acquisition trial to precede a correct acquisition trial response. Therefore, it would be of interest to know whether correct acquisition trial responses are more likely following acquisition or mastered stimuli. Such analyses may reveal that it is equally likely that the learner makes a correct acquisition trial response following an acquisition trial than if a mastered trail was administered. This type of result could call into question the assumed causal mechanism as to why interspersed training produces more correct responses on acquisition trials than another task sequencing procedure such as concurrent training. This type of research question can only be addressed if data are analysed on a trial-by-trial basis.

It appears that there is a gap in the DTT task sequencing literature as it relates to conducting within-session analyses. The question of whether correct unprompted acquisition trial responses are more likely following mastered trials than acquisition trials is a question which consists of analysing sequences of two or more events. Therefore, it appears appropriate to analyse the data using one or more sequential analysis methods in order to investigate sequences of interest that may have clinical importance. This study will provide such analyses.

Aims of Study

Study 2 had three aims. First, to extend the findings from Study 1 and investigate how within-trial error-correction treatment integrity errors affected learner responding on acquisition trials; second, to introduce a method of evaluating between-trial treatment integrity for the error-correction procedures that had a between-trials component; third, to compare the relative effectiveness of concurrent and interspersed task sequencing procedures using extended Markov chain analyses. The main question for the task sequencing analyses was whether correct unprompted acquisition trial responses were more or less likely immediately following one or more mastered trials than they were when they followed one or more acquisition trials.

Method

Participants

Dyads 1, 3, 4, 6, 7, and 8 participated in this study. The dyads consisted of the same therapists, learners and consultants described in Study 1.

Event Definitions

The events recorded during the baseline phase for Study 1 were combined so that they formed individual trials. Once the individual trials were created they were coded according to trial type (mastered or acquisition), learner response (correct or incorrect), whether the response was prompted or not, and it was noted whether all within-trial components were administered as prescribed. If within-trial treatment integrity errors occurred, the type of error was noted (e.g., self-corrections, the ITI not lasting for its prescribed duration etc.). The sequential order of the trials was maintained. Not all parts of the trial were used in each of the analyses. For example, if the event of interest was learner response (correct or incorrect), then other information such as type of S^D, whether the responses was prompted, and whether the trial was administered as prescribed, were not included in the analyses. The following paragraphs contain a description of how trials were coded for each of the analyses.

Within-Trials Error-Correction Treatment Integrity Analyses. Only acquisition trials were used for the within-trial error-correction treatment integrity analyses, however not all acquisition trials were used. Only acquisition trials that were immediately followed by another acquisition trial were included in the analyses. Trials in which the learner made correct responses were coded as CORRECT RESPONSE. If a prompt was administered and the learner made a correct response it was coded as PROMPTED CORRECT RESPONSE. If the learner made an incorrect response it was coded as INCORRECT RESPONSE. For incorrect learner responses it was also indicated whether the error-correction procedure was administered as prescribed. If the within-trial error correction procedure was administered as prescribed, the trial was coded as INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT. If the within-trial error-correction procedure was not administered as prescribed, the trial was coded as INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT. Within-trial error-correction treatment integrity errors (as identified in Study 1) consisted of learner self-corrections, not providing the error-consequence following incorrect responses, providing prompts following incorrect responses, not allowing ITI of prescribed duration, and providing a correct consequence when the learner made incorrect responses. If a trial contained any of these errors it was scored as containing a treatment integrity error.

Between-Trials Error-Correction Treatment Integrity Analyses. The between-trial error-correction treatment integrity analyses were conducted only for incorrect responses that occurred with acquisition S^D. Incorrect acquisition trial responses were scored as INCORRECT RESPONSE, and if the response was prompted it was scored as PROMPTED INCORRECT RESPONSE. Once incorrect responses occurred, the trial immediately following the incorrect response was coded as to whether it contained the same S^D, a different S^D or a mastered S^D. If the trial contained the same S^D, it was coded as SAME SD. If the trial contained a different S^D, it was coded as DIFFERENT SD. If the trial was a mastered trial, it was coded as MASTERED. If the same S^D was re-presented with a prompt it was coded as SAME SD WITH PROMPT. If a different S^D was presented with a prompt, it was coded as DIFFERENT SD WITH PROMPT. Incorrect responses that occurred when the error-correction procedure was to provide verbal-feedback only were coded as INCORRCT RESPONSE (VOCAL FEEDBACK).

Task Sequencing Analyses. For the task sequencing analyses, correct learner responses that occurred during mastered trials were coded as MASTERED CORRECT RESPONSE, and incorrect mastered trial responses were coded as MASTERED INCORRECT RESPONSE. Correct acquisition trial responses were coded as ACQUISITION CORRECT RESPONSE, and incorrect acquisition trial responses were coded as ACQUISITION INCORRECT RESPONSE. A prompted correct acquisition trial response was coded as ACQUISITION CORRECT RESPONSE (P).

For the task sequencing analyses, trials were separated (for each dyad) on the basis that they were administered using concurrent or interspersed training. Once trials for each task sequencing procedure had been separated, the trials from each task sequencing procedure were combined. The sequential order of the trials was maintained.

For all analyses, prompted trials included all trials that contained intra or extra-stimulus prompts. Intra-stimulus prompts were used for Dyads 3, 7, and 8. Intra-stimulus prompts consisted of printing the name of the stimulus onto the stimulus. Incorrect responses across all analyses included trials that consisted of multiple learner responses (such as self-corrections) and non-responses.

Analyses

All analyses were conducted using Obswin 3.4 software, and Microsoft Excel 2010 ©. One-step Markov transition matrices (MTM) were used to analyse data for the within-trial error-correction treatment integrity errors. Modified conditional probability matrices were used when analysing the between-trials treatment integrity of the error-correction procedures. Three different types of analyses were conducted for the DTT task sequencing procedures. The first was a summary table of learner responses on acquisition trials; the second was calculating the conditional probabilities of correct acquisition trial responses following mastered or acquisition S^D; the third consisted of extended three-event Markov chain analyses. Chi-square tests were conducted to assess whether sequences resulting in unprompted correct and incorrect acquisition trial responses were occurring more or less than expected.

Results

Within-Trials Error-Correction Treatment Integrity Analyses. The analyses were conducted for Dyads 3, 4, 7, and 8. The data for Dyad 1 were excluded because only two instances of INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT were coded. The Dyad 6 results were excluded because only six instances of INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT were recorded. These were not enough for valid and reliable within-subject comparisons to be made.

The data were organised into one-step MTM, and these can be seen in Tables 6.1 to 6.4. The Markov matrices contain the transitional probability and frequency of each transition (in brackets). The probabilities were calculated by dividing the number of times the transition occurred by the number of times the first event in the sequence occurred (as indicated in the SUM column of the matrices). The events in the first column were the first trial in a two-trial sequence, and the events in columns two to five were the second trial in the sequence. It was of interest to investigate if the learner was more likely to make correct responses on the trial following the correct application of the within-trial error-correction procedure, compared to if the therapist made a within-trial error-correction treatment integrity error.

The Dyad 3 results can be seen in Table 6.1. The first two rows display all possible outcomes on the trials following unprompted and prompted correct learner responses. Three hundred and twenty three correct responses (166 + 157) were included in the analyses. The third and fourth rows display all possible outcomes on the trial following incorrect learner responses. One hundred and forty incorrect responses were used in the analyses (125 + 15). The analyses concern events following incorrect learner responses.

As seen in Table 6.1, the probability of the learner making a correct response following the correct application of the error-correction procedure was calculated by adding the probabilities in the first (.35) and second (.40) cells of the third row of the table. The sum of these probabilities was .75. The probability of the learner making an incorrect response following the correct application of the error-correction procedure was calculated by adding the probabilities of the third (.19) and fourth (.06) cells in the third row. The sum of these probabilities was .25. To calculate the probability of a correct response following the incorrect application of the within-trial error-correction procedure, the probabilities of the first (.33) and second (.40) cells of the fourth row were summed. The probability was .73. The probability of an incorrect response following the incorrect application of the within-trial error-correction procedure was calculated by adding the third (.20) and fourth (.07) cells of the fourth row. The probability was .27. Thus, there was little difference (.02) in the probability of correct learner responses whether Therapist 1 provided accurate error-correction or not.

Table 6.2 displays the Dyad 4 results. Table 6.2 shows that 98 correct (56 + 42) and 56 (29 + 27) incorrect responses were included in the analyses. The probability of the learner making correct acquisition trial responses following the correct therapist application of the error-correction procedure was .94 (.66 + .28), and the probability of making incorrect responses was .06 (.03 + .03). The probability of correct learner responses following the incorrect therapist application of the within-trial error-correction procedure was .63 (.30 + .33), and the probability of incorrect responses was .37 (.26 + .11). The probability was .31 higher that the learner would make a correct trial response if the therapist administered the within-trial error-correction procedure as prescribed.

Table 6.1

Dyad 3 Markov transition matrix for within-trial error-correction treatment integrity errors analyses. The table contains the transitional probabilities and frequencies (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to five were the second trial in the sequence.

First Trial	Second Trial				
	CORRECT RESPONSE	PROMPTED CORRECT RESPONSE	INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	SUM
CORRECT RESPONSE	.64 (106)	.04 (7)	.30 (49)	.02 (4)	1.00 (166)
PROMPTED CORRECT RESPONSE	.11 (18)	.69 (109)	.16 (25)	.03 (5)	1.00 (157)
INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	.35 (44)	.40 (50)	.19 (24)	.06 (7)	1.00 (125)
INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	.33 (5)	.40 (6)	.20 (3)	.07 (1)	1.00 (15)
SUM					463

Table 6.2

Dyad 4 Markov transition matrix for within-trial error-correction treatment integrity errors analyses. The table contains the transitional probabilities and frequencies (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial				
	CORRECT RESPONSE	PROMPTED CORRECT RESPONSE	INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	SUM
CORRECT RESPONSE	.52 (29)	.13 (7)	.18 (10)	.18 (10)	1.00 (56)
PROMPTED CORRECT RESPONSE	.24 (10)	.52 (22)	.12 (5)	.12 (5)	1.00 (42)
INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	.66 (19)	.28 (8)	.03 (1)	.03 (1)	1.00 (29)
INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	.30 (8)	.33 (9)	.26 (7)	.11 (3)	1.00 (27)
SUM					154

Table 6.3 shows the results for Dyad 7. The analyses included 126 (76 + 50) correct and 31 (15 + 16) incorrect responses. The probability of correct learner responses on the trial following the correct therapist application of the within-trial error-correction procedure was .93 (.40 + .53), and the probability of another incorrect response was .07 (.00 + .07). The probability of correct learner responses following the incorrect therapist application of the within-trial error-correction procedure was .69 (.44 + .25), and the probability of another incorrect response was .31 (.06 + .25). The probability of the learner making correct responses was .24 higher if the therapist administered the within-trial error-correction procedure correctly.

The Dyad 8 results can be seen in Table 6.4. The analyses included 103 (69 + 34) correct and 88 (51 + 37) incorrect responses. The probability of a correct learner response following the correct therapist application of the within-trial error-correction procedure was .67 (.47 + .20), and the probability of the learner responding incorrectly was .34 (.24 + .10). The probability of correct learner responses following the incorrect therapist application of the within-trial error-correction procedure was .52 (.38 + .14) which was .15 less than if the error-correction procedure was administered correctly. The probability of an incorrect response following the incorrect therapist application of the error-correction procedure on a within-trial basis was .49 (.14 +.35).

The results from the analyses showed that for Dyads 4, 7, and 8, the learners were more likely to make correct acquisition trials if the within-trial error-correction procedure was administered correctly. For Dyad 3 there was a small difference.

Between-trials Error-Correction Treatment Integrity Analyses

The data from Dyads 1, 3, 4, 7, and 8 were included in the analyses. The Dyad 6 data were excluded from the analyses because not enough incorrect acquisition trial responses

Table 6.3

Dyad 7 Markov transition matrix for within-trial error-correction treatment integrity errors analyses. The table contains the transitional probabilities and frequencies (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial				
	CORRECT RESPONSE	PROMPTED CORRECT RESPONSE	INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	SUM
CORRECT RESPONSE	.75 (57)	.09 (7)	.07 (5)	.09 (7)	1.00 (76)
PROMPTED CORRECT RESPONSE	.20 (10)	.70 (35)	.06 (3)	.04 (2)	1.00 (50)
INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	.40 (6)	.53 (8)	.00 (0)	.07 (1)	1.00 (15)
INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	.44 (7)	.25 (4)	.06 (1)	.25 (4)	1.00 (16)
SUM					157

Table 6.4

Dyad 8 Markov transition matrix for within-trial error-correction treatment integrity errors analyses. The table contains the transitional probabilities and frequencies (in brackets) of each transition. The event in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial				
	CORRECT RESPONSE	PROMPTED CORRECT RESPONSE	INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	SUM
CORRECT RESPONSE	.61 (42)	.03 (2)	.23 (16)	.13 (9)	1.00 (69)
PROMPTED CORRECT RESPONSE	.15 (5)	.56 (19)	.21 (7)	.09 (3)	1.00(34)
INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT	.47 (24)	.20 (10)	.24 (12)	.10 (5)	1.00 (51)
INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT	.38 (14)	.14 (5)	.14 (5)	.35 (13)	1.00 (37)
SUM					191

were recorded to produce meaningful results. For all other dyads at least 30 incorrect learner responses were recorded. The data were organised into conditional probability matrices and these can be seen in Tables 6.5 to 6.9. The conditional probability matrices contain the conditional probabilities and the frequencies (in brackets) of each transition. The probabilities were calculated by dividing the number of times the transition occurred by the total number of times the first event in the sequence occurred (as indicated in the SUM column of the matrices). The events in the first column were the first trial in a two-trial sequence, and the events in columns two to six were the second trial in the sequence. In the conditional probability matrices all prescribed sequences are italicised and in bold text. A probability criterion of .90 was selected (the same as for the within-trial treatment integrity) to indicate satisfactory levels of treatment integrity for the between-trials error-correction procedures.

The results for Dyads 1, 3, and 4 can be seen in Tables 6.5 to 6.7. The error-correction procedure for Dyads 1, 3, and 4 consisted of two one-step transitions. Unprompted incorrect responses had to be followed by a prompted re-presentation of the same S^D from the previous trial (i.e., INCORRECT RESPONSE – SAME SD WITH PROMPT, second cell of first row). If the learner made a correct response on the prompted re-presentation of the S^D then the second transition was to re-present the same S^D without the prompt (i.e., SAME SD WITH PROMPT (IF CORRECT RESPONSE) – SAME SD, first cell in the second row). All prompted incorrect learner responses had to be followed by a prompted re-presentation of the same S^D on the next trial (i.e., PROMPTED INCORRECT RESPONSE – SAME SD WITH PROMPT (second cell of the third row).

The results for Dyad 1 can be seen in Table 6.5. Following unprompted incorrect learner responses, the probability of the therapist re-presenting the same S^D with a prompt was .51, showing that approximately half of all unprompted incorrect responses were

Table 6.5

Dyad 1 between-trials error-correction treatment integrity conditional probability matrix. The table contains the transitional probabilities and frequencies (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence The prescribed sequences are italicised and in bold text.

First Trial	Second Trial					
	SAME SD	SAME SD WITH PROMPT	DIFFERENT SD	DIFFERENT SD WITH PROMPT	MASTERED SD	SUM
INCORRECT RESPONSE	.08 (3)	***.51 (20)***	.23 (9)	.08 (3)	.08 (3)	1.00 (38)
SAME SD WITH PROMPT (IF CORRECT RESPONSE)	***.40 (8)***	.00 (0)	.35 (7)	.00 (0)	.25 (5)	1.00 (20)
PROMPTED INCORRECT RESPONSE	.00 (0)	***1.00 (4)***	.00 (0)	.00 (0)	.00 (0)	1.00 (4)

followed by the prescribed procedure (INCORRECT RESPONSE – SAME SD WITH PROMPT, n = 20). The most likely treatment integrity error following unprompted incorrect learner responses was to present a different S^D (INCORRECT RESPONSE – DIFFERENT SD, p = .23, n = 9). Once the learner had made a correct response on the prompted re-presentation of the S^D, the most likely event to follow was for the therapist to re-present the same S^D. The probability of the sequence SAME SD WITH PROMPT (IF CORRECT RESPONSE) – SAME SD was .40 (n = 8). The two most likely treatment integrity errors were to present a different S^D or a mastered $S^{D.}$ The sequence SAME SD WITH PROMPT – DIFFERENT SD had a probability of .35 (n = 7), and the probability of the sequence SAME SD WITH PROMPT – MASTERED was .25 (n = 5). Perfect treatment integrity was observed following prompted incorrect responses.

The Dyad 3 results can be seen in Table 6.6. The prescribed sequence of INCORRECT RESPONSE – SAME SD WITH PROMPT occurred 51 times with a probability of .36. Approximately one-third of all unprompted incorrect learner responses were followed by the prescribed event. The two most likely treatment integrity error sequences were not administering the prompt when re-presenting the S^D and presenting a different S^D. The probability of the sequence INCORRECT RESPONSE – SAME SD was .32 (n = 45), and the probability of the sequence INCORRECT RESPONSE – DIFFERENT SD was .18 (n = 25). Once the learner had made a correct response, the prescribed sequence of SAME SD WITH PROMPT (IF CORRECT RESPONSE) – SAME SD occurred with a probability of .60 (n = 30). The most likely treatment integrity error was presenting a mastered S^D instead of re-presenting the same S^D without a prompt. The probability of the sequence SAME SD WITH PROMPT – MASTERED was .38 (n = 19). For prompted incorrect learner responses the most likely sequence was the prescribed sequence of PROMPTED INCORRECT RESPONSE – SAME SD (p = .71, n = 5).

Table 6.6

Dyad 3 between-trial error-correction treatment integrity conditional probability matrix. The table contains the transitional probabilities and frequencies (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence. The prescribed sequences are italicised and in bold text.

First Trial	Second Trial					
	SAME SD	SAME SD WITH PROMPT	DIFFERENT SD	DIFFERENT SD WITH PROMPT	MASTERED SD	SUM
INCORRECT RESPONSE	.32 (45)	***.36 (51)***	.18 (25)	.05 (7)	.10 (14)	.100 (142)
SAME SD WITH PROMPT (IF CORRECT RESPONSE)	***.60 (30)***	.00 (0)	.00 (0)	.02 (1)	.38 (19)	1.00 (50)
PROMPTED INCORRECT RESPONSE	.00 (0)	***.71 (5)***	.00 (0)	.29 (2)	.00 (0)	1.00 (7)

The results from Dyad 4 can be seen in Table 6.7. Following unprompted learner responses the probability of the therapist re-presenting the same S^D with a prompt was .55 (*n* = 28). This indicates that approximately half of all unprompted incorrect learner responses were followed by the prescribed event. The most likely treatment integrity error was to re-present the same S^D but without the prompt (INCORRECT RESPONSE – SAME SD, *n* = 21, p = .41). Once the learner had made the correct response the probability of the prescribed sequence SAME SD WITH PROMPT (IF CORRECT RESPONSE) – SAME SD was .80 (*n* = 24). The most likely treatment integrity error was to present a mastered S^D (SAME SD WITH PROMPT – MASTERED, p = .13, *n* = 4). Following prompted incorrect learner responses the prescribed event of presenting the same S^D with a prompt followed with a probability .40 (*n* = 2). The most likely treatment integrity error sequence was to present the same S^D without a prompt (PROMPTED INCORRECT RESPONSE – SAME SD, p = .60, *n* = 3).

The results showed that only one of the prescribed sequences scored above 90% treatment integrity for Dyad 1, and none of the prescribed sequences scored above 90% for Dyads 3 and 4. The most likely treatment integrity errors included re-presenting the S^D without the prompt following incorrect responses, or presenting a different S^D following prompted incorrect learner responses.

The results for Dyads 7 and 8 can be seen in Table 6.8 and 6.9. Two error-correction procedures were used for Dyads 7 and 8. The first procedure involved re-presenting the same S^D from the previous trial with a prompt, regardless of whether the previous trial was prompted or not. The prescribed sequences were RESPONSE INCORRECT – SAME SD WITH PROMPT (second cell in the first row), and PROMPTED INCORRECT RESPONSE– SAME SD WITH PROMPT, second cell in the second row). The second error-correction procedure consisted of verbal-feedback only. Following the verbal-

Table 6.7

Dyad 4 between-trials error-correction treatment integrity conditional probability matrix. The table contains the probability and frequency (in brackets) of each transition. The event in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence. The prescribed sequences are italicised and in bold text.

First Trial	Second Trial					
	SAME SD	SAME SD WITH PROMPT	DIFFERENT SD	DIFFERENT SD WITH PROMPT	MASTERED SD	SUM
INCORRECT RESPONSE	.41 (21)	***.55 (28)***	.02 (1)	.00 (0)	.02 (1)	1.00 (51)
SAME SD WITH PROMPT (IF CORRECT RESPONSE)	***.80 (24)***	.03 (1)	.03 (1)	.00 (0)	.13 (4)	1.00 (30)
PROMPTED INCORRECT RESPONSE	.60 (3)	***.40 (2)***	.03 (1)	.00 (0)	.00 (0)	1.00 (5)

Table 6.8

Dyad 7 between-trial error-correction treatment integrity conditional probability matrix. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence. The prescribed sequences are italicised and in bold text

First Trial	Second Trial					
	SAME SD	SAME SD WITH PROMPT	DIFFERENT SD	DIFFERENT SD WITH PROMPT	MASTERED SD	SUM
INCORRECT RESPONSE	.59 (13)	***.36 (8)***	.00 (0)	.00 (0)	.05 (1)	1.00 (22)
SAME SD WITH PROMPT (IF CORRECT RESPONSE)	.00 (0)	***1.00 (2)***	.00 (0)	.00 (0)	.00 (0)	1.00 (2)
INCORRECT RESPONSE (VOCAL FEEDBACK)	.13 (1)	.00 (0)	***.88 (7)***	.00 (0)	.00 (0)	1.00 (8)

Table 6.9

Dyad 8 between-trial error-correction treatment integrity conditional probability matrix. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence. The prescribed sequences are italicised and in bold text

First Trial	Second Trial					
	SAME SD	SAME SD WITH PROMPT	DIFFERENT SD	DIFFERENT SD WITH PROMPT	MASTERED SD	SUM
INCORRECT RESPONSE	.59 (26)	***.18 (8)***	.16 (7)	.02 (1)	.05 (2)	1.00 (44)
SAME SD WITH PROMPT (IF CORRECT RESPONSE)	.18 (2)	***.64 (7)***	.09 (1)	.09 (1)	.00 (0)	1.00 (11)
INCORRECT RESPONSE (VOCAL FEEDBACK)	.03 (1)	.00 (0)	***.91 (29)***	.00 (0)	.06 (2)	1.00 (32)

feedback, the therapist had to present a different S^D. The prescribed sequence was INCORRECT RESPONSE (VOCAL FEEDBACK) – DIFFERENT SD (third cell in third row).

The results for Dyad 7 can be seen in Table 6.8. The probability of the prescribed sequence INCORRECT RESPONSE – SAME SD WITH PROMPT was .36 (n = 8) indicating that approximately one third of all unprompted incorrect learner responses were followed by the prescribed event. The most likely treatment integrity error was to re-present the same S^D without the prompt (INCORRECT RESPONSE – SAME SD, p = .59, n = 13). Perfect treatment integrity was observed following prompted incorrect learner responses. For the verbal-feedback only error-correction procedure, the probability of the prescribed sequence was .88 (INCORRECT RESPONSE (VOCAL FEEDBACK) – DIFFERENT SD, n = 7).

The Dyad 8 results can be seen in Table 6.9. The probability of the prescribed sequence INCORRECT RESPONSE – SAME SD WITH PROMPT was .18 (n = 8). The results showed that approximately every fifth unprompted incorrect learner response was followed by the prescribed event. The most likely event to follow unprompted incorrect learner response was a re-presentation of the same S^D without the prompt (INCORRECT RESPONSE – SAME SD, p = .59, n = 26).

The most likely event following prompted incorrect learner responses was prompted re-presentation of the same S^D (PROMPTED INCORRECT RESPONSE, n = 7, p = .64). The most likely treatment integrity error was to re-present the same S^D without the prompt. This sequence was observed twice and the probability was .18. For the verbal-feedback only error-correction procedure the probability of the prescribed sequence was .91 (n = 29).

The Dyad 7 and 8 results showed that two out of the six prescribed sequences occurred with a probability of at least .90. The treatment integrity error occurring most often for both dyads was re-presenting the S^D without the prompt following unprompted and prompted incorrect responses.

Task Sequencing Analyses

The aim of the analyses was to investigate and compare the relative effectiveness of concurrent and interspersed training. Table 6.10 contains a summary of learner performance for all dyads on acquisition trials during each task sequencing procedure. Table 6.10 contains the frequencies and probabilities of each acquisition trial outcome during concurrent and interspersed training. The probabilities were calculated by dividing the number of times an outcome was observed by the total number of acquisition trials that were administered for each procedure.

The first acquisition trial outcome to be described is ACQUISITION CORRECT RESPONSE. Dyads 3, 4, and 8 scored a higher proportion of unprompted correct acquisition trial responses during interspersed training. The probability of ACQUISITION CORRECT RESPONSE during interspersed training was .51 for Dyad 3, .46 for Dyad 4 and .48 for Dyad 8. The corresponding probabilities for these dyads during concurrent training were .32 for Dyad 3, .20 for Dyad 4, and .38 for Dyad 8.

For Dyads 6 and 7 the probability of ACQUISITION CORRECT RESPONSE was higher during concurrent training. The probabilities were .83 for Dyad 6 and .54 for Dyad 7. The corresponding probabilities for these dyads during interspersed training were .56 for Dyad 6 and .46 for Dyad 7. The probability of unprompted correct learner responses on acquisition trials were approximately the same between the two procedures for Dyad 1.

Table 6.10

Summary of results for Dyads 1, 3, 4, 6, 7, and 8 comparing learner performance between concurrent and interspersed training for correct, incorrect, and prompted correct acquisition trial responses. The table contains the probability of a trial, and the frequency (in brackets) for the task sequencing procedures. The highest probability values between procedures for each trial type are in bold text.

Dyad	Procedure	Trial Type			
		ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	SUM
1	Concurrent	.41 (44)	.36 (38)	**.23 (25)**	1.00 (107)
	Interspersed	**.43 (45)**	**.37 (39)**	.20 (21)	1.00 (105)
3	Concurrent	.32 (100)	**.47 (148)**	.21 (64)	1.00 (312)
	Interspersed	**.51 (155)**	.19 (56)	**.30 (91)**	1.00 (302)
4	Concurrent	.20 (11)	**.54 (29)**	**.26 (14)**	1.00 (54)
	Interspersed	**.46 (80)**	.29 (50)	.25 (43)	1.00 (172)

Table 6.10 (continued)

Dyad	Procedure	Trial Type			
		ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	SUM
6	Concurrent	**.83 (43)**	.12 (6)	.06 (3)	1.00 (52)
	Interspersed	.56 (77)	**.29 (40)**	**.15 (21)**	1.00 (138)
7	Concurrent	**.54 (59)**	.30 (33)	**.16 (17)**	1.00 (109)
	Interspersed	.46 (46)	**.38 (38)**	**.16 (16)**	1.00 (100)
8	Concurrent	.38 (45)	**.25 (30)**	**.36 (43)**	1.00 (118)
	Interspersed	**.48 (72)**	.19 (29)	.33 (50)	1.00 (151)

The next acquisition trial outcome to be described is ACQUISITION INCORRECT RESPONSE. The results from Table 6.10 shows that the probability of ACQUISITION INCORRECT RESPONSE was higher during interspersed training for Dyad 3 (p = .30) and Dyad 6 (p = .15). The corresponding probabilities during concurrent training were .21 for Dyad 3 and .06 for Dyad 6. Minimal differences in the probabilities between procedures were detected for Dyads 1, 4, 7, and 8.

The probability of prompted correct acquisition trial responses (ACQUISITION CORRECT RESPONSE (P)) was higher during interspersed training for Dyad 6 (p = .29) and Dyad 7 (p = .38). The corresponding probabilities during concurrent training were .12 for Dyad 6 and .30 for Dyad 7. The probability of ACQUISITION CORRECT RESPONSE (P) was higher during concurrent training for Dyad 3 (p = .47), Dyad 4 (p = .54), and Dyad 8 (p = .25). The corresponding probabilities during interspersed training were .19 for Dyad 3, .29 for Dyad 4, and .19 for Dyad 8. The probabilities were approximately the same between procedures for Dyad 1. The summary results from Table 6.10 showed that there were no consistent and systematic differences between the procedures in terms of learner outcomes on acquisition trials.

Task Sequencing: Conditional Probability Analyses

Conditional probability analyses were conducted to compare the probabilities of unprompted correct acquisition trial responses following ACQUISITION CORRECT RESPONSE and MASTERED CORRECT RESPONSE during interspersed training. These analyses were conducted to investigate whether presenting mastered S^D made it more likely for the learner to respond correctly on a subsequent acquisition S^D compared to when an acquisition S^D had been presented. The conditional probability of the ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE sequence for concurrent training was also calculated. The results from the analyses can be seen in Table 6.11.

Table 6.11

The conditional probabilities of sequences resulting in correct unprompted acquisition trial responses following ACQUISITION CORRECT RESPONSE and MASTERED CORRECT RESPONSE during interspersed training, and following ACQUISITION CORRECT RESPONSE during concurrent training. The frequency of each transition is in brackets. The highest probability for each dyad is in bold text.

	Interspersed		Concurrent
Dyad	ACQUISITION CORRECT RESPONSE FOLLOWING ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE FOLLOWING MASTERED CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE FOLLOWING AQCUISITION CORRECT RESPONSE
1	**.78 (7/9)**	.64 (25/39)	.46 (17/37)
3	.58 (40/69)	**.73 (73/100)**	.72 (68/94)
4	.67 (20/30)	**.76 (42/55)**	.40 (4/10)
6	.82 (28/34)	.82 (33/40)	**.93 (39/42)**
7	.73 (11/15)	.71 (22/31)	**.84 (47/56)**
8	.55 (11/19)	.60 (32/53)	**.72 (31/43)**

Table 6.11 displays the probabilities and the frequencies (in brackets) of the transitions for each dyad.

For interspersed training, the sequence MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE was higher for Dyad 3 ($p = .73$), Dyad 4 ($p = .76$), and Dyad 8 ($p = .60$). The probability of ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE during interspersed training was higher for Dyad 1 ($p = .78$), and Dyad 7 ($p = .73$). For Dyad 6, the probability of unprompted correct acquisition trial responses during interspersed training was .82 regardless of whether it was preceded by a mastered or acquisition S^D.

For Dyads 6, 7, and 8 the probability of the ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE during concurrent training was higher than the probabilities of both interspersed training procedures. The probability of the ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE sequence during concurrent training was .93 for Dyad 6, .84 for Dyad 7, and .72 for Dyad 8.

For Dyad 3 the probability of the ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE during concurrent training ($p = .72$) was approximately the same as the MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE during interspersed training. The results from Table 6.11 showed that only for Dyad 4 did it appear that presenting a mastered S^D made it more likely that the learner would respond correctly on a subsequent acquisition S^D during interspersed training.

Task Sequencing: Extended Markov Chain Analyses

Exploratory analyses were conducted to investigate the most frequently occurring three-trial sequences resulting in unprompted correct and incorrect acquisition trial responses during interspersed training. The baseline phase data for Dyads 1, 3, 4, 6, 7, and 8 were used in the analyses. The Markov chains were extended to include three trials, that is, the analyses were used to investigate the degree with which a trial outcome was associated with the two trials that preceded it. The MTM used to construct the Markov probability trees (Figures 6.1 to 6.6) can be seen in Tables 6.12 to 6.17. The probabilities were calculated by dividing the number of times the transition occurred by the number of times the first event in the sequence occurred (as indicated in the SUM column). The frequency of each transition is in brackets. The events in the first column were the first trial in a two-trial sequence, and the events in columns two to six were the second trial in the sequence.

Chi-square tests were conducted to investigate whether the three-event sequences occurred significantly more or less than expected by chance. That is, was there a statistically significant difference between the observed and expected values. The expected value of a sequence represents the number of times a sequence would be observed if events were occurring independently (i.e., randomly). The expected value of a sequence is calculated by calculating the unconditional probability of the sequence. The unconditional probability of each event in the sequence was calculated by dividing the number of times each event was observed by the total number of events. The unconditional probability of the sequence was calculated by multiplying the unconditional probabilities of each event in the sequence. Once the unconditional probability of the sequence was calculated, the expected value was calculated by multiplying the unconditional probability of the sequence with the total number of three-event sequences.

Table 6.12

Dyad 1 Markov transition matrix for interspersed training. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial					
	ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	MASTERED CORRECT RESPONSE	MASTERED INCORRECT RESPONSE	SUM
ACQUISITION CORRECT RESPONSE	.16 (7)	.11 (5)	.05 (2)	.59 (26)	.09 (4)	1.00 (44)
ACQUISITION CORRECT RESPONSE (P)	.13 (5)	.05 (2)	.03 (1)	.69 (27)	.10 (4)	1.00 (39)
ACQUISITION INCORRECT RESPONSE	.24 (5)	.48 (10)	.14 (3)	.14 (3)	.00 (0)	1.00 (21)
MASTERED CORRECT RESPONSE	.14 (25)	.12 (21)	.08 (14)	.55 (97)	.10 (18)	1.00 (175)
MASTERED INCORRECT RESPONSE	.07 (2)	.03 (1)	.03 (1)	.76 (22)	.10 (3)	1.00 (29)
SUM						308

Table 6.13

Dyad 3 Markov transition matrix for interspersed training. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial					
	ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	MASTERED CORRECT RESPONSE	MASTERED INCORRECT RESPONSE	SUM
ACQUISITION CORRECT RESPONSE	.26 (40)	.01 (2)	.19 (29)	.42 (65)	.12 (19)	1.00 (155)
ACQUISITION CORRECT RESPONSE (P)	.11 (6)	.07 (4)	.27 (15)	.38 (21)	.16 (9)	1.00 (55)
ACQUISITION INCORRECT RESPONSE	.35 (32)	.35 (32)	.20 (18)	.05 (5)	.04 (4)	1.00 (91)
MASTERED CORRECT RESPONSE	.33 (73)	.07 (15)	.12 (27)	.34 (75)	.13 (29)	1.00 (219)
MASTERED INCORRECT RESPONSE	.14 (3)	.04 (3)	.03 (2)	.69 (53)	.21 (16)	1.00 (77)
SUM						597

Table 6.14

Dyad 4 Markov transition matrix for interspersed training. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial					
	ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	MASTERED CORRECT RESPONSE	MASTERED INCORRECT RESPONSE	SUM
ACQUISITION CORRECT RESPONSE	.25 (20)	.01 (1)	.13 (10)	.57 (45)	.04 (3)	1.00 (79)
ACQUISITION CORRECT RESPONSE (P)	.24 (12)	.12 (6)	.18 (9)	.44 (22)	.02 (1)	1.00 (50)
ACQUISITION INCORRECT RESPONSE	.12 (5)	.65 (28)	.23 (10)	.00 (0)	.00 (0)	1.00 (43)
MASTERED CORRECT RESPONSE	.29 (42)	.10 (14)	.09 (13)	.40 (57)	.13 (18)	1.00 (144)
MASTERED INCORRECT RESPONSE	.00 (0)	.04 (1)	.04 (1)	.80 (20)	.12 (3)	1.00 (25)
SUM						341

Table 6.15

Dyad 6 Markov transition matrix for interspersed training. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial					
	ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	MASTERED CORRECT RESPONSE	MASTERED INCORRECT RESPONSE	SUM
ACQUISITION CORRECT RESPONSE	.36 (28)	.03 (2)	.08 (6)	.47 (36)	.06 (5)	1.00 (77)
ACQUISITION CORRECT RESPONSE (P)	.18 (7)	.26 (10)	.08 (5)	.44 (17)	.05 (2)	1.00 (39)
ACQUISITION INCORRECT RESPONSE	.43 (9)	.14 (3)	.24 (5)	.19 (4)	.00 (0)	1.00 (21)
MASTERED CORRECT RESPONSE	.34 (33)	.26 (25)	.07 (7)	.30 (29)	.04 (4)	1.00 (98)
MASTERED INCORRECT RESPONSE	.00 (0)	.00 (0)	.00 (0)	.91 (10)	.09 (1)	1.00 (11)
SUM						246

Table 6.16

Dyad 7 Markov transition matrix for interspersed training. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial					
	ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	MASTERED CORRECT RESPONSE	MASTERED INCORRECT RESPONSE	SUM
ACQUISITION CORRECT RESPONSE	.24 (11)	.09 (4)	.09 (4)	.52 (24)	.07 (3)	1.00 (46)
ACQUISITION CORRECT RESPONSE (P)	.19 (7)	.35 (13)	.05 (2)	.38 (14)	.03 (1)	1.00 (37)
ACQUISITION INCORRECT RESPONSE	.31 (5)	.56 (9)	.06 (1)	.06 (1)	.00 (0)	1.00 (16)
MASTERED CORRECT RESPONSE	.22 (22)	.11 (11)	.09 (9)	.46 (46)	.11 (11)	1.00 (99)
MASTERED INCORRECT RESPONSE	.06 (1)	.06 (1)	.00 (0)	.76 (13)	.12 (2)	1.00 (17)
SUM						215

Table 6.17

Dyad 8 Markov transition matrix for interspersed training. The table contains the probability and frequency (in brackets) of each transition. The events in the first column were the first trial in the sequence, and the events in columns two to six were the second trial in the sequence.

First Trial	Second Trial					
	ACQUISITION CORRECT RESPONSE	ACQUISITION CORRECT RESPONSE (P)	ACQUISITION INCORRECT RESPONSE	MASTERED CORRECT RESPONSE	MASTERED INCORRECT RESPONSE	SUM
ACQUISITION CORRECT RESPONSE	.15 (11)	.01 (1)	.11 (8)	.58 (42)	.14 (10)	1.00 (72)
ACQUISITION CORRECT RESPONSE (P)	.17 (5)	.07 (2)	.17 (5)	.48 (14)	.10 (3)	1.00 (29)
ACQUISITION INCORRECT RESPONSE	.36 (18)	.22 (11)	.32 (16)	.10 (5)	.00 (0)	1.00 (50)
MASTERED CORRECT RESPONSE	.27 (32)	.12 (14)	.18 (21)	.31 (36)	.13 (15)	1.00 (118)
MASTERED INCORRECT RESPONSE	.18 (6)	.00 (0)	.00 (0)	.67 (22)	.15 (5)	1.00 (33)
SUM						302

The chi-square statistic was calculated taking into account both the number of times the sequence occurred and the number of time it did not. Therefore, all sequences were categorized as either being the sequence of interest or not being the sequence of interest. The chi-square statistic was calculated using the formula:

$$\chi^2 = \sum \frac{(Observed-Expected)^2}{Expected}$$

The chi-square critical value at the .05 level of significance is 3.84 with one degree of freedom. The results from the analyses can be seen in Figure 6.1 to 6.6. Figures 6.1 to 6.6 contain the frequency of the sequence (*n*) and the probability of the sequence (p). The probabilities were calculated by multiplying the relevant transitional probabilities from the MTM. Figures 6.1 to 6.6 contain the most frequently occurring sequences which resulted in correct and incorrect acquisition trial responses.

The Dyad 1 results from the extended Markov chain analyses can be seen in Figure 6.1, and the corresponding MTM can be seen in Table 6.12. The sequence MASTERED CORRRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE was observed 15 times with a probability of .08. The expected value of the sequence was 14.34. The difference was not statistically significant (χ^2 (1) = 0.02, $p > .05$). The sequence that occurred with the second highest frequency was ACQUISTION CORRECT RESPONSE (P) – MASTERED CORRECT – ACQUISITION CORRECT RESPONSE which was observed 5 times with a probability of .10. The difference was not significant (χ^2 (1) = 0.98, p > .05).

In terms of sequences resulting in incorrect acquisition trial responses, the sequence that occurred most often was MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE. The sequence was

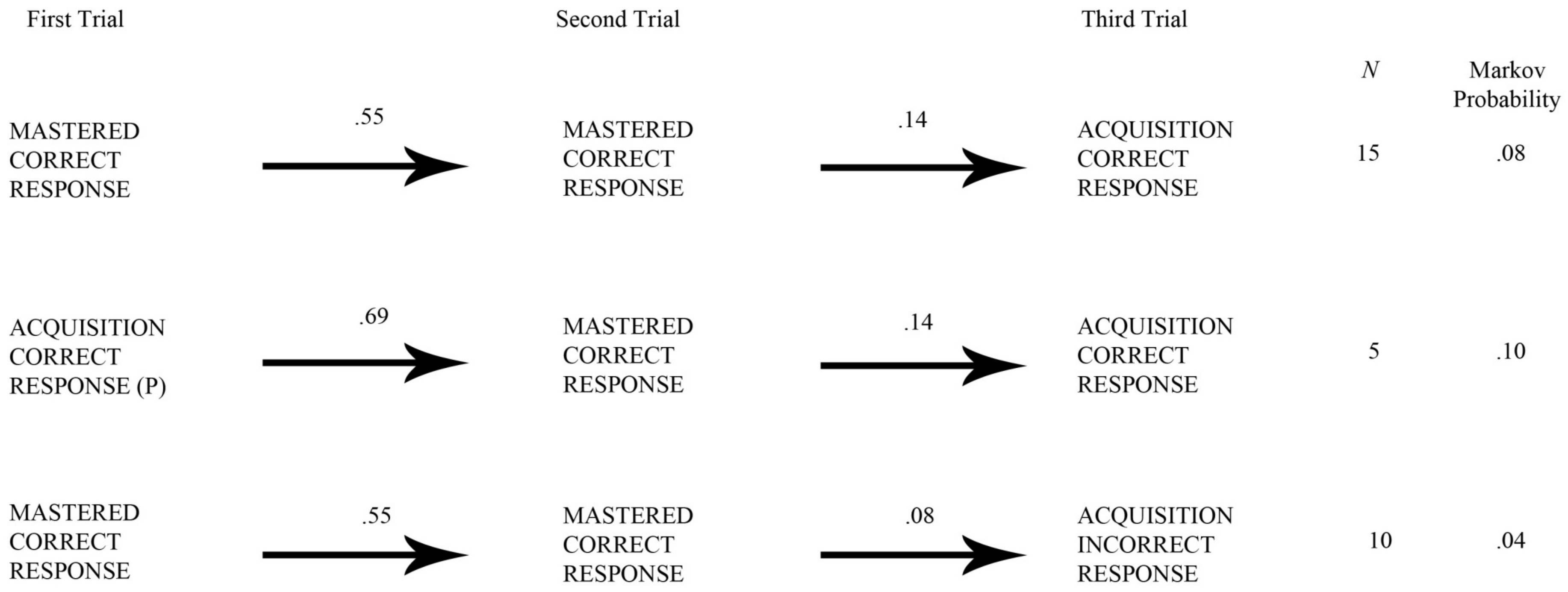

Figure 6.1: Extended Markov probability chain analyses for the most frequently observed three-trial sequences resulting in correct and incorrect acquisition trial responses for Dyad 1. The graph contains the frequency of the sequence and the Markov probability

observed 10 times with a probability of .04. The expected value of the sequence was 6.69, and the difference between the observed and expected frequencies was not significant (χ^2 (1) = 1.59, $p > .05$). No other three-event sequences resulting in incorrect acquisition trial responses occurred more than twice.

The Dyad 3 results from the extended Markov chain analyses can be seen in Figure 6.2 and the corresponding MTM in Table 6.13. Figure 6.2 shows that the most frequently observed three-event sequence resulting in unprompted correct acquisition trial responses was ACQUISITION CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. The sequence was observed 36 times with a probability of .14. The expected value of the sequence was 14.66. The difference between the observed and expected values was significant (χ^2 (1) = 31.45, $p < .05$).

The second most observed sequence was MASTERED CORRECT – MASTERED CORRECT – ACQUISITION CORRECT, the sequence was observed 21 times and the probability of the sequence was .11. The expected value was 20.72. This difference was not significant (χ^2 (1) ≈ 0, $p > .05$). The sequence MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE was observed 14 times and the probability was .09. The expected value of the sequence was 14.79 and the difference was not significant (χ^2 (1) = 0.04, $p > .05$).

For sequences resulting in incorrect acquisition trial responses the two sequences observed most frequently were MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE. The sequence was observed 15 times and the probability of the sequence was .06. The expected value of the sequence was 8.61 and this difference was significant indicating the sequence was occurring significantly more than expected (χ^2 (1) = 4.73, $p < .05$).

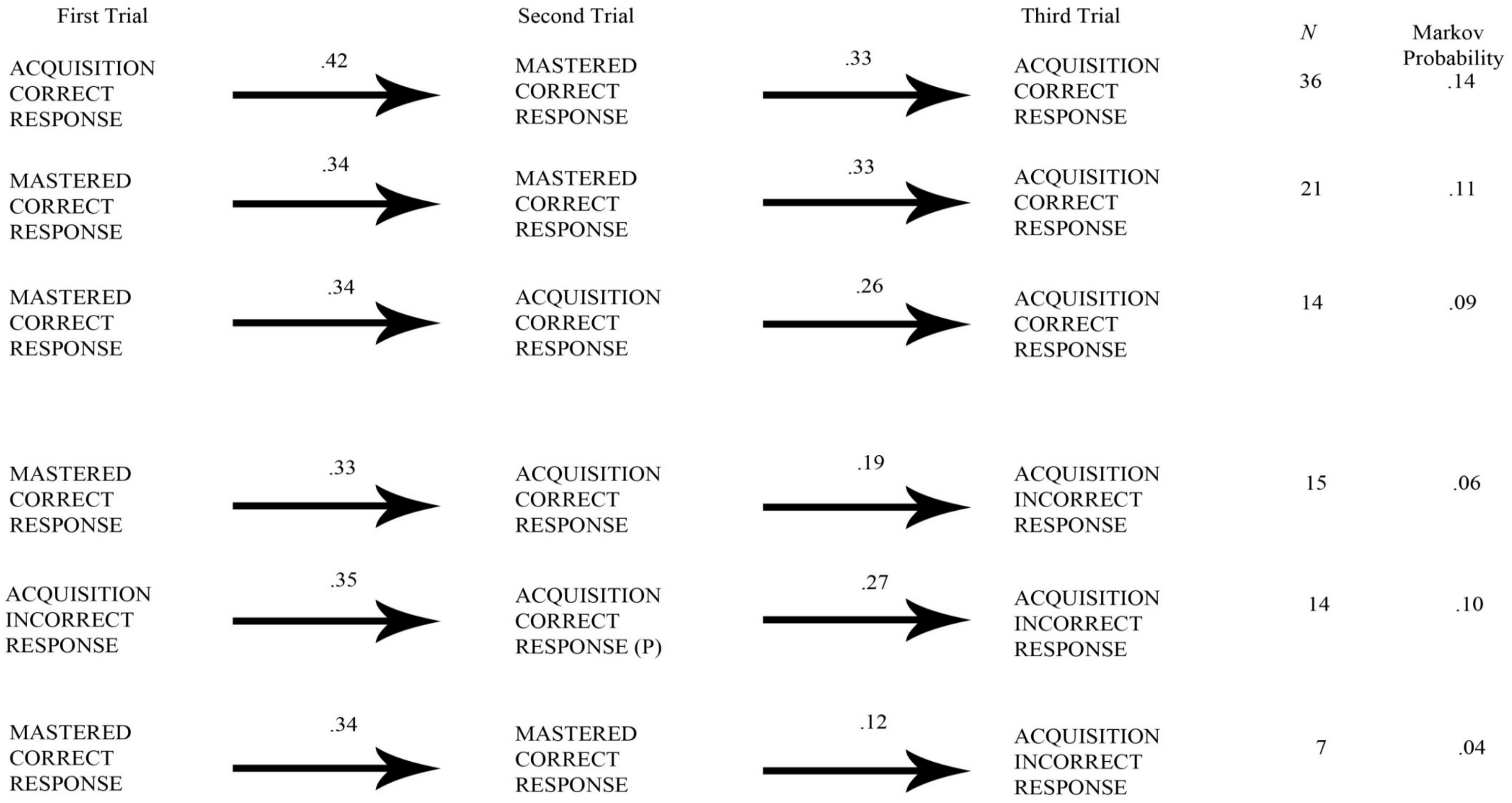

Figure 6.2: Extended Markov probability chain analyses for the most frequently observed three-trial sequences resulting in correct and incorrect acquisition trial response for Dyad 3. The graph contains the frequency of the sequence and the Markov probability

The second most observed sequence was ACQUISITION INCORRECT RESPONSE – ACQUISITION CORRECT RESPONSE (P) – ACQUISITION INCORRECT RESPONSE. The sequence occurred 14 times with a probability of .10. The expected value of the sequence was 1.30 and the difference was significant (χ^2 (1) = 124.46, $p < .05$). The sequence was occurring significantly more than what was expected. The sequence MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE was observed 7 times and the probability was .04. The expected frequency was 12.16. This difference was not significant (χ^2 (1) = 2.28, $p > .05$).

The Dyad 4 results from the extended Markov probability analyses can be seen in Figure 6.3 and the corresponding MTM in Table 6.14. The three-event sequence which most often resulted in unprompted correct acquisition trial responses was MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. The sequence was observed 18 times with a probability of .12. The observed frequency did not differ significantly from the expected frequency of 14.12 (χ^2 (1) = 1.03, $p > .05$). The sequence in which the acquisition and mastered S^D alternated (ACQUISITION CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE) was observed 16 times with a probability of .17, and there was a statistically significant difference between the observed and expected values (χ^2 (1) = 8.96, $p < .05$). The expected value of the sequence was 7.69. The sequence ACQUISITION INCORRECT RESPONSE – ACQUISITION CORRECT RESPONSE (P) – ACQUISITION INCORRECT occurred with a probability of .16 and was observed 11 times. The expected value of the sequence was 1.44 and occurred significantly more than expected (χ^2 (1) = 62.81, $p < .05$).

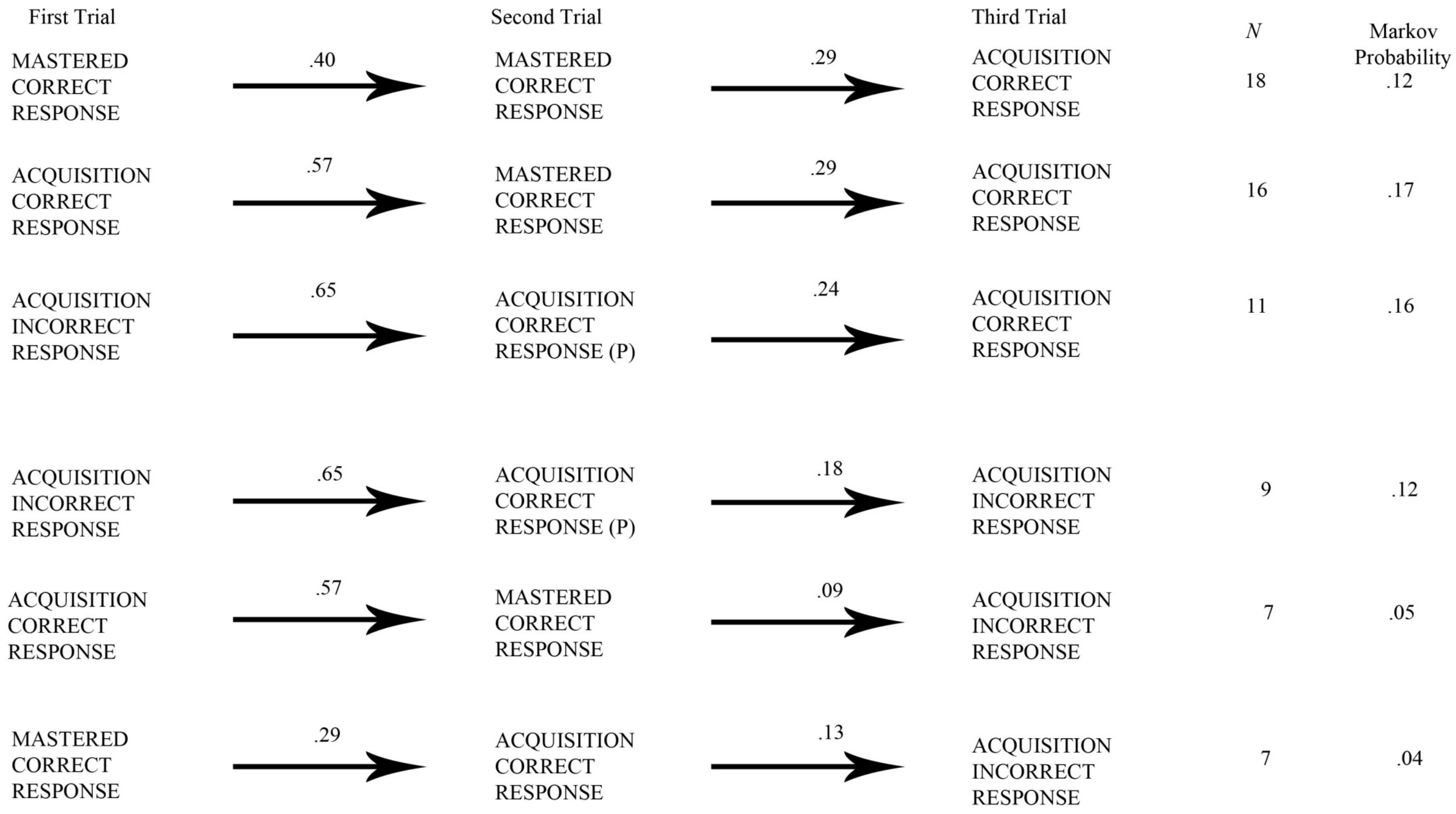

Figure 6.3: Extended Markov probability chain analyses for the most frequently observed three-trial sequences resulting in correct and incorrect acquisition trial responses for Dyad 4. The graph contains the frequency of the sequence and the Markov probability

When investigating the three-event sequences resulting in incorrect acquisition trial responses, Figure 6.3 shows that the sequence observed most often was ACQUISITION INCORRECT RESPONSE – ACQUISITION CORRECT RESPONSE (P) – ACQUISITION INCORRECT RESPONSE. The sequence occurred 9 times with a probability of .12. The expected value of the sequence was .80, and the difference between the observed and expected frequencies was significant (χ^2 (1) = 85.17, $p < .05$). The sequences ACQUISITION CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE and MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE both occurred 7 times with respective probabilities of .05 and .04. These two sequences were not significant.

The Dyad 6 results from the extended analyses can be seen in Figure 6.4 and the corresponding MTM can be seen in Table 6.15. For three-event sequences resulting in unprompted correct acquisition trial responses, the sequence with the highest observed frequency was MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. The sequence was observed 17 times with a probability of .10. The expected frequency was 12.08. The difference was not significant (χ^2 (1) = 1.96, $p > .05$).

The second most frequently observed sequence was ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. The sequence was observed 11 times with a probability of .13 and the expected value was 7.02. The difference was not significant (χ^2 (1) = 1.96, $p > .05$). The third most observed sequence was MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE which was observed 10 times and the probability was .12. The expected value of the sequence was 9.21 and the expected value

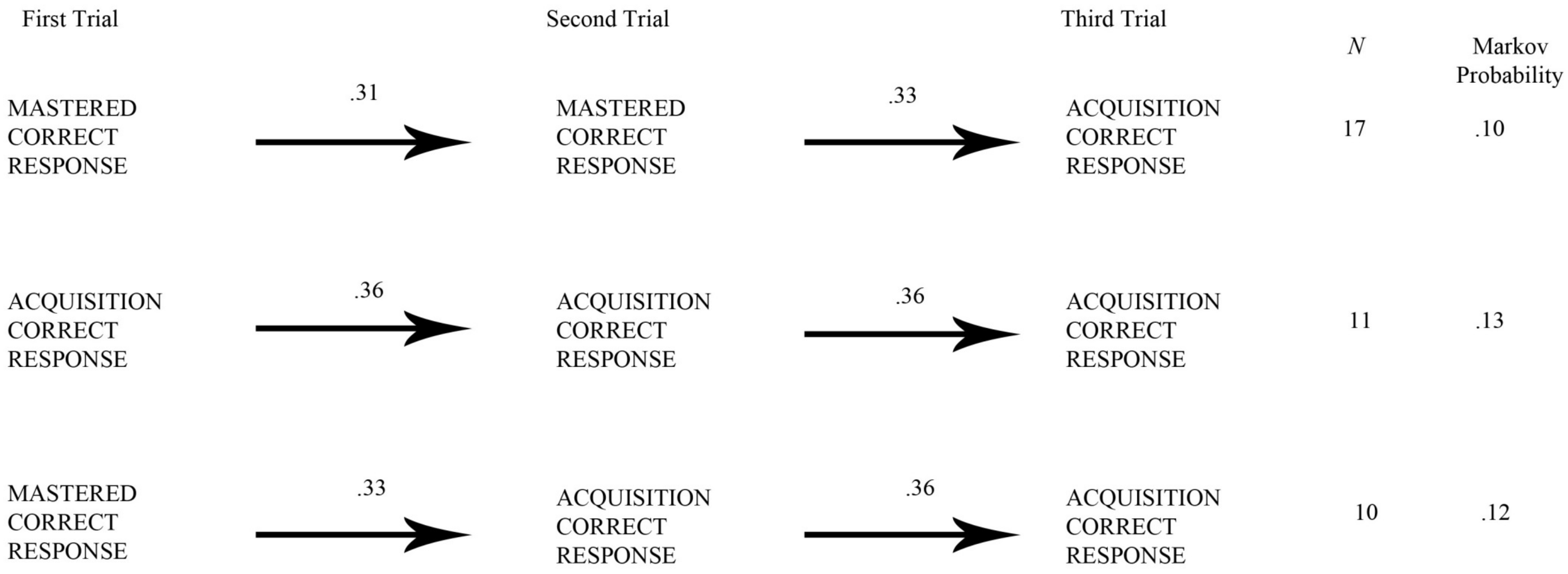

Figure 6.4: Extended Markov probability chain analyses for the most frequently observed three-trial sequences resulting in correct and incorrect acquisition trial responses for Dyad 6. The graph contains the frequency of the sequence and the Markov probability.

did not significantly differ from its expected value (χ^2 (1) = 0.05, p > .05). No sequence resulting in incorrect acquisition trial responses occurred with a frequency of greater than 4.

The Dyad 7 results from the extended Markov analyses can be seen in Figure 6.5 and the transition matrix in Table 6.16. First to be discussed are the sequences ending in unprompted correct acquisition trial responses. The sequence observed most often was MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. The sequence was observed 12 times with a probability of .10. The expected value was 9.57, and this difference was not significant (χ^2 (1) = 0.57, p > .05). No other sequences occurred with a frequency of greater than 4.

For sequences resulting in incorrect acquisition trial response the sequence that occurred most often was MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE. The sequence was observed 6 times, and the probability was .04. The expected value of the sequence was 3.33, and the difference was not significant (χ^2 (1) = 2.07, p > .05).

The Dyad 8 results from the extended analyses can be seen in Figure 6.6 and the MTM in Table 6.17. The sequence that occurred most frequently was the sequence ACQUISITION CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE which was observed 14 times with a probability of .16. The expected frequency of the sequence was 6.67 and the sequence did occur significantly more than expected (χ^2 (1) = 6.79, p < .05).

The second most observed sequence was MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. The sequence occurred 11 times, and the probability of the sequence was .08. The expected value of the sequence was 11.03, and the difference was not significant (χ^2 (1) ≈ 0, p > .05). The third most frequently observed sequence was MASTERED CORRECT RESPONSE –

ACQUISITION INCORRECT RESPONSE – ACQUISITION CORRECT RESPONSE which occurred 9 times with a probability of .06. The expected value was 4, and did not occur significantly more than expected (χ^2 (1) = 4.71, p < .05).

In terms of sequences resulting in incorrect learner responses the sequence MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE was observed 10 times with a probability of .05, and the expected value of the sequence was 7.71. This difference was not significant (χ^2 (1) = 0.68, p > .05). The sequence MASTERED CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE (n = 8, p = .06, expected value = 3.22) did occur significantly more than expected (χ^2 (1) = 7.01, p < .05).

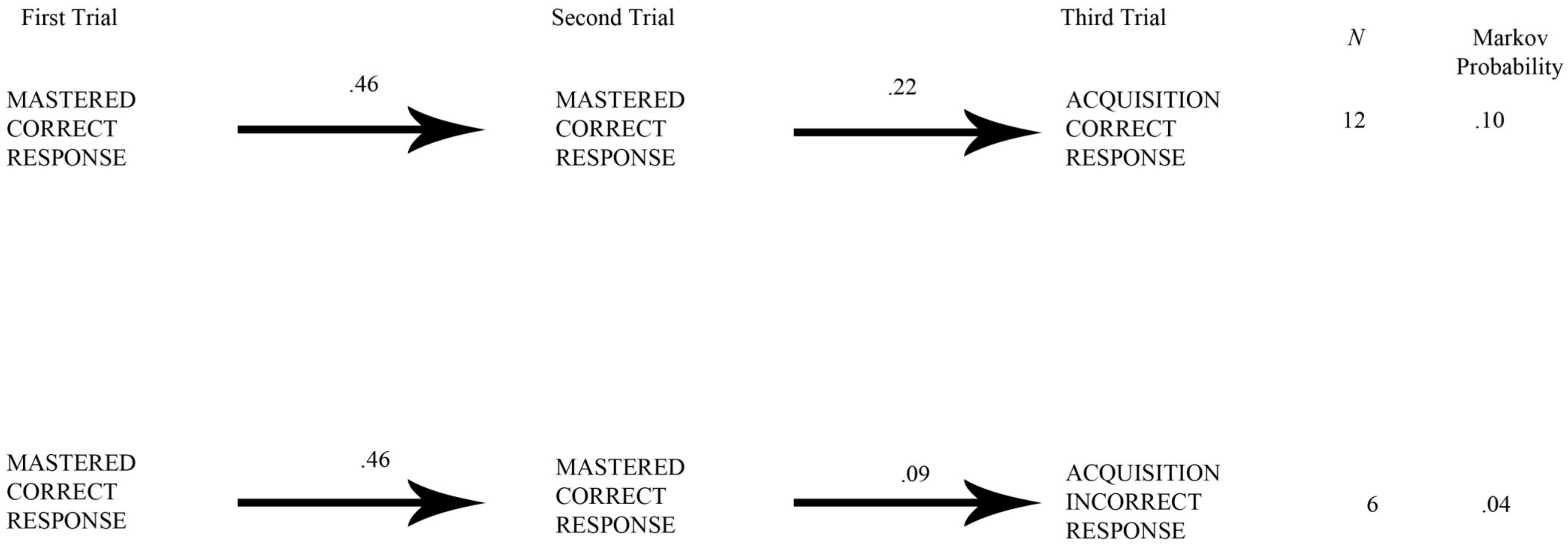

Figure 6.5: Extended Markov probability chain analyses for the most frequently observed three-trial sequences resulting in correct acquisition trial responses for Dyad 7. The graph contains the frequency of the sequence and the Markov probability.

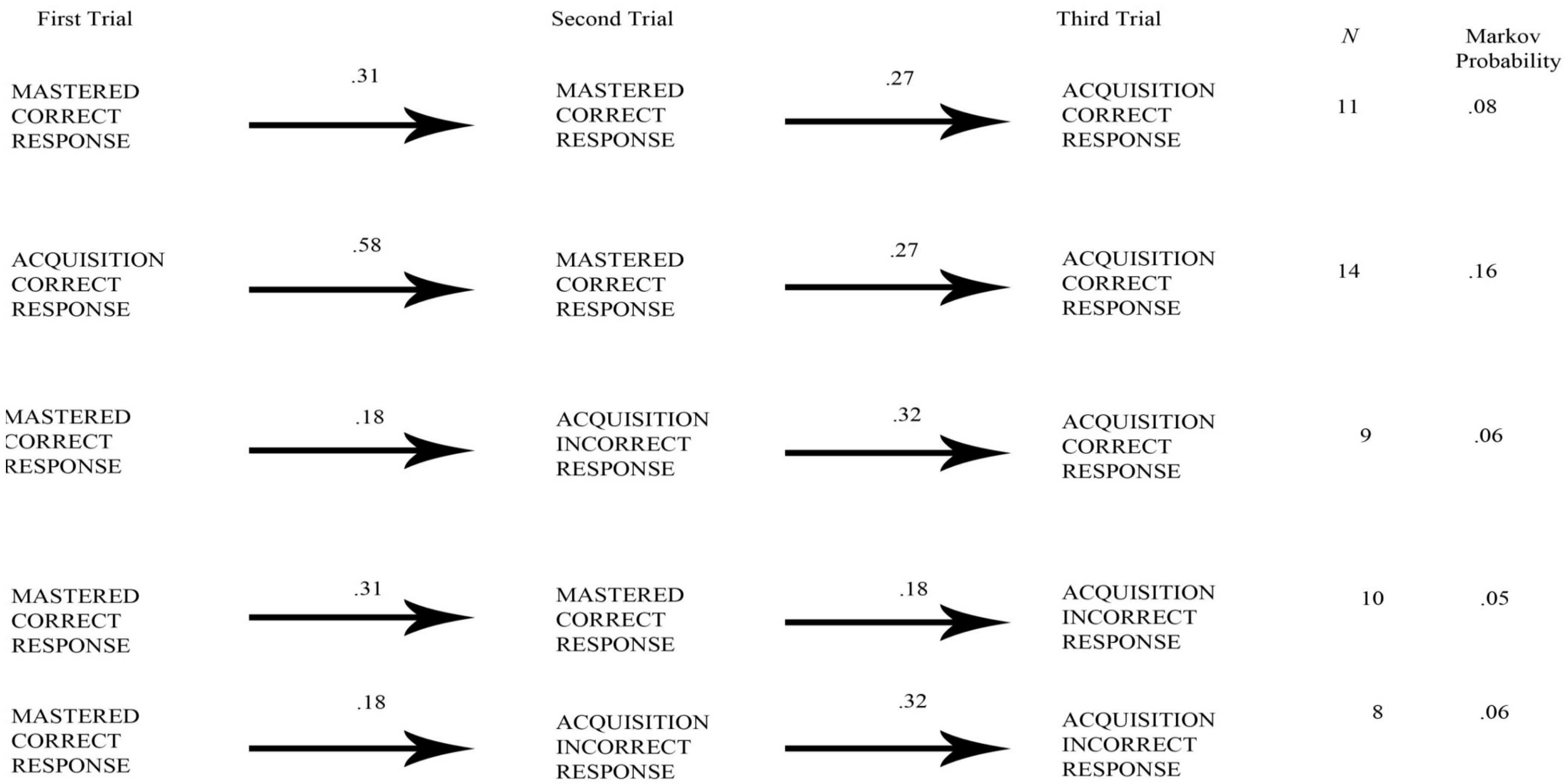

Figure 6.6: Extended Markov probability chain analyses for the most frequently observed three-trial sequences resulting in correct and incorrect acquisition trial responses for Dyad 8. The graph contains the frequency of the sequence and the Markov probability.

Discussion

General Findings

The aim of the analyses involving within-trial error-correction procedure errors was to extend the findings from Study 1 and investigate how within-trial error-correction treatment integrity errors affected learner performance. For three dyads, the probability of correct acquisition trial responses was lower following the incorrect within-trial application of the error-correction procedure, compared to when the within-trial error-correction procedure was administered in the prescribed manner. For one dyad, no differences were found. The results showed that using one-step MTM was an effective method of evaluating learner performance following the within-trial error-correction treatment integrity errors that were reported to consultants, and for which therapists were re-trained.

The second aim of this study was to introduce a method of evaluating the treatment integrity of error-correction procedures that have a between-trials component. The analyses consisted of constructing probability matrices, similar to the one-step MTM used in Study 1, so that between-trial error-correction treatment integrity errors could be identified. The matrices were similar to MTM in that they were exhaustive; they displayed all possible transitions that could be made following a trial of interest. Setting up the matrices in this way meant that the type of treatment integrity errors, along with their frequencies and probabilities could easily be identified. Treatment integrity errors were identified which can be reported to consultants so that therapists can be re-trained.

The aim of the task sequencing analyses was to compare the relative effectiveness of concurrent and interspersed task sequencing procedures. The main research question was whether correct unprompted acquisition trial responses were more or less likely immediately following one or more mastered trial than they were following one or more acquisition trials.

The results from the task sequencing analyses showed that interspersed training did not result in a higher probability of unprompted correct acquisition trial responses across all of the dyads. For some dyads, concurrent training appeared to be the more effective task sequencing procedure. Dyad 4 was the only dyad for which clear benefits of interspersed training were detected.

Within-Trial Error-Correction Treatment Integrity Errors

The findings from the within-trial error-correction treatment integrity error analyses showed that correct acquisition trial responses were less likely following within-trial error-correction errors for three dyads. This supports the findings from other studies that showed that learner performance on acquisition trials was adversely affected by within-trial treatment integrity errors (DiGennaro Reed, Reed, Baez, & Maguire, 2011; Downs, Downs, & Rau, 2008; Koegel, Russo, & Rincover, 1977; Sarokoff & Sturmey, 2008; Vladescu, Carroll, Paden, & Kodak, 2012). The results further justified the therapist re-training that was initiated as part of Study 1. The results from the current analyses are important because the overall acquisition trial treatment integrity scores across all within-trial components for these three dyads ranged between 76% and 87% which is relatively high. Although the treatment integrity scores were high, learner performance on acquisition trials appeared to be adversely affected by these errors. The results emphasise the point that high overall treatment integrity scores may mask infrequently occurring treatment integrity errors that may adversely affect learner performance.

It is not immediately apparent why there were no differences detected for Dyad 3. The findings may be procedural in that for Dyad 3 125 INCORRECT RESPONSE FOLLOWED BY CORRECT TREATMENT were coded and 15 INCORRECT RESPOSE RESPONSE FOLLOWED BY INCORRECT TREATMENT were recorded. It is possible that if approximately 125 INCORRECT RESPONSE FOLLOWED BY INCORRECT TREATMENT were also recorded, that the probability of correct acquisition trial responses may decrease, although it is also plausible that they might increase.

Another reason may be that the frequencies and probabilities of the within-trial error-correction errors were not high enough to affect learner performance. The learner self-correction error occurred 16 times with a probability of .06, and the error sequence of not allowing ITI of prescribed duration following the error-consequence was .06, and occurred 50 times. These frequencies and probabilities were relatively small and may not have reached the threshold where they were adversely affecting learner performance. The results from the Davison and McCarthy (1980) study showed that substantial decreases in the number of correct discriminations were not detected until errors were reinforced with a probability .40. Therefore, it is possible that the error sequences identified for Dyad 3 did not reach the threshold where they were adversely affecting learner performance.

The other possibility is that the within-trial error-correction procedure was not effective in increasing the probability of correct responding on the next acquisition trial. The within-trial error-consequence for Dyad 3 was to say no and to remove all stimuli from table. Informal discussions with the consultant revealed that they suspected that the learner found the verbal utterance of saying no reinforcing, and that the within-trial error-correction was going to change so that no verbal or social interaction with the learner was made following incorrect responses. The last possibility was that the types of within-trial error-correction procedure errors that were identified for Dyad 3 were not functionally related to learner performance.

The results from these analyses need to be interpreted with caution for a number of reasons. The results only show that incorrect acquisition trial responses were correlated with within-trial error-correction treatment integrity errors. The results do not show that the incorrect acquisition trial responses were caused by the within-trial error-correction errors. In order to establish a functional relationship between within-trial error-correction treatment integrity errors and learner responding, a controlled study must be conducted.

Given how the trials were coded, it cannot be stated with certainty which specific within-trial error-correction treatment integrity errors were correlated with incorrect acquisition trial responses. These errors were learner self-corrections, not providing the error-consequence following incorrect responses, providing prompts following incorrect responses, not allowing ITI of prescribed duration, and providing the correct consequence following incorrect responses. These errors did not always occur in isolation. For example, it was the case for some trials that the learner self-corrected, and the therapist did not allow an ITI of prescribed duration during the same trial. Therefore two errors occurred and this study was not able to determine which one of these within-trial error-correction errors was more detrimental for learner performance. Such analyses could be conducted, however there were not enough instances of these errors to produce meaningful results.

Despite the limitations of the analyses, the results indicated that impaired learner performance on acquisition trials were correlated with the incorrect application of the within-trial error-correction procedures for three dyads. Given the several within-trial error-correction treatment integrity errors that were identified controlled studies can be conducted to assess their individual effects on learner performance.

Between-Trial Error-Correction Treatment Integrity

The biggest finding from the analyses was that high-levels of the within-trial treatment integrity did not equate to high degrees of treatment integrity of the between-trials component of the error-correction procedures. Very few of the prescribed sequences occurred with a probability of at least .90. The results showed that it should not be assumed that all aspects of the DTT procedure are occurring with a high degree of treatment integrity because the within-trial treatment integrity is high. Each aspect of the procedure should be examined and evaluated on a regular basis in order for the entire procedure to be conducted with a high degree of treatment integrity.

The results showed several areas where therapists could be re-trained. All therapists need to be re-trained to administer a prompt along with the re-presentation of the S^D following incorrect acquisition trial responses. This is an omission error, and research has shown that learner performance on acquisition trials is sensitive to omission errors (e.g., St. Peter Pipkin, Vollmer, & Sloman, 2010; Wilder, Atwell, & Wine, 2006). The highest probability of the prescribed INCORRECT RESPONSE – SDA (SAME) P sequence was .56. So across all dyads, at best, the prescribed procedure is administered following slightly more than half of all incorrect acquisition trial responses.

For Dyads 7 and 8, no error-correction procedure was in place for certain programmes. The prescribed sequence was that a different S^D should be administered on the next trial. The probability of the sequence for Dyad 7 was .88, and for Dyad 8 it was .91. The results showed that this procedure was administered with a higher degree of treatment integrity than the error-correction procedure requiring prompted re-presentations of the S^D. Providing the vocal-feedback only without the presentation of an error-correction trial requires less input from the therapist than the procedures requiring multiple prompted and unprompted re-presentations of the S^D. It is interesting that the error-correction procedure requiring less therapist input was implemented with a higher degree of treatment integrity than the procedures that required more. It could be that that more complex error-correction procedures are more susceptible to treatment integrity errors than simpler error-correction procedures (Cooper, Heron, & Heward, 2007; Peterson, Homer, & Wonderlich, 1982). This is an empirical question which can be assessed in controlled studies.

There are a few possible reasons as to why the probabilities of the prescribed sequences were lower than the optimal 90% treatment integrity. One reason is related to the within-trial error-correction treatment integrity errors that were identified during Study 1. Some of these errors included presenting prompts immediately following incorrect learner responses. Within-trial error sequences such as INCORRECT – GESTURAL, INCORRECT – PHYSICAL and INCORRECT – VERBAL reached the .05 criterion for Dyads 1, 4, 7, and 8. In these cases the therapist provided the prompt immediately following the incorrect response instead of ending the trial with the error-consequence and re-presenting the S^D with a prompt. The within-trial treatment integrity error sequences indicate that therapists tended to alter the error-correction procedure and conduct the error-correction procedure on a within-trial basis instead of across trials. This could explain why on the trial following the incorrect response the therapist re-presented the S^D without the prompt because they prompted immediately following the incorrect response.

It is not clear as to why therapists presented different acquisition S^D following incorrect acquisition trial responses when they should have re-presented the S^D with a prompt. The within-trial treatment integrity errors identified in Study 1 cannot account for that type of between-trial treatment integrity error.

The effectiveness of the different error-correction procedures could not be compared for two reasons. One reason was related to the frequency with which therapy sessions were video-recorded. Video-recordings could not be made every time the learner received therapy, and learner performance on individual acquisition stimuli could not be monitored. The second reason relates to the compromised levels of treatment integrity of the error-correction procedures. The decreased levels of treatment integrity introduces confounds into analyses regarding procedure effectiveness. Any statements regarding procedure effectiveness would not be reliable.

The aim of the analyses was to introduce a procedure for evaluating the treatment integrity with which error-correction procedures consisting of a between-trials component were administered. The usage of a probability matrix similar to MTM was effective in identifying treatment integrity errors that can be targeted for re-training. The errors that were identified in this study can now be empirically tested in controlled studies to assess their effects on learner performance. Conducting such studies will increase the ecological validity of studies investigating DTT treatment integrity errors and their effects on learner performance (Carroll, Kodak, & Fisher, 2013).

Task Sequencing

The results from the task sequencing analyses showed that interspersed training did not result in superior learner performance on acquisition trials across for all dyads. The results do not appear to be consistent with that of Dunlap (1984) and Rowan and Pear (1985) which found that interspersed training was superior to concurrent training in terms of learner performance on acquisition trials. The results for Dyad 4 were the only results that showed clear benefits of interspersed training. Table 6.11 showed that for Dyad 4, the probability of correct unprompted acquisition responses was more likely following mastered trials. Both the interspersed training probabilities were higher than the concurrent training probability, although results should be interpreted with caution given the low frequency of the concurrent training sequence.

The conditional probability analyses (Table 6.11) revealed results that could not be obtained using the results from Table 6.10. The following two paragraphs will provide two such examples. Table 6.10 showed that the Dyad 3 learner made proportionately more unprompted correct acquisition trial responses during interspersed training than concurrent training. On that basis, it can be assumed that interspersed training was the more effective procedure in terms of correct unprompted acquisition trial responding. However, the results from the conditional probability analyses showed that the probability of successive correct unprompted acquisition trial responses during concurrent training was approximately equal to the MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE sequence during interspersed training. The results from the conditional probability showed that when given an independent opportunity to respond on acquisition trials, the learner is just as likely to respond correctly during concurrent training as they were following mastered trials during interspersed training.

The conditional probability analyses also revealed patterns obscured by the Table 6.10 results for Dyad 6. The results from Table 6.10 showed that the learner made 77 correct unprompted acquisition trial responses during interspersed training. The results from the conditional probability analyses showed that the learner was just as likely to respond correctly ($p = .82$) on acquisition trials following a mastered or acquisition S^D. Interspersing the mastered trials did not increase the probability of unprompted correct acquisition trial responses above the probability of presenting successive acquisition S^D.

The results from the extended Markov chain analyses should be interpreted with caution given that the observed frequencies of some sequences were relatively low. The accuracy of the analyses would be improved with larger samples. The findings from the extended Markov probability chains analyses were not consistent across all dyads, especially in terms of whether correct unprompted acquisition trial responses were observed more often than expected if preceded by two correct mastered trial responses. For Dyads 1, 3, and 8, the observed frequency of the MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE sequence was approximately the same as the expected frequency, indicating that the sequence occurred about as often as expected. However, for Dyads 4, 6, and 7, there were differences between the observed and the expected values although these differences were not statistically significant.

In terms of the sequence MASTERED CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION INCORRECT RESPONSE the observed frequencies of the sequence were higher than the expected counts for Dyads 1, 7, and 8, indicating that incorrect acquisition trial responses were observed more than was expected following two correct mastered trial responses. The results from the extended Markov probability chains further support the results from the conditional probability analyses that the intended benefits of interspersed training were not experienced by all learners

A majority of the most frequently observed sequences resulting in unprompted correct acquisition trial response did not occur significantly more than their expected frequencies. However, the sequence ACQUISITION CORRECT RESPONSE – MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE, which involved strictly alternating mastered and acquisition S^D, did occur significantly more than expected. This was the case for Dyads 3, 4, and 8. A possible clinical application of the results would be to present mastered and acquisition S^D in alternation instead of presenting multiple mastered S^D in succession.

One suggestion why interspersed training was not more effective than concurrent training for some dyads is that mastered stimuli were not actually mastered or that mastery was not maintained. Such a conclusion is warranted for Dyad 3 and Dyad 8. The mastery criterion for all dyads in the task sequencing analyses was that they had to respond correctly on at least 80% of trials for three consecutive sessions. Dyad 3 responded correctly on 74% of mastered trials administered during interspersed training, and Dyad 8 responded correctly on 78% of mastered trials. The percentage correct mastered trial responding was less than 80% for Dyads 3, and 8 which suggests that some of the stimuli which have been assumed to be mastered may actually not have been. Therefore, the effectiveness of the interspersed procedure may have been compromised for these dyads. Dyad 1 responded correctly on 86% of all mastered trials, Dyad 4 on 85%, Dyad 6 on 86%, and Dyad 7 on 85%.

Some of the findings from the task sequencing analyses can be accounted for by the findings from the Charlop, Kurtz, and Milstein (1992) study. Charlop et al. investigated whether the effectiveness of interspersed training would be affected when edible reinforcers were provided on different reinforcement schedules depending on whether the trial was mastered or acquisition. Five learners diagnosed with autism participated in a multiple baseline study in which conditions were counterbalanced across learners. Social praise and edible items were provided following every correct acquisition trial response across all conditions. The reinforcement schedules for edible items during mastered trials were manipulated across conditions, but social praise was provided following each correct response. During baseline edible items were provided on a variable ratio 3 (VR 3) schedule. During the no-reinforcers condition, no consequences were provided for correct mastered trial responses, and during the praise-only condition only social praise was provided following every correct mastered trial response. These conditions assessed whether removing all reinforcers (no-reinforcers) and removing the primary reinforcer (praise-only) for correct mastered trial responding affected acquisition trial responding.

The Charlop et al. (1992) results showed that skills were not acquired during baseline when edible items were available for correct responses during both trial types. Learners acquired skills once the no-reinforcement and praise-only conditions were initiated, and different types of reinforcers were available depending on what type of trial was administered. Charlop et al. stated that interspersed training had adverse effects on skill acquisition when edible items were available for correct responses during both trial types, as evident by the lack of skill acquisition during baseline.

For Dyads 1, 6, 7, and 8, learners received the same types of correct consequences following mastered and acquisition trials. Therapists provided social praise following correct responses on either trial type. Therapists did conduct preference assessments. Preferred stimuli, items and activities were sometimes provided at the end of a programme rather than following a particular trial type. Therefore, learners always received social praise following each correct response no matter the trial type, and only received preferred items or activities at the end of lessons. Although the Charlop et al. (1992) study and the current study differed in terms of the type of reinforcers (food versus social praise), the results may be applicable because the type of reinforcer was the same for mastered and acquisition trials.

An established operant condition principle is the differential outcomes effect (e.g., Brodigan & Peterson, 1976; Trapold, 1970). The differential outcomes effect is defined as, "the increase in the speed of acquisition or terminal accuracy that occurs in discrimination training when each of two or more discriminative stimuli is correlated with a particular outcome," (Goeters & Blakely, 1992, p. 389). Brodigan and Peterson (1976) found that rats made more correct discriminations and learned discriminations quicker when there were differential outcomes associated with each type of response. One possible recommendation that can be made from the findings is that therapists differentially reinforce correct mastered and acquisition trial responses during interspersed training.

Another factor that must be taken into consideration is that the results from the error-correction treatment integrity analyses showed that the integrity of all error-correction procedures was compromised for all dyads. Therefore some aspects of the DTT task sequencing procedures have been compromised in terms of the integrity with which they were administered. These compromises in treatment integrity could have affected the effectiveness of the procedures.

Another interesting finding from the analyses was the difference in the results from Dyads 1 and 4 (Learner 1 was part of both dyads). For Dyad 1, during interspersed training, the probability of unprompted correct acquisition trial responses was higher following unprompted correct acquisition trial responses. The results were the opposite for Dyad 4. There are several reasons why this may be the case. The within-trial treatment integrity score for acquisition trials was 86% for Dyad 4 and 34% for Dyad 1. The programmes had also changed between the times the therapy sessions were recorded for Dyad 1 and Dyad 4. The results may also suggest some measure of therapist stimulus control, and it could be case that some unmeasured aspect of therapist behaviour was influencing learner behaviour.

The results from the task sequencing analyses need to be interpreted with caution given the low frequencies with which some of sequences were observed. The results from the analyses are correlational and not functional. Some of the patterns and sequences identified in the task sequencing analyses suggest that at the very least consultants and therapists should consider some of the issues identified in this study as to why interspersed training does not appear to be the more effective procedure.

Chapter VII

General Discussion

Summary of Findings

The results from this thesis provided a demonstration of how clinically relevant information can be obtained when discrete-trial teaching data (DTT) data are analysed on a within-sessions basis. These within-sessions analyses provided a thorough evaluation of various parts of the DTT procedure for eight learner-therapist dyads. Collectively the results from this thesis showed that the quality of therapy was diminished by various treatment integrity errors occurring on both a within and between-trials basis. Some of these errors adversely affected learner performance. The results from the analyses also identified concerns regarding the procedures used to re-train therapists, and how the task sequencing procedures were prescribed. These results were obscured when data were analysed using whole-sessions methods.

The findings from this thesis are important from both a clinical and statistical perspective. The section *Clinical Recommendations* will discuss the broader clinical issues which the results from this thesis address. The sections labeled *Markov Transition Matrices and Extended Markov Chains*, and *Level of Analysis* will discuss the importance of the results from a statistical perspective.

Clinical Recommendations

Two important issues in relation to DTT that is administered as part of a learner's regular teaching programme in the natural environment are the quality of therapy, and the quality of the supervision that is provided for these programmes (Eikeseth, Klintwall, Jahr, & Karlsson, 2012; Stock, Mirenda, & Smith, 2013). Studies evaluating the effectiveness of DTT in the natural environment often find that there is a lesser degree of emphasis placed on constantly evaluating treatment integrity, and that supervision is inadequate and infrequent

(Eikeseth et al., 2012; Magiati, Charman, & Howlin, 2007; 2007; Stock et al., 2013). In this thesis treatment integrity was used to assess the quality of the therapy on various parts of the DTT procedure. The quality of supervision was determined by evaluating the frequency of supervision, the effectiveness of the therapist re-training procedures, and how the DTT programmes and procedures were prescribed.

Quality of Therapy

Consultants and therapists have a professional obligation to provide the highest quality service to learners receiving these services, and to the parent/caregivers who pay for them. Estimated costs of early intensive behavioural interventions consisting of DTT range between $20 000 and $60 000 per learner per year (Chasson, Harris, & Neely, 2007; Jacobson, Mulick, & Green, 1998; Sallows & Graupner, 2005). The results from this thesis indicated that learners were not receiving the highest quality therapy. This supports the findings of Eikeseth et al. (2012), Eikeseth, Hayward, Gale, Gitleson, and Eldevik (2009), and Magiati et al. (2007) who found similar results when evaluating the quality of DTT programmes administered in the natural environment.

The quality of the DTT programmes for all dyads was compromised due to the low levels of treatment integrity that were observed on both a within-trials and between-trials basis. Compromised levels of treatment integrity are undesirable for a number of reasons. From a clinical perspective they decrease the accuracy with which clinical decisions can be made regarding the effectiveness of the procedure, and concerning learner performance. Also, research has shown that learning can be adversely affected by treatment integrity errors (e.g. DiGennaro Reed, Reed, Baez, & Maguire, 2011; Downs, Downs, & Rau, 2008; Koegel et al., 1977; Sarokoff & Sturmey, 2008; Vladescu, Carroll, Paden, & Kodak, 2012). The finding from the within-trial error-correction treatment integrity analyses in Study 2 supported these findings.

Six of the dyads had overall within-trial treatment integrity scores of at least 80%. However, the overall treatment integrity scores were still less than the recommended at least 90%, which has been shown to be the optimal level of within-trial treatment integrity in terms of producing best learner outcomes (Koegel, Russo, & Rincover, 1977). High overall treatment integrity scores can be misleading for two reasons. First, as was found in Study 1, high overall treatment integrity scores can obscure treatment integrity errors occurring on an event-by-event basis. Second, the results from the Study 2 within-trial error-correction treatment integrity analyses showed that these errors did adversely affect learner performance on acquisition trials even when the overall treatment integrity scores were at least 80%. The results suggest that it should not be assumed that learners are receiving adequate levels of therapy based on high overall treatment integrity scores. All within-trial components should be examined separately on a regular basis to ensure that treatment integrity errors that adversely affect learner performance do not become a part of the procedure.

Traditionally DTT treatment integrity studies (e.g., Carroll, Kodak, & Fisher, 2012; Downs et al., 2008; Koegel et al., 1977; Vladescu et al., 2012) have focused almost exclusively on evaluating treatment integrity on a within-trial basis and almost no attention has been given to evaluating the treatment integrity of other parts of the procedure. The DTT literature demonstrates a lack of methods for identifying treatment integrity errors on other parts of the DTT procedure. This thesis introduced a sequential analysis method (i.e., conditional probability matrices) of evaluating the treatment integrity of prescribed error-correction procedures, and was successful at identifying treatment integrity errors for these procedures.

One of the main findings from Study 2 was that high levels of treatment integrity for the within-trial components did not indicate high levels of treatment integrity scores for the between-trials error-correction procedure. The results from Study 2 showed that the various error-correction procedures were not being administered according to consultant protocols, which suggests that therapists were not receiving adequate supervision on the error-correction part of the procedure. It is recommended that DTT programme consultants and supervisors attend to all parts of the DTT procedures as part of the supervision process. The errors identified in Study 2 can be targeted for therapist re-training. The low levels of treatment integrity for the error-correction procedures meant that the effectiveness of the error-correction procedures could not be evaluated given that the interval validity of the procedures had been compromised.

Quality of Supervision

The thesis identified a number of issues in relation to the supervision that was provided for the DTT programmes. These issues are the frequency of supervision, the procedures used to re-train therapists, and how certain parts of the DTT procedures have been prescribed.

Frequency of Supervision. The frequency of supervision is one factor that influences the quality of DTT programmes. Eikeseth et al. (2009) found significant positive correlations between the number of hours of supervision received per month and increases in learner IQ scores. The study consisted of 20 pre-school learners who had been diagnosed with autism. Bibby, Eikeseth, Martin, Mudford, & Reeves (2002) conducted a study evaluating learner outcomes of learners who received parent-managed intensive interventions which consisted of DTT. Bibby et al. reported that approximately 80% of the teaching programmes were supervised by consultants who were not adequately qualified to provide supervision for the type of teaching programmes that were being administered (i.e., UCLA-model programmes, Lovaas, 1987). In addition, supervision occurred approximately once every three months indicating that supervision was irregular and infrequent. The Bibby et al. results showed that the intervention had little or no effectiveness in terms of learner outcomes. Magiati et al. (2007) conducted a similar study and reported that supervision for some families occurred every 2 to 4 months, and for other families as infrequent as every 5 to 6 months. The results from the Magiati et al. study showed similar learner outcomes to those of Bibby et al.

The therapists for seven dyads received regular supervision every two weeks or once per month. The overall treatment integrity score was at least 80% for six of those seven dyads. Although the therapist for Dyad 1 also received regular supervision the overall treatment integrity score was 42%. As was discussed in Study 1 the low levels of treatment integrity for Dyad 1 were primarily caused by one particular treatment integrity error sequence (i.e., the ITI not lasting for its prescribed duration). It appears that the supervision and training procedures that the Dyad 1 therapist received were not sensitive (or thorough) enough to detect and correct this error. It was not until analyses of all the within-trial components were conducted (i.e. one-step Markov transition matrices, MTM) that the treatment integrity error was identified. The consultant for Dyad 1 expressed surprise at the frequency of this error.

For Dyad 2 the therapist received irregular and infrequent supervision. The therapist for Dyad 2 had not received any supervision during the 10 weeks it took to make the first 16 video-recordings. The within-trial treatment integrity for the Dyad 2 DTT programmes had deteriorated to such an extent that the therapist committed frequent errors across almost all parts of the procedure. In addition the low levels of within-trial treatment integrity meant that the Dyad 2 data had to be excluded from all of the analyses conducted in Study 2 because no valid or reliable results could be obtained due compromises in the internal validity of the procedure.

Therapist Re-Training Procedures. The results from Study 1 showed that the procedure used to re-train therapists produced short-term increases in within-trial treatment integrity, but that it was not sufficient to maintain it. This finding supports Koegel, Glahn, and Nieminen (1978) who stated that "certain training techniques may produce large initial changes in parent or teacher behavior, but do not seem to produce durable results" (p. 96). Only one dyad, Dyad 4, reached the ≥ 90% treatment integrity criterion following therapist re-training. Within-trial treatment integrity errors identified in the baseline phase were presented to the programme consultant and therapist during a meeting which took approximately 30 min. The prescribed procedures were discussed and following the meeting the researcher emailed the findings from the baseline phase to both the consultant and therapist. No further therapist re-training was reported. The follow-up phase results showed evidence of treatment drift and the probabilities of certain error sequences were at least as high during the follow-up phase as they were during the baseline phase. Based on the results, consultants for these DTT programmes can be advised to consider other more effective therapist re-training methods.

The DTT literature contains a few examples of effective therapist re-training methods that have resulted in maintenance of within-trial treatment integrity gains. Downs, Downs, and Rau (2008), and LeBlanc, Ricciardi, and Luiselli (2005) were therapist re-training studies that included a maintenance component. Following an initial baseline phase Downs et al. (2008) re-trained therapists using didactic feedback, live modelling, and corrective feedback. The baseline treatment integrity scores ranged between 63% and 80%. Following therapist re-training treatment integrity scores were above 90%. The maintenance data, recorded across a 10-week follow-up period following therapist re-training, showed that increases in treatment integrity were maintained at above 95%. Supervision sessions took place 2, 4, 6 and 10 weeks following the intervention.

LeBlanc et al. (2005) re-trained therapists using abbreviated performance feedback which consisted of reinforcing correct application of DTT components with social praise, and the therapist being corrected every time a treatment integrity error was made. The initial baseline phase data showed that treatment integrity scores ranged between 32% and 45%. Following re-training treatment integrity increased to above 90%. These increases in treatment integrity were maintained for up to 11 weeks following therapist re-training. Another effective therapist re-training procedure is behavioural skills training which consists of written instructions, verbal instructions, rehearsal of procedures, and modelling the correct application of the DTT components (Sarokoff & Sturmey, 2004). These are some of the possible therapist re-training methods that we can advise consultants to use instead of the procedures used in the current study which did not result in maintenance of treatment integrity increases for certain dyads based on the finding of treatment drift in the follow-up phase.

Thomson, Martin, Arnal, Fazio, and Yu (2009) conducted a review of 17 research articles in which therapists were re-trained on how to administer the within-trial components of DTT correctly following data collection during an initial baseline phase. The Thomson et al. review found that all of the studies in which increases in treatment integrity were reported consisted of training packages involving multiple re-training methods. These methods included a combination of antecedent methods (e.g., lectures, video-modeling, live modelling, and written instructions), and consequence methods (e.g., corrective feedback). None of the studies included in the review used verbal-feedback only. The methods used by the consultants to re-train therapists were not consistent with what has been reported in the literature.

How the Various Parts of The DTT Procedure Have Been Prescribed. The thesis also identified problems with how some aspects of the DTT programmes, such as the error-correction procedures and task sequencing procedures, were prescribed. Consultants and therapists need to use caution in how these procedures are prescribed and take steps to ensure that all parts of the DTT procedure are programmed in such a way that the learner can benefit from the applications of these procedures.

For some DTT programmes the prescribed error-correction procedure was verbal-feedback only. For these DTT programmes only two possible correct answers (e.g., he or she) were being trained, and learners were presented with two picture cards; one was the correct answer and the other was the distractor stimulus. If the error-correction procedure consists of re-presenting the S^D during the next trial the learner only has to switch their response between trials to make a correct response. The consultants stated that they did not want to reinforce switching between stimuli from one trial to the next. In these cases the error-correction procedure was selected based on how the programmes were designed. The error-correction procedures were not prescribed based on formal assessment. The video-recordings showed that therapists conducted frequent reinforcer preference assessments but it never was observed nor reported by the consultants that any formal assessments were conducted to select error-correction procedures.

However, selecting error-correction procedures in the aforementioned manner may not necessarily be best for the learner in terms of increasing the frequency and the probability of correct responses during acquisition trials. Research suggests that error-correction procedures requiring greater active learner participation appear to be more effective (Barbetta et al., 1993; McGhan & Lerman, 2013; Rogers & Iwata, 1991; Worsdell et al., 2005).

McGhan and Lerman (2013) investigated the use of a rapid error-correction identification procedure to identify appropriate and effective error-correction procedures. The aim of the assessment was to identify error-correction procedures that were the least intrusive but also the most effective. Four error-corrections were compared in the study. First, verbal-feedback only (all stimuli were removed and no error-correction trials were administered); second, model (the therapist models the correct response on the next trial, and no error-correction trials were administered); third, active student response (therapist provides a gestural prompt and vocal feedback about what the correct response was and re-presents the S^D); and fourth, direct rehearsal (the therapist provides a gestural prompt and vocal feedback about what the correct response was should be and re-presented the S^D from the previous trial. Following the prompted trial, if the learner responded correctly, the therapist re-presented the S^D until the learner made three consecutive unprompted correct responses).

The study consisted of two phases; an assessment phase and a validation phase. The assessment phase consisted of comparing the number of trials administered during each error-correction procedure until skills were mastered. During the validation phase the error-correction procedure identified as the most effective in the assessment phase was compared to two other procedures; one that was more intrusive and one that was less intrusive. The results showed that for four of the five participants the results from the validation phase was consistent with that of the assessment phase. The study provided evidence that initial error-correction assessment checks can be beneficial. It is recommended here that the consultants consider incorporating procedures, such as the one introduced by McGhan and Lerman (2013), to identify effective error-correction procedures.

The results from the task sequencing procedure analyses raise further questions about how some parts of the DTT programmes had been prescribed. The results from the task sequencing analyses suggest that not all of the learners benefited from interspersed training. It is not possible to determine from the task-sequencing analyses whether it was the interspersed training procedure itself that was ineffective. However, several possible factors that could be contributing to the diminished effectiveness of the interspersed training procedure were discussed in the Study 2 discussion section. Those factors indicate that there are problems with the way in which the interspersed training procedure has been prescribed.

It could be that interspersed training is a procedure that is not effective for all learners. Consultants should not assume that interspersed training will be effective simply because it has been effective with other clients. Research such as Charlop, Kurtz, and Milstein (1992) has shown that the effectiveness of the interspersed training procedure could be compromised if learners received edible reinforcers for both acquisition and mastered trials. Therefore, it does not appear to be the case that a procedure will be effective simply because it is being administered or because it has been effective with other clients. The procedure has to be prescribed and administered in a manner that will maximize the benefits of the procedure.

The manner in which the within-trial components have been prescribed generally did not deviate from what was described in Chapter II (p. 32 – 40). The only concern was with the ITI duration for Dyads 6, 7 and 8. The prescribed length of the ITI was 0.5 s, which is less than the minimum recommended duration of 1 s, and in practice a 0.5 s ITI is effectively no ITI.

Markov Transition Matrices and Extended Markov Chains

One of the specific aims of the thesis was to investigate the clinical utility of MTM and extended Markov probability chains with clinically important populations. The DTT procedure was used to evaluate the clinical utility of Markov probability chains because it is a procedure consisting of multiple components that are administered in a particular order. This feature of the procedure makes it suitable for sequential analysis, especially MTM.

The advantage of methods such as MTM is that the matrices are exhaustive (Gottman & Notarius, 1978). In other words, they display the frequencies and probabilities of all possible transitions between all components of the procedure which allowed exploratory analyses to be conducted so that within-trial treatment integrity errors could be identified. One of the criticisms of MTM and extended Markov probability chains has been that they can produce a large amount of data output, especially when several variables are being coded

(Sackett, 1979). However, for the purposes of Study 1 this feature of MTM proved to be useful rather than problematic. In addition the results from the social validity questionnaires showed that the programme consultants and therapists all either agreed or strongly agreed that the results were presented in an easy-to-understand manner.

MTM were preferred over other sequential analysis methods to conduct the treatment integrity analyses because the matrices are exhaustive, and the meaning of the transitional probability values are relatively easy to understand. The transitional probabilities also provide more descriptive information than a method such as Yule's Q which provides a single index of association (Lloyd, Kennedy, & Yoder, 2013). Lloyd et al. stated that "Yule's Q is the recommended method for statistically quantifying sequential associations between two events," (p. 480).

As a comparison to the Markov probabilities the within-trial treatment integrity data were analysed using Yule's Q. In the within-trial treatment integrity analyses Yule's Q provided results that were difficult to interpret. For example the Dyad 3 (Study 1) data showed that the transitional probability of the sequence INCORRECT – CORRECT (learner self-correction) was .05, which indicated that approximately 5% of all incorrect responses were followed by correct responses. The Yule's Q statistic for the same sequence was -0.59 indicating that there was a moderate negative association between these events. The negative Yule's Q value is the result of the relatively low frequency of the INCORRECT – CORRECT sequence compared to the frequencies of other (i.e., not correct responses) events following incorrect responses. The Yule's Q value is difficult to interpret in terms of whether the association between the two events was large enough to warrant intervention. From looking at the Yule's Q value in isolation it may be concluded that no intervention is necessary given that there was a moderate negative association between the events. However, the transitional probability of the sequence, along with the other transitional probabilities presented as part of an exhaustive MTM provided results were easier to interpret. Using a .05 criterion it was decided that the probability of the sequence was large enough to warrant intervention.

The results from the thesis provided two examples of where MTM were used in novel ways to produce results of clinical significance while mitigating some the concerns associated with their usage. For Study 1 they were used to identify within-trial treatment integrity errors. The one-step MTM were also useful in evaluating the local effects of specific types of within-trial treatment integrity errors; error sequences associated with the incorrect application of the within-trial error-correction procedures. The research question of interest for those analyses was whether these within-trial error-correction errors would adversely affect learner performance when the error was followed by an acquisition trial. In other words the analyses isolated the local effects of the error on learner performance. This is an advantage over the between-sessions methods which cannot isolate the effects of individual trial outcomes. The results showed that within-trial error-correction errors were correlated with impaired performance on acquisition trials if the acquisition trial immediately followed the error.

One of the original aims of Study 1 was to plot the data using extended Markov probability chains of three or four events. The intent was to examine longer chains for treatment integrity errors and to take an exploratory approach to see if any extended sequences could be identified that could improve the quality of the DTT programmes. The results from the extended Markov chains analyses were somewhat redundant in terms of increasing treatment integrity. For example the sequence INCORRECT – CORRECT – CORRECT CONSEQUENCE indicates two treatment integrity errors; the first error is the learner self-correction, and the second error is providing a correct consequence following the self-correction. If the one-step error sequence of INCORRECT – CORRECT can be reduced to zero then it is no longer possible for the correct consequence to follow learner self-corrections. Therefore, it may be useful and interesting to know the results from the longer Markov probability chains analyses but not necessary to improve treatment integrity. The one-step transitions produced meaningful results and identified treatment integrity for which the therapist could be retrained and treatment integrity improved.

The extended Markov probability chain analysis appeared to be an appropriate analysis when evaluating interspersed training. The analyses were exploratory and intended to identify the most frequently occurring three-trial sequences that resulted in unprompted correct and incorrect acquisition trial responses. The analyses also consisted of conducting significance tests to investigate if sequences were occurring more or less often than expected by chance. The sequences identified as occurring most frequently varied between dyads which made it difficult to make general conclusions.

However, the results from the extended Markov probability chains analyses did uncover unexpected sequences. The results suggest that for some dyads there may be some benefit from strictly alternating mastered and acquisition S^D. Future controlled studies can be conducted to assess the effectiveness of interspersed training when strictly alternating the type of S^D versus administering repeated mastered S^D. Therefore although extended Markov probability chains analyses were not necessary to address the research question, they did uncover unexpected sequential dependencies which may have clinical value.

There were two cases in which the one-step MTM and extended Markov probability chains were not the most effective method available to answer the research question. One of the aims of Study 2 was to conduct exploratory analyses, similar to those conducted in Study 1, to identify treatment integrity errors for error-corrections procedures that consisted of a between-trials component. The one-step MTM were not suitable for this analysis given how the variables were coded (p. 165). Conditional probability matrices similar to the one-step MTM were constructed. The conditional probability matrices were similar to the one-step

MTM in that they were exhaustive; they displayed all possible trial outcomes given a particular error-correction procedure event. Setting up the conditional probability analyses in this way still allowed for exploratory analyses to be conducted so that treatment integrity errors could be identified. The errors that were identified could be targeted for therapist re-training. Therefore although the traditional Markov matrix setup was not appropriate to answer the research question the modified matrices which maintained the exhaustive coding feature of the MTM produced results of clinical significance.

The other example where the MTM and extended Markov probability chains were not the most effective method of analysing the data were the task sequencing analyses. In order to address the research question adequately the analysis had to isolate the probabilities of unprompted correct and incorrect acquisition trial responses following one or more mastered or acquisition trials. The exhaustive nature of the MTM meant that it was not suitable as the main method of analysis because it displays the frequencies and probabilities between all components of the procedure. For example, one of the sequences of interest was MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE. To conduct the analyses using the conditional probability analyses all instances where correct mastered trial responses were followed by an acquisition trial were isolated and the probability of making correct or incorrect acquisition trial responses were calculated. When using the MTM the probability of the MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE sequence was influenced by the number of times other sequences occurred. If 100 correct mastered trial responses were recorded and there were 20 MASTERED CORRECT RESPONSE – ACQUISITION CORRECT RESPONSE sequences then the Markov probability of the sequence would be .20. However, the learner only had 25 opportunities to make independent acquisition trial responses following MASTERED CORRECT RESPONSE. A conditional probability analysis would produce a probability value of .80 (20/25). The value is a better indicator of the effects of interspersing the mastered trials because it states that out of the 25 opportunities that the learner had to make independent acquisition trial responses following a correct mastered trial response they responded correctly on 20 such occasions. On the other hand the MTM shows that 100 times MASTERED CORRECT RESPONSE occurred and 20 times it was followed by ACQUISITION CORRECT RESPONSE. On the other 80 occasions it was followed by other events. The presence of these other events decreases the probability of the sequence, not because interspersed training was necessarily ineffective but because it took into account all components of the procedure.

Based on the findings from this thesis it is recommended that in order to conduct the MTM analyses and produce meaningful results approximately 300 – 400 trials worth of data, which equates to between 1500 and 2000 events, need to be collected. Collecting data from between 300 to 400 trials was enough for the within-trial treatment integrity checks and in most cases enough for all of the Study 2 analyses. Collecting data on approximately 300 to 400 trials can be done relatively quickly as was the case for Dyad 4. In the baseline phase 413 trials were collected across three baseline sessions, and 447 trials across three follow-up phase trials. For Dyad 4 the 413 baseline trials were administered in approximately 55 min (approximately 7 – 8 trials per minute), and the 447 follow-up phase trials were administered in approximately 77 min (approximately 5 – 6 trials per minute). When the researcher wants to target a specific type of transition or part of the procedure it is recommended that they continue video-recording until they have enough instances of that transition. Although the aforementioned recommendation comes with the caveat that more video-recordings mean that the delay in reporting the data will be increased. This was a concern expressed by one of the therapists in the social validity questionnaire.

No formal mathematical criteria were used to determine how much data were needed to conduct the analyses. Data were collected until meaningful and interpretable probability values were produced. Consider a situation in which the results from a Markov transition matrix reveals that two two-event sequences occurred with a probability of .80. However, one of the transitions occurred four out of a possible five times, and the other sequence occurred 80 out of a possible 100 times. Although the transitional probabilities are the same the researcher can have more confidence in the transitional probability resulting from the transition that occurred 80 times. Given the way the transitional probabilities within a Markov matrix are calculated, higher than expected probability values can be produced simply because not enough instances of a particular event was observed. It may not be prudent to recommend or suggest changes regarding a learner's DTT programme based on high probability values that were the result of not collecting enough data.

One way to get an idea of how much data are needed to conduct these Markov type analyses is to test whether the data series is stationary (data series here means the changing probabilities of a transition as the denominator of the probability increases). Stationary in this context means, "that the sequential structure of the data is the same independent of where in the sequence we begin," (Bakeman & Gottman, 1997, p. 138). One way of determining this is to divide the data series into two halves, and constructing a transition matrix for data from the first half, a transition matrix for data from the second half, and a transition matrix for data from the entire data series. A test for stationarity involves conducting parametric tests to determine if the values from the two transition matrices differ significantly from the data in transition matrix for the entire data series. One such parametric test is the likelihood ratio chi-squared test for stationarity (Gottman & Roy, 1990). A stationary data series (i.e., the transition matrices from the two transition matrices do not significantly differ from the full data series transition matrix) indicates that the probability values are fairly stable across the entire data series, and this may indicate that the extent of the dataset is sufficient.

The results from sequential analyses provide information only about the correlations between events and conclusions regarding causal relations cannot be drawn from any of the analyses conducted in this thesis (Anderson & Long, 2002; Bijou et al., 1968; Rahman, Oliver, & Alderman, 2010; Woods, Borrero, Laud & Borrero, 2010). This is particularly relevant for the analyses conducted in Study 2 evaluating the local effects of the within-trial error-correction treatment integrity errors, and the conditional probability analyses that were conducted for the task sequencing procedures. The fact that sequential analyses provide correlational information is less of a concern for the treatment integrity analyses that were conducted in both studies. The MTM from Study 1 and the conditional probability matrices from Study 2 were used in an exploratory fashion to indicate the frequencies and probabilities of sequences. The matrices were investigated for any error sequences that were occurring with a probability greater than .05 and were not used to make any inferences about function.

Level of Analysis

In this thesis the data were analysed on three levels; within-trials, between-trials, and between-sessions. Based on the results from this thesis one question that can be asked is what the appropriated level of analysis is for DTT data. The level of analysis should be determined by the aim of the research question, the type of data collected, and by which method is best suited to display the important relations within the data (Fahmie & Hanley, 2008). Fahmie and Hanley (2008) used the terms *distant* and *intimate* as metaphors for how data are analysed and displayed. These terms represents a continuum. At one end of the continuum is *intimate* which indicates that the analyses and data display methods are in close proximity to the raw data (i.e., event-by-event or second-by-second). At the other end of the continuum is *distant* which means that data are aggregated into units or bins and as such are distant from the raw data (i.e., percentage correct responses or blocks-of-trials analyses). The results, analyses, and data depiction methods presented in this thesis were representative of both ends of the continuum.

The advantage of analysing the data on a within-trials basis was that it allowed for specific types of treatment integrity errors to be identified. The disadvantage of the within-trial analyses is that they require large amounts of detailed and precise data recording of all prescribed events of the DTT procedure, which may be difficult and time-consuming. From a practical perspective it may not always be possible to collect data in such a way as to conduct these within-trial analyses. For example, within-trial analyses would not be possible if data were collected using discontinuous data collection methods (e.g., Cummings & Carr, 2009; Lerman, Dittlinger, Fentress, & Lanagan, 2011; Najdowski et al., 2009).

Between-trials analyses provided clinically useful findings that could not be obtained using the within-trials analyses or the between-sessions analyses. The between-trials analyses were conducted to evaluate the local effects of the within-trial error-correction treatment integrity errors. Evaluating the local effects of specific types of within-trials treatment integrity errors was not possible using an event-by-event analysis because it only looked at events occurring within a trial. The relatively low frequency of these within-trial errors meant that the between-sessions methods were also not appropriate because aggregating the trials into bins obscured the local effects of these errors. The between-sessions methods were useful for displaying general patterns and trends.

The focus of this thesis was to use various sequential analysis methods, specifically MTM and extended Markov probability chains to analyse DTT data. Sequential analysis methods represent the intimate end of the data analysis continuum. Using these sequential analysis methods several clinically relevant results were obtained that could not have been obtained had the data been analysed using only the traditional between-sessions methods.

However, it would be erroneous to suggest that DTT data should be analysed using these sequential analyses methods at the expense of the traditional whole-session methods. It is possible to conduct whole-session and sequential analysis simultaneously. It may even be necessary in some cases as it may not be appropriate to prolong data collection in order to perform sequential analysis if clinically relevant information can be obtained more quickly using the whole-session methods. For example, a whole-session analysis such as percentage correct per block of 20 trials could show within three or four data points a pattern of treatment integrity scores less than 40%. In that case it may be necessary to combine all of the within-session data and conduct the Markov analyses to try and identify the errors correlated with the decreases in the overall integrity scores. The only caveat is that conducting Markov analyses with a limited number of data points could lead to probability values that are difficult to interpret because these values may be inflated. The results from this thesis provided a demonstration of how clinically relevant information can be obtained when data are analysed on a more intimate basis.

Limitations and Future Research

The research presented in this thesis has a few limitations in addition to those mentioned in the Discussion sections of each study. The first limitation was the small number of participants. Eight therapist-learner dyads consisting of five learners, eight therapists, and four consultants participated in this study. Another limitation of the research was that in some cases not enough data could be collected to conduct certain analyses for all dyads. For example in Study 1 insufficient data were collected for Dyad 5 to conduct the within-trial analyses to the same extent as the other dyads. The same limitation was encountered for some of the Study 2 analyses. The Dyad 5 data could also not be used in any of the Study 2 analyses and not enough data were collected for Dyad 6 to conduct the between-trials error-correction treatment integrity analyses. The small number of participants and insufficient data for some dyads limits the extent to which generalized conclusions can be made.

Another limitation of the research as it relates to evaluating the clinical utility of MTM and extended Markov probability chains is that the research involved only DTT. There are other procedures that also consist of multiple components which also appear to be appropriate for Markov chain analyses. Some of these procedures include incidental teaching (Hart & Risley, 1975), and task analyses. Both of these procedures consist of multiple components in which events follow each other in a particular order. Analysing the data from these two procedures using a one-step MTM could produce clinically significant results. Future research can be conducted with these types of procedures analysing the data in the same way as in this thesis.

The Study 1 analyses can be expanded to include other within-trial components that were not included in the study. Other components that can be included are whether the therapists selected an effective reinforcer or conducted a reinforcer preference assessment (Koegel et al., 1977; Lafasakis & Sturmey, 2007; Sarokoff & Sturmey, 2004), or whether the learner is attending when the S^D is delivered (Koegel et al., 1977; Leblanc et al., (2005); Ryan & Hemmes, 2005). However, it must be noted that every time a new variable is added to the analysis the resulting output will increase.

The results from Studies 1 and 2 suggest several avenues for future research. Observational studies such as the ones conducted in this thesis are advantageous because they identify treatment integrity errors and other procedural concerns of clinical relevance as they occur in the natural environment. How these errors affect learner performance during acquisition trials can now be investigated in controlled studies (e.g., the repeated presentations of S^D). Controlled studies similar to Carroll, Kodak, and Fisher (2013) will increase the ecological validity of studies investigating the functional relationship between within-trial treatment integrity errors and learner performance during acquisition trials. A recent paper by DiGennaro Reed and Codding (2014) stressed the importance of conducting treatment integrity studies with increased levels of ecological validity.

It is yet unknown the extent to which between-trial error-correction errors identified in Study 2 affect learner performance on acquisition trials. Controlled studies can be conducted to assess how the errors identified in Study 2 affect learner performance.

Future research can involve developing apps for devices such as Smartphones so that data can be collected as DTT is taking place. Such an app could contain pre-programmed variables and analyses. The Obswin 3.4™ software could not be used to collect data in real time for a couple of reasons. The first reason was that the programme is best suited for devices such as laptops and PC computers, meaning it was impractical to record events as DTT was taking place. The second reason is that it required a wireless web camera and an internet connection to produce a live feed as DTT was taking place. These two reasons made it impractical to use the software to record DTT data in real time. A smaller device such as a Smartphone, with an app containing pre-programmed variables and analyses would be more convenient, and could allow for results to be produced much quicker than they were during Study 1. The Obswin 3.4 ™ software requires the video-recording to be downloaded onto PC computer before events could be coded and data analysed. Any device or app that can reduce the number of steps needed to produce the data needed for these analyses would be useful.

Conclusions

Given the findings of this thesis a useful evaluation tool would consist of three main features. The first feature is keeping the structure of the MTM. This feature is useful because the matrices are exhaustive; they can simultaneously display the frequencies and probabilities of all possible transitions between all components of the procedure. This feature is especially useful when analysing transitions from an intervention with multiple components such as DTT.

The second feature is a method that is less intensive in terms of data collection. The MTM required a large number of data points to produce meaningful and interpretable probability values. One possible solution would be to contract some of the variables, such as the different types of response prompts into one variable called PROMPTS. However, each time variables are contracted the analysis moves away from being a true event-by-event analysis, and potentially important details can be lost.

The third feature is having empirically established criteria for deciding when the probability of specific error sequences are high enough to warrant intervention. For example, two error sequences are identified; one occurs with a probability of .10, and the other with a probability of .05. It is possible that the sequence occurring with a probability of .10 does not affect learning until it occurs with a probability of .30, whereas the sequence occurring with a probability of .05 does affect learning when it occurs with a probability of .02. Having empirically established criteria will assists therapists and consults in making accurate clinical decisions when deciding how best to deal with treatment integrity errors, especially when more than one error is occurring. Future research which seeks to further develop within-session tools for assessing DTT data should take into consideration these three features.

The MTM, extended Markov probability chains, and other sequential analysis methods used in this thesis provided a thorough analysis of several aspects of DTT programmes occurring in the natural environment as part of the regular teaching programme for eight therapist-learner dyads. The use of sequential analysis to evaluate DTT programmes has been sparse. The results from this thesis provided a demonstration of how clinically relevant results that can be obtained when these analyses are conducted. The research and methods presented in this thesis provide a starting point for further research developing analyses of this kind.

Appendix A

Social Validity Questionnaire: Consultants

The purpose of this social validity questionnaire is to evaluate your experience with the method and procedures used to identify treatment integrity errors for the discrete-trial teaching (DTT) programmes you administer. If you have any additional feedback or comments you can write them in the blank space beneath Item 8. Thank you for your participation.

Please mark (X) the degree to which you agree with the following statements:

1. It is important to administer DTT with a high degree of treatment integrity.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

2. It is important for DTT treatment integrity errors to be identified and corrected.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

3. The results from video-recordings were presented to me in a clear and easy-to-understand manner.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

4. I was surprised by the treatment integrity errors that were identified as part of the DTT programmes for which I am the consultant.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

5. I found the results useful in helping in correcting treatment integrity errors for the therapists which I supervise.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

6. The results from the study were beneficial to the learner.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

7. I would recommend this method as an effective way to evaluate treatment integrity errors in DTT.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

8. Any further comments or feedback?

Appendix B

Social Validity Questionnaire: Therapists

The purpose of this social validity questionnaire is to evaluate your experience with the method and procedures used to identify treatment integrity errors for the discrete-trial teaching (DTT) programmes you administer. If you have any additional feedback or comments you can write them in the blank space beneath Item 10. Thank you for your participation.

Please mark (X) the degree to which you agree with the following statements:

1. It is important to administer DTT with a high degree of treatment integrity.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

2. It is important for DTT treatment integrity errors to be identified and corrected.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

3. The results from video-recordings were presented to me in a clear and easy-to-understand manner.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

4. I was surprised by the treatment integrity errors that were identified as part of the DTT that I administer.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

5. On days when video-recordings took place the learner exhibited more off-task or non-compliant behaviour than on days for which video-recordings were not made.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

6. I was comfortable being video-recorded.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

7. I found the results useful in helping me to correct treatment integrity errors.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

8. The results from the study were beneficial to the learner

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

9. I would recommend this method as an effective way to evaluate treatment integrity errors in DTT.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Don't Know
[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

10. Any further comments or feedback?

References

Alber, S. R., & Nelson, J. S., & Brennan, K. B. (2002). A comparative analysis of two homework study methods on elementary and secondary school students' acquisition and maintenance of social studies content. *Education and Treatment of Children, 25,* 172 – 196.

Alison, P.D., & Liker, J.K. (1982). Analysing sequential categorical data on dyadic interaction. *Psychological Bulletin, 91,* 393 – 403.

Altman, S.A. (1965). Sociobiology of Rhesus Monkeys II: Stochastics of social communication. *Journal of Theoretical Biology, 8,* 490-522.

Anderson, C.M., & Long, E.S. (2002). Use of a structured descriptive assessment methodology to identify variables affecting problem. *Journal of Applied Behavior Analysis, 35,* 137 – 154.

Anderson, S.R., Taras, M., & Cannon, B.O. (1996). Teaching new skills to young children with autism. In C. Maurice, G. Green, & S. Luce (Eds.), *Behavioral interventions for young children with autism: A manual for parents and professionals*. Austin, TX: Pro-Ed.

Arkoosh, M. K., Derby, K. M., Wacker, D. P., Berg, W., McLaughlin, T. F., & Barretto, A. (2007). A descriptive evaluation of long-term treatment integrity. *Behavior Modification, 31,* 880 – 895.

Armstrong, K. J., Ehrhardt, K. E., Cool, R. T., & Poling, A. (1997). Social validity and treatment integrity data: Reporting in articles in the Journal of Developmental and Physical Disabilities, 1991 – 1995. *Journal of Developmental and Physical Disabilities, 9,* 359 – 367.

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

Bakeman, R. (2001). Behavioral observation and coding. In H.T. Reis & C.M. Judd (Eds.). *Handbook of research methods in social and personality psychology* (138 – 160). Cambridge University Press.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to Sequential Analysis* (2nd Ed), Cambridge University Press.

Barbetta, P. M., Heron, T. E., & Heward, W. L. (1993). Effects of active student response during error correction on the acquisition, maintenance, and generalisation of sight words by students with developmental disabilities. *Journal of Applied Behavior Analysis, 26,* 111-119.

Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effect of two treatments in a single behavior. *Journal of Applied Behavior Analysis, 12,* 199 – 210.

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

Benavides, C. A., & Poulson, C. L. (2009). Task interspersal and performance of matching tasks by preschoolers with autism. *Research in Autism Spectrum Disorders, 3,* 619 – 629.

Berchtold, A., & Sackett, G.P. (2002). Markovian models for the development study of social behavior. *American Journal of Primatology, 58,* 149 – 167.

Bibby, P., Eikeseth, S., Martin, N. T., Mudford, O. C., Reeves, D. (2002). Progress and outcomes for children with autism receiving parent-managed intensive interventions. *Research in Developmental Disabilities, 23,* 81-104.

Bijou, S.W., Peterson, R.F., & Ault, M.H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis, 1,* 175 – 191.

Bondy, A., & Frost, L. (2002). *The picture exchange communication system*. Newark, DE: Pyramid Educational Products.

Bonett, D.G., & Price, R.M. (2007). Statistical inference for generalized Yule coefficients in 2 x 2 contingency tables. *Sociological Methods & Research, 35, 429 – 445.*

Borrero, C.S.W., & Borrero, J.C. (2008). Descriptive and experimental analyses of potential precursors to problem behavior. *Journal of Applied Behavior Analysis, 41,* 83 – 96.

Carr, E. G., & Dores, P. A. (1981). Patterns of language acquisition following simultaneous communication with autistic children. *Analysis and Intervention in Developmental Disabilities, 1,* 347 - 361.

Carr, E. G., Kologinsky, E., & Leff-Simon, S. (1987). Acquisition of sign language III: Generalized descriptive phrases. *Journal of Autism and Developmental Disorders, 17,* 217 – 229.

Carroll, R. A., Kodak, T., & Fisher, W. W. (2013). An evaluation of programmed treatment integrity errors during discrete-trial instruction. *Journal of Applied Behavior Analysis,* 46, 379 – 394.

Catania, A. C. (1963). Concurrent performances: A baseline for the study of reinforcement magnitude. *Journal of the Experimental Analysis, 6,* 299 – 300.

Catania, C. N., Almeida, D., Liu-Constant. B, & DiGennaro Reed, F. D. (2009). Video modeling to train staff to implement discrete-trial instruction. *Journal of Applied Behavior Analysis, 42,* 387 - 392.

Charlop, M. H., Kurtz, P. F., & Milstein, J. P. (1992). Too much reinforcement, too little behavior: Assessing task interspersal procedures in conjunction with different reinforcement schedules with autistic children. *Journal of Applied Behavior Analysis, 25*, 795 – 808.

Chasson, G. S., Harris, G. E., & Neely, W. J. (2007). Cost comparison of early intensive behavioural interventions and special education for children with autism. *Journal of Children and Family Studies, 16,* 401 – 413.

Chen, C.C., McComas, J.J., Hartman, E., & Symons, F.J., (2011). A prospective sequential analysis of the relation between physical aggression and peer rejection acts in a high-risk preschool sample. *Early Education and Development, 22*, 574 – 592.

Chong, I. M., & Carr, J. E. (2005). An investigation of the potentially adverse effects of task interspersal. *Behavioral Interventions, 20*, 285 – 300.

Coe, D., Matson, J., Fee, V., Manikam, R., & Linarello, C. (1990). Training nonverbal and verbal play skills to mentally retarded and autistic children. *Journal of Autism and Developmental Disorders, 20*, 177 - 187.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimental design and analysis issues for field settings.* Boston: Houghton Miffin.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd Ed). Upper Saddle River, New Jersey: Pearson.

Crockett, J. L., Fleming, R. K., Doepke, K. J., & Stevens, J. S. (2007). Parent training: Acquisition and generalisation of discrete trials teaching skills with parents of children with autism. *Research in Developmental Disabilities, 28,* 23 - 36.

Cummings, A. R., & Carr, J. E (2009). Evaluating programs for children with autism spectrum disorder via continuous and discontinuous measurement. *Journal of Applied Behavior Analysis, 42*, 57 - 71.

Cuvo, A. J., Klevans, L., Borakove, S., Borakove, L. S., Van Landuyt, J., & Lutzker, J. R. (1980). A comparison of three strategies for teaching object names. *Journal of Applied Behavior Analysis, 13*, 249 – 257.

Davison, M., & McCarthy, D. (1980). Reinforcement for errors in a signal detection procedure. *Journal of the Experimental Analysis of Behavior, 34*, 35 - 47.

Davison, M. C., & Tustin, R. D. (1978). The relation between the generalised matching law and signal-detection theory. *Journal of the Experimental Analysis of Behavior, 29*, 331-336.

Delprato, D. J. (2001). Comparisons of discrete-trial and normalized behavioral language intervention for young children with autism. *Journal of Autism and Developmental Disorders, 31*, 315 - 325.

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

DiGennaro, F. D., Martens, B. K., & Kleinman, A. E. (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.

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

DiGennaro Reed, F. D., Reed, D. D., Baez, C. N., & Maguire, H. (2011). A parametric analysis of errors of commission during discrete-trial training. *Journal of Applied Behavior Analysis, 44,* 611 - 615.

DiGennaro Reed, F. D., & Codding, R. (2014). Advancements in procedural fidelity assessment and intervention: Introduction to the special issue. *Journal of Behavioral Education, 23,* 1 – 18.

Dogan, O. S., & Tekin-Iftar, E. (2002). The effects of simultaneous prompting on teaching receptively identifying occupations from picture cards. *Research in Developmental Disabilities, 23,* 237 - 252.

Downs, A., Downs, R. C., & Rau, K. (2008). Effects of training and feedback on discrete-trial teaching skills and student performance. *Research in Developmental Disabilities, 29,* 235 - 246.

Doyle, P. M., Wolery, M., Ault, M. J., Gast, D. L., & Wiley, K. (1989). Establishing conditional discriminations: Concurrent versus isolation-intermix instructions. *Research in Developmental Disabilities, 10,* 3469 – 362.

Dunlap, G. (1984). The influence of task variation and maintenance tasks on the learning and affect of autistic children. *Journal of Experimental Child Psychology, 37,* 41 – 64.

Dunlap, G., & Koegel, R. L. (1980). Motivating autistic children through stimulus variation. *Journal of Applied Behavior Analysis, 13,* 619 – 627.

Eikeseth, S., Hayward, D., Gale, C., Gitlesen, J. P., & Eldevik, S. (2009). Intensity of supervision and outcome for preschool aged children receiving early and intensive behavioral interventions: A preliminary study. *Research in Autism Spectrum Disorders, 3,* 67 – 73.

Eikeseth, S., Klintwall, L., Jahr, E., & Karlsson, P. (2012). Outcome for children with autism receiving early and intensive behavioral intervention in mainstream preschool and kindergarten settings. *Research in Autism Spectrum Disorders, 6,* 829-835.

Fahmie, T.A., & Hanley, G.P. (2008). Progressing towards data intimacy: A review of within-session data analysis. *Journal of Applied Behavior Analysis, 41,* 319 – 331.

Ferster, T.A., Skinner, B.F. (1957). *Schedules of reinforcement*. New York: Appleton-Century-Crofts.

Foster, S. L., & Cone, J. D (1986). Design and use of direct observation systems. In A. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), Handbook of Behavioral Assessment. (pp. 253 – 324). New York: John Wiley.

Fox, R. M. (1982). *Increasing behaviors of persons with severe retardation and autism.* Research Press: Illinois.

Fryling, M. J., Wallace, M. D., & Yassine, J. N. (2012). Impact of treatment integrity on intervention effectiveness. *Journal of Applied Behavior Analysis, 45,* 449 – 453.

Ghezzi, P.M. (2007). Discrete trials teaching. *Psychology in the Schools, 44,* 667 - 679.

Gottman, J.M., & Notarius, C. (1978). Sequential analysis of observational data using Markov chains. In Kratochwill, T.R. (Ed), *Single subjects research: Strategies for evaluating change* (237 – 286). Academic Press: New York.

Gottman, J.M., & Roy, A.K. (1990). *Sequential analysis: A guide for behavioural researchers.* Cambridge University Press.

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

Grindle, C. F., & Remington, B. (2002). Discrete trial training for autistic children when reward is delayed: A comparison of conditioned cue value and response marking. *Journal of Applied Behavior Analysis, 35,* 187 - 190.

Grow, L. L., Carr, J., Gunby, K. V., Charania, S. M., Gonslaves, L., Ktaech, I. A., & Kisamore, A. N. Deviations from prescribed prompting procedures: Implications for treatment integrity. *Journal of Behavioral Education, 18,* 142-156.

Hart, B., & Risley, T. R. (1975). Incidental teaching in language in the preschool. *Journal of Applied Behavior Analysis, 8,* 411 – 420.

Heckaman, K., A., Alber, S. R., Hooper, S., & Heward, W. L. (1998). A comparison of least-to-most prompts and progressive time delay on the disruptive behavior of students with autism. *Journal of Behavioral Education, 8*, 171 – 201.

Holcombe, A., Wolery, M., & Snyder, E. (1994). Effects of two levels of procedural fidelity with constant time delay on children's learning. Journal of Behavioral Education, 4, 49-73.

Holding, E., Bray, M., & Kehle, T. J. (2011). Does speed matter? A comparison of the effectiveness of fluency and discrete trial training for teaching noun labels to children with autism. *Psychology in the Schools, 48,* 166 – 183.

Hollard, V., & Davison, M. C. (1972). Preference for qualitatively different reinforcers. *Journal of the Experimental Analysis of Behavior, 16*, 375 – 380.

Jacobson, J. W., Mulick, J. A., & Green, G. (1998). Cost-benefit estimates for early intensive behavioral intervention for young children with autism: General models and single state case. *Behavioral Interventions, 13,* 201 – 226.

Jennett, H. K., Harris, S. L., & Delmolino, L. (2008). Discrete trial instruction vs. mand training for teaching children with autism to make requests. *The Analysis of Verbal Behavior, 24,* 69 - 85.

Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics for human behavioral research.* Hillside, NJ: Erlbaum.

Johnston, J.M., & Pennypacker, H.S. (2009). *Strategies and tactics of behavioral research* (3rd Ed.). New York: NY.

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

Jose, P.E. (1988). Sequentiality of speech acts in conversational structure. *Journal of Psycholinguistic Research, 17,* 65 – 88.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10,* 141-150.

Kazdin, A. E. (1998). Research Design in Clinical Psychology. (3rd Ed). Homewood: Dorsey Press.

Koegel, R. L., Dunlap, G., & Dyer, K. (1980). Intertrial interval duration and learning in autistic children. *Journal of Applied Behavior Analysis, 13,* 91 - 99.

Koegel, R. L., & Glahn, T. J., & Nieminen, G. S. (1978). Generalisation of parent-training results. *Journal of Applied Behavior Analysis, 11,* 95 - 109.

Koegel, L. K., & Koegel, R. L. (1986). The effects of interspersed maintenance tasks on academic performance in a severe childhood stroke victim. *Journal of Applied Behavior Analysis, 19*, 425 – 430.

Koegel, R. L., Russo, D. C., & Rincover, A. (1977). Assessing and training teachers in the generalized use of behavior modification with autistic children. *Journal of Applied Behavior Analysis, 10*, 197 - 205.

Lafasakis, M., & Sturmey, P. (2007). Training parent implementation of discrete-trial teaching: Effects on generalisation of parent teaching and child correct responding. *Journal of Applied Behavior Analysis, 40,* 685-689.

Leaf, R., & McEachin, J. (1999). *A work in progress: Behavior management strategies and curriculum for intensive behavioral treatment of autism*. New York: DRL Books

Leaf, J.B., Sheldon, J.B., & Sherman, J.A. (2010). Comparison of simultaneous prompting and no-no prompting in two-choice discrimination learning with children with autism. *Journal of Applied Behavior Analysis, 43,* 215 - 228.

LeBlanc, M., Ricciardi, J. N., & Luiselli, J. K. (2005). Improving discrete trial instruction by paraprofessional staff through an abbreviated performance feedback intervention. *Education and Treatment of Children, 28,* 76 - 82.

Lerman, D. C., Dittlinger, L. H., Fentress, G., & Lanagan, T. (2011). A comparison of methods for collecting data on performance during discrete-trial teaching. *Behavior Analysis in Practice, 4,* 53 – 63.

Lerman, D.C., & Iwata, B.A. (1993). Descriptive and experimental analyses of variables maintaining self-injurious behavior. *Journal of Applied Behavior Analysis*, *26*, 293 - 319.

Liaupsin, C. J., Ferro, J. B., & Umbreit, J. (2012). Treatment integrity in intervention research: Models, measures, and future directions. *Advances in Learning and Behavioral Disabilities, 25,* 301-322.

Lloyd, B. P., Kennedy, C. H., & Yoder, P. J. (2013). Quantifying contingent relations from direct observation data: Transitional probability comparisons versus Yule's Q. *Journal of Applied Behavior Analysis, 46,* 479 – 497.

Lovaas, I, O., (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, 55, 3 -9.

Lovaas, O.I. (2003). *Teaching individuals with developmental delays: Basic intervention techniques*. Austin, TX: Pro-Ed.

Love, J. R., & Carr, J. E., Almason, S., & Petursdottir, A., I. (2009). Early and intensive behavioral intervention for autism: A survey of clinical practices. *Research in Autism Spectrum Disorders, 3,* 421 – 428.

Lutzket, J. R., & Sherman, J. A. (1974). Producing generative sentence usage by imitation and reinforcement procedures. *Journal of Applied Behavior Analysis, 7,* 447 – 460.

Mackintosh, N. J. (1974). *The psychology of animal learning.* New York: Academic Press.

MacLean W.E, Tapp J.T, Johnson W.L. (1985). Alternate methods and software for calculating interobserver agreement for continuous observation data. *Journal of Psychopathology and Behavioral Assessment; 7,* 65–73.

Magiati, I., Charman, T., Howlin, P. (2007). A two-year prospective follow-up study of community-based early intensive behavioural intervention and specialist nursery provision for children with autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 48,* 803-812.

Martens, B.K., DiGennaro, F.D., Reed, D.D., Szczech, F.M., & Rosenthal, B.D. (2008). Contingency space analysis: An alternative method for identifying contingent relations from observational data. *Journal of Applied Behavior Analysis, 41,* 69 – 81.

Martin, N., Oliver, C., & Hall, S. (1999). ObsWin: Observational Data Collection & Analysis for Windows. *CTI Psychology Software News,* ***9***, 14-16.

Matthews, B. K., Shimoff, E., & Catania, C. A. (1987). Saying and doing: A contingency-space analysis. *Journal of Applied Behavior Analysis, 20,* 69 – 74.

McEachin, J. J., Smith, T., & Lovaas, O. I. (1993). Long-term outcome for children with autism who received intensive behavioural treatment. *American Journal of Mental Retardation, 97,* 359-372.

McComas, J.J., Moore, T., Dahl, N., Hartman, E., Hoch, J., & Symons, F. (2009). Calculating contingencies in natural environments: Issues in the application of sequential analysis. *Journal of Applied Behavior Analysis, 42,* 413 – 423.

McGee, G. G., Krantz, P. J., & McClannahan, L. E. (1985). The facilitative effects of incidental teaching on preposition use by autistic children. *Journal of Applied Behavior Analysis, 18,* 17 - 31.

McGhan, A. C., Lerman, D.C. (2013). An assessment of error-correction procedures for learners with autism. *Journal of Applied Behavior Analysis, 46*, 626 – 639.

Miranda-Linne, F., & Melin, L. (1992). Acquisition, generalisation, and spontaneous use of color adjectives: A comparison of incidental teaching and traditional discrete-trial procedures for children with autism. *Research in Developmental Disabilities, 13*, 191 – 210.

Monchar, F. J., & Prinz, R. J. (1991) Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247 – 266.

Montague, E., Xu, J., Chen, P., Asan, O., Barrett, B.P., & Chewing, B. (2011). Modeling eye gaze patterns in clinician-patient interaction with lag sequential analysis. *Human Factors: The Journal of the Human Factors and ergonomics Society, 53,* 502 – 516.

Mueller, M. M., Piazza, C. C., Moore, J. W., Kelley, M. E., Bethke, S. A., Pruett, A. E., Oberdorff, A. J., & Layer, S. A. Training parents to implement pediatric feeding protocols. *Journal of Applied Behavior Analysis, 36,* 545 – 562.

Najdowski, A. C., Chilingaryan, V., Bergstrom, R., Granpeesheh, D., Balsanyan, S., Aguila, B., & Tarbox, J. (2009). Comparison of data-collection methods in a behavioral intervention program for children with pervasive developmental disorders: A replication. *Journal of Applied Behavior Analysis, 42,* 827 – 832.

Neef, N. A., & Iwata, B. A., & Page. (1980). The effects of interspersal training versus high-density reinforcement on spelling acquisition and retention. *Journal of Applied Behavior Analysis, 13*, 153 – 158.

Noell, G. H., Gresham, F. M., & Gansle, K. A. (2002). Does treatment integrity matter? A preliminary investigation of instructional implementation and mathematics performance. *Journal of Behavioral Education, 1*, 51 – 67.

Oliver, C., Woodcock, K.A., & Humphreys, G.W. (2009). The relationship between components of the behavioural phenotype in Prader-Willi syndrome. *Journal of Applied Research in Intellectual Disabilities, 22,* 403 – 407.

Panyan, M. C., & Hall, R. V. (1978). Effects of serial versus concurrent task sequencing on acquisition, maintenance and generalisation. *Journal of Applied Behaviour Analysis, 11*, 67 – 74.

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

Rahman, B., Oliver, C., & Alderman, N. (2010). Descriptive analysis of challenging behaviours shown by adults with acquired brain injury. *Neuropsychological Rehabilitation, 20*, 212 – 238.

Rodgers, T. A., & Iwata, B. A. (1991). An analysis of error-correction procedures during discrimination training. *Journal of Applied Behavior Analysis*, *24,* 775-781.

Rowan, V. C., & Pear, J. J. (1985). A comparison of the effects of interspersal and concurrent training sequences on acquisition, retention, and generalisation of picture names. *Applied Research in Mental Retardation, 6*, 127 – 145.

Ryan, C.S., & Hemmes, N. S. (2005). Post-training discrete-trials teaching performance by instructors of young children with autism in early intensive behavioral intervention. *The Behavior Analyst Today, 6,* 1 – 6.

Sackett, G.P. (1979). The lag sequential analysis of contingency and cyclicity in behavioural interaction research. In J.D. Osofsky (Eds.), *Handbook of infant development* (623 – 649). New York.

Sallows, G. O., & Graupner, D. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. *American Journal of Mental retardation, 6,* 417 – 438.

Samaha, A.L., Vollmer, T.R., Borrero, C., Sloman, K., Pipkin, C.S., & Bourret, J. (2009). Analyses of response-stimulus sequences in descriptive observations. *Journal of Applied Behavior Analysis, 42,* 447 – 468.

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

Sarokoff, R., & Sturmey, P. (2004). The effects of behavioral skills training on staff implementation of Discrete-trial teaching. Journal of Applied Behavior Analysis, 37, 535 – 538.

Sarokoff, R. A., & Sturmey, P. (2008). The effects of instructions, rehearsal, modeling and feedback on acquisition and generalisation of staff use of discrete-trial teaching student correct responses. *Research in Autism Spectrum Disorders, 2,* 125 - 136.

Schroeder, G. L., & Baer, D. M. (1972). Effects of concurrent and serial training on generalized vocal imitation in retarded children. *Developmental Psychology*, *6*, 293 – 301.

Schroeder, C. M., Schuster, J. W., & Hemmeter, M. L. (1998). Efficiency in programming generalisation: Comparison of two methods for teaching expressive labeling to preschoolers with developmental delays. *Journal of Developmental and Physical Disabilities*, *10*, 109 – 131.

Schuster, J. W., Griffen, A. K., & Wolery, M. (1992). Comparisons of simultaneous prompting and time delay procedures in teaching sight words to elementary students with moderate mental retardation. *Journal of Behavioral Education, 2*, 305 - 325.

Smith, T. (2001). Discrete trial training in the treatment of autism. *Focus on Autism and Other Developmental Disabilities, 16*, 86 - 92.

Smith, R.G., & Churchill, R.M. (2002). Identification of environmental determinants of behavior disorders through functional analysis of precursor behaviors. *Journal of Applied Behavior Analysis*, *35,* 125 – 136.

Smith, T**.,** Mruzek, D. W., Wheat, L. A., & Hughes, C. (2006). Error correction in discrimination training for children with autism. *Behavioral Interventions, 21,* 245-263.

Stock, R., Mirenda, P., & Smith, I. M. (2013). Comparison of community-based verbal behavior and pivotal response treatment programs for young children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 7,* 1168 – 1181.

St. Peter Pipkin, C., Vollmer, T. R., & Sloman, K. N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: A translation model. *Journal of Applied Behavior Analysis, 43,* 47 – 70.

Stuart, B. (1971). Behavioral contracting within the families of delinquents. *Journal of Behavior Therapy and Experimental Psychiatry, 2,* 1 – 11.

Summers, J., & Szatmari, P. (2009). Using discrete trial instruction to teach children with Angelman Syndrome. *Focus on Autism and Other Developmental Disabilities, 24,* 216 - 226.

Sundberg, J.W., & Partington, M.L. (1998). *Teaching language to children with autism or other developmental disabilities.* Pleasant Hill, CA: Behavior Analysts, Inc

Thomson, K., Martin, G. L., Arnal, L., Fazzio, D., & Yu, C.T. (2009). Instructing individuals to deliver discrete-trials teaching to children with autism spectrum disorders: A review. *Research in Autism Spectrum Disorders, 3,* 590-606.

Thompson, R.H., & Iwata, B.A. (2001). A descriptive analysis of social consequences following problem behavior. *Journal of Applied Behavior Analysis, 34,* 169 – 178.

Touchette, P.E., MacDonald, R.F., & Langer, S.N. (1985). A scatterplot for identifying stimulus control of problem behavior. *Journal of Applied Behavior Analysis, 18,* 343 – 351.

Valcante, G., Roberson, W., Reid, W. R., & Wolking, W. D. (1989). Effects of wait-time and intertrial interval durations on learning by children with multiple handicaps. *Journal of Applied Behavior Analysis, 22*, 43 - 55.

Vismara, L. A., & Rogers, S. J. (2010). Behavioral treatments in autism spectrum disorder: What do we know? *Annual Review of Clinical Psychology, 6,* 447 – 468.

Vladescu, J. C., Carroll, R., Paden, A., & Kodak, T. M. (2012). The effects of video modeling with voiceover instruction on accurate implementation of discrete-trial instruction. *Journal of Applied Behavior Analysis, 45*, 419 – 423.

Vollmer, T.R., Borrero, J.C., Wright, C.S., Van Camp, C., & Lalli, J.S. (2001). Identifying possible contingencies during descriptive analyses of severe disorders. *Journal of Applied Behavior Analysis, 34,* 269 – 287.

Waldo, L., Guess, D., & Flannagan, B. (1982). Effects of concurrent and serial training on receptive labeling by severely retarded individuals. *Journal of the Association for the Severely Handicapped, 6*, 56 – 65.

Watson, P. J., & Workman, E. A. (1981). The non-concurrent multiple baseline across-individuals design: An extention of the traditional multiple baseline design. *Journal of Behavior Therapy and Experimental Psychiatry, 12,* 257-259.

Weber, R. C., & Thorpe, J. (1992). Teaching children with autism through task variation in physical education. *Exceptional Children,* 59, 77 – 86.

Weingart, L.R., Prietula, M.J., Hyder, E., & Genovese, C. (1999). Knowledge and the sequential processes of negotiation: A Markov chain analysis of response-in-kind. *Journal of Experimental Social Psychology, 35,* 366 – 393.

Wheeler, J. J., Baggett, B. A., 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.

White, K. G., & Wixted, J. T. (1999). Psychophysics of remembering. *Journal of the Experimental Analysis of Behavior, 71*, 91-113.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1,* 80 – 83.

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

Williams, J. A., Koegel, R. L., & Egel, A. L. (1981). Response-reinforcer relationships and improved learning in autistic children. *Journal of Applied Behavior Analysis, 14*, 53 - 60.

Wilson, N.C. (2008). Analysis of observational data on the behaviour of individuals: Conditional probability analysis and time-base lag sequential analysis. Unpublished Masters Thesis. University of Auckland.

Wolf, M.M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. Journal of Applied Behavior Analysis, 11, 203 – 214.

Wolery, M., Holcombe, A. M., Werts, M. G., & Cipolloni, R. M. (1993). Effects of simultaneous prompting and instructive feedback. *Early Education and Development, 4*, 20-31.

Woods, J.N., Borrero, J.C., Laud, R.B., & Borrero, C.S.W. (2010). Descriptive analyses of paediatric food refusal: The structure of parental attention. *Behavior Modification, 34*, 35 – 56.

Worsdell, A. S., Iwata, B. A., Dozier, C. L., Johnson, A. D., Neidert, P. LO., & Thomason, J. L. (2005). Analysis of response repetition as an error-correction strategy during sight-wording reading. *Journal of Applied Behavior Analysis, 38*, 511-527.

Wright, A. A., Cook, R. G., Rivera, J. J., Sands, S. F., & Delius, J. D. (1988). Concept learning by pigeons: Matching to sample with trial-unique video picture stimuli. *Animal Learning and Behavior, 16*, 436-444.

Yoder, P.J., Short-Meyerson, K., & Tapp, J. (2004). Measurement of behavior with a special emphasis on sequential analysis of behavior. In E. Emerson., C Hatton, T. Thompson & T.R Parmenter (Eds). *Applied Research in intellectual disabilities* (179 – 190). San Francisco.

Yoder, P. J., & Symons, F. (2010). *Observational measurement of behavior*. New York: Springer Publishing Company.

Yoder, P.J., & Tapp, J. (2004). Empirical guidance for time-window sequential analysis of single cases. *Journal of Behavioural Education, 13*, 227 – 246.

Young, J. M., Krantz, P. J., McClannahan, L. E., & Poulson, C. L. (1994) Generalized imitation and response-class formation in children with autism. *Journal of Applied Behavior Analysis, 27*, 685 - 697.

Yule, G. V., & Kendall, M. G. (1957). *An introduction to the theory of statistics*. London, England: Griffin.