CHAPTER 2

LITERATURE REVIEW

2.1 Behavioral Modification on Academic Behaviors

Behavioral modification is based on the assumption that maladaptive behaviors are governed by antecedent and consequent events (Simpson, 1998). Thus, it is important to identify and manipulate relevant antecedent and consequent stimuli to rebuild new targeted behaviors. To enhance academic behavior changes, applied behavior investigators have conducted numerous functional behavioral analyses to discover effective intervention strategies.

Intervention strategies which have been used in promoting academic behaviors are outlined as follows:

Computer Assisted Instruction

Higgins and Boone (1990) explored the use of computer-assisted study guides in an A-B-A design. Their results showed that computer-assisted study guide treatment was an effective mean of instruction. Howell, Sidorenko, and Jurica (1987) conducted two investigations to evaluate the effectiveness of the use of drill-and-practice software and tutorial-based software on the acquisition of multiplication facts separately. They found that both drill-and practice and tutorial software were effective in increasing correct responses of students with learning disabilities.

Cooperative Learning

Pigott, Fantuzzo, and Clement (1986) used a peer-mediated group contingencies intervention to evaluate the effect on the arithmetic performance of underachievers. In their peer-mediated group contingencies intervention, students in a small group performed separately in the roles of coach, scorekeeper, referee, and manager, each with a stake correspondent with the operation of a peer tutoring (i.e., peer instruction, peer observation, peer evaluation, and peer reinforcement), and gained group-determined backup reinforcers if they met their goals four times. Results demonstrated that the intervention improved the average number of arithmetic problems completed accurately and the mean accuracy percentage for all treated students.

An intervention that embraced self-management, cooperative learning, and interdependent group-oriented contingencies was conducted by Olympia, Sheridan, Jenson, and Andrews (1994). As well as evaluating the potency of the intervention, they compared the relative effectiveness of interventions under the conditions of self-selected goals versus teacher-selected goals on homework completion, accuracy, and standardized achievement and curriculum-based measures. For homework completion, averaged percentage of arithmetic homework assignments returned across all students under the intervention with students-selected goals was 74.1% over the performance of the baseline phase of 40.6%. A similar conclusion appeared in the treatment with teacher-selected goals, but the improvement between baseline phases and treatment phases were slightly inferior to that in the treatment with student-selected goals. Also, the interventions disregarding students-selected goals or teacher-selected goals improved academic achievement and arithmetic homework accuracy and there was little difference between the two conditions of goals selection.

Cushing and Kennedy (1997) investigated whether peer support strategies have a positive or negative academic effect on adolescents without disabilities who served as peer supports for students with disabilities by using single-subject design. Their results indicated that all participants providing peer support improved their academic engagement.

Reinforcement

In conducting a single-subject behavior analysis, Noell, et al. (1998) found that the intervention combining modeling and practice with contingent reward was more effective than that of contingent reward only on oral reading fluency for three participants with ADHD (Attention Deficit Hyperactivity Disorder). Lloyd, Eberhardt, and Drake (1996) used a reversal design to examine the relative effects of individual and group-oriented reinforcement contingencies on academic performance within a cooperative learning context. Their results showed that group-oriented rewards produced higher mean Spanish quiz scores than did individual contingencies under group study.

Gillat and Sulzer-Azaroff (1994) explored the relationship between principals' involvement in instructional processes and students' academic performance. In the first experiment, results showed that all students improved their performance on quizzes on multiplication problems while a trained elementary school principle regularly visited the students' classroom and also set goals and gave feedback and praise to students. In a second experiment, they also found all students in two different classes increased the number of pages reported read when a middle school principle was involved.

Self-Control Training

Stevenson and Fantuzzo (1986) inspected the effect, generality and social validation of a self-control training package. Under the context of a reversal design, they found that three fifth-grade underachieving students improved their arithmetic performances while they accepted the self-control intervention. In addition, generalization effects and social validity of self-control intervention were confirmed.

Harris (1986) evaluated the relative contribution of self-monitoring of attention and self-monitoring of productivity on on-task behavior and academic productivity of four learning disabled children with attentional problems by using a multiple baseline across subjects design with a counterbalance to two kinds of self-monitoring intervention. Results indicated that mean on-task scores increased to 91%, 77%, 90%, and 91% during the self-monitoring of attention phase and 87%, 75%, 89%, and 98% during the self-monitoring of productivity phase from 57% and 32%, 44%, and 52% in baseline for four treated children respectively. Additionally, mean academic productivity scores also increased regardless of self-monitoring interventions. Thus Harris concluded that both self-monitoring of attention and self-monitoring of productivity were effective. However, it is difficult to prove which aspect is better than another due to the limitation of the research in lacking a common frame of reference to compare the two different self-monitoring interventions.

Roberts, Nelson, and Olson (1987) evaluated the effectiveness of self-instructional training with reinforcement in the use of self-instruction only, accuracy only, or both. This investigation documented that self-instruction intervention independent of any reward contingent condition produced a substantial improvement in the accuracy of arithmetic problem solving for all treated subjects.

Dunlap and Dunlap (1989) applied a self-monitoring package including self-monitoring, reinforcement, and feedback under a multiple baseline across students design to evaluate whether it could promote three learning disabled students' correct responses for subtraction problems. Their results showed that the academic performances of all students during intervention sessions were superior to that during preceding baseline sessions.

Lloyd, Bateman, Landrum, and Hallahan (1989) examined the relative contribution of self-recording of attentive behavior and academic productivity on arithmetic performance and attention to task with five special students. The results showed that the self-recording procedures ameliorated both arithmetic performance and on-task behavior, but it was not clear if there could be discrimination as to which type of self-recording was more effective.

Strategy Instruction

Montague (1992) employed a multiple baseline design to examine the effects of cognitive and metacognitive strategy instruction for mathematical problem solving. The result suggested that cognitive and metacognitive strategy instruction can profit middle school participants with learning disabilities.

O’Shaughnessy and Swanson (2000) indicated that phonological awareness training and word analogy training are both effective ways to improve oral reading fluency. Also, several investigations (Babyak, Koorland, & Mathes, 2000; Gardill & Jitendra, 1999; Gurney, Gersten, Dimino, & Carnine, 1990; Idol, 1987; Newby, Caldwell, & Recht, 1989) demonstrated that a story mapping technique can improve reading comprehension.

Knapczyk (1989) used a question-asking training incorporating videotaped exemplars, practice, and feedback in a resource room to evaluate its influence on three learning disabled students’ question asking behaviors and performances for seatwork assignments in a regular mathematics class. Results demonstrated that the level of question-asking and the accuracy scores on assignments were raised for all participants while they accepted the training. Knapczyk concluded that there was a positive relationship between question-asking skill and academic performance, in that students could improve academic attainment by using the strategy of question-asking appropriately to acquire comprehensive understanding and feedback.

Summary of 2.1

As reviewed above, considerable literature regarding academic behavior change has been accumulated and several educational intervention strategies have been proposed.

But individual investigations are too specific to determine whether particular treatments are consistently effective across different conditions. For this reason, it may be useful to aggregate and compare the results, and a quantitative research synthesis may contribute to a development of understanding in this domain.

2.2 Methods for Analyzing Single-Subject Outcomes

In the search for a suitable method to analyze outcomes of single-subject researches and provide a metric of effect sizes, this present section focuses on methodology. Several common procedures used in analyzing single-subject data can be described as follows:

Visual Analysis

Visual inspection has long been proposed as the foremost method for evaluating intervention effects on target behaviors within single-subject researches. Graphic presentation provides concise and comprehensive information about a continuous development and a causal relationship and also visual inspection possesses the advantage of assessing data of individuals and small groups immediately (Tawney & Gast, 1984; Parsonson & Baer, 1992).

While interpreting graphed data from a single-subject design, three popular principles are utilized as an explanatory basis (Franklin, Gorman, Beasley, & Allison, 1996). First, central location is utilized within phases and its shift between phases is noted to judge whether treatment effects have occurred. Then, to account for variability, a degree of departures from central location and switches of the deviation over time should be considered as they may affect accuracy of a judgment and confidence of an inference. If levels of variability within a phase decrease, it signifies that an observed series of data has been stable. In addition, the larger the variability between differential phases, the easier it may be determined. Finally, trend symbolizes how data transfer over time. Many raters consider trend as a guide. But it is difficult to discriminate an angle that represents trend changes accurately (Parsonson & Baer, 1992). Methods for superimposing trend lines such as least-squares regression, split-middle method and so on have been proposed for judgmental aids (Franklin, Gorman, Beasley, & Allison, 1996).

However, the reliability of visual inspection is unclear as interpretations of results are always not equivalent among different evaluators. Jones, Weinrott, and Vaught (1978) found low inter-judge agreement for visual analysis. Park, Marascuilo, and Gaylord-Ross (1990) indicated that the magnitude of agreement among judges in their study was not large and just somewhat better than chance.

In reviewing some results of examinations on how mean shift influences inter-judge agreement, Parsonson and Baer (1992) indicated that graphing technique becomes more debatable under a condition of decreasing levels of mean shift. But they commented that a detection of mean shift in such studies is not always clear as it is often confounded with serial dependency and changes in variability and trend. By the same token, changes in variability and trend are also associated with inconsistent interpretations. Therefore, the three interpretive aspects, central location, variability, and trend as mentioned above should be used together when interpreting a graph display.

In light of the above, some commentators advocate the use of statistical procedures that can provide objective and consistent results to supplement visual analysis in evaluating single-subject data (Jones, Weinrott, & Vaught, 1978; Park, Marascuilo, & Gaylord-Ross, 1990; Richards, Taylor, & Ramasamy, 1997). But there is still a challenge in relation to poor agreement between inferences of visual analysis and statistical procedures. For example, Jones, et al., (1978) showed that uniformity between illations of visual inspection and time-series analysis was not much greater than chance. Ottenbacher (1990) asked 30 raters to evaluate whether a significant change occurred across two phases of a set of 24 stimulus graphs of an AB single subject design and compared conclusions of visual inference with that of the split-middle method of trend estimation. The results of his study also revealed a weak agreement on the outcomes of two procedures. Richards, Taylor, and Ramasamy (1997) requested 42 undergraduate and 20 graduate special education pre-service and in-service teachers to evaluate 24 graphs which were the same as those used in Ottenbacher's investigation (1990). The researchers also contrasted visual analysis with statistical determination using the split-middle method of trend estimation and found that group means in scoring by visual inspection were inaccurate when referred to the results of the split-middle method.

In contrast, Park, Marascuilo, and Gaylord-Ross (1990) claimed a considerable agreement between visual inspection and randomization test at 80%. However, it is difficult to conclude that visual analysis is more congruent with a randomization test than other statistical procedures for two reasons. First, most of the agreement (67%) was on non-significant experimental effects. Second, the analysis of consistency was only based on a small sub-sample of graphs (n=15) that had sufficient data points (at least 25) on which to conduct a randomization test.

Furthermore, some researchers have indicated that experienced visual analysts are more likely to generate conservative judgments and commit Type II errors than those adopting statistical procedures (Ottenbacher, 1990; Parsonson & Baer, 1992). Matyas and Greenwood (1990) commented that previous empirical studies on one issue of false alarm (Type I errors) and miss rates (Type II errors) in visual analysis were not appropriate because the performance of visual analysis was compared with that of statistical procedures as an absolute criterion. In the study by Matyas and Greenwood, three factors (the effect of intervention, the amount of random error, and the degree of serial dependency) were manipulated on 27 charts and 37 beginning practitioners were asked to respond as to whether intervention effects existed. The results showed high false alarm rates (Type I errors) especially under a condition where positive autocorrelation and soaring random variation were integrated, and also relatively low miss rates (Type II errors).

Hence, a classical visual inspection approach cannot provide a fair and unifying metric for single-subject investigations at aggregated level.

Statistical Analysis

In search of an objective and appropriate outcome metric, several methods of statistical analysis have been proposed. Gentile, Roden, and Klein (1972) suggested an analysis of variance (ANOVA) model to deal with single-subject data by analogy with those in multiple-subject experimental designs. But adopting this method results in a serious violation of the assumptions of parametric statistical tests, especially in the requirement for independent observation (Crosbie, 1993).

With regard to the existence of serial dependency (autocorrelation) in single-subject time series data, a question arises in relation to an empirical work by Huitema in 1985 (Matyas & Greenwood, 1996). In view of Huitema's conclusion that serial dependency in single-subject time series data is not an acute issue, Center, et al. (1985-86) advocated a piecewise regression technique that allows for changes in level, changes in slope, and joint effects of both to obtain an index of effect size. They compared the effect size from a piecewise regression model with those from an ANOVA model considering the mean only and claimed that treatment effects were overestimated by the ANOVA model (the mean effect size was 3.2) in contrast with the piecewise regression model (the mean effect size for level was 1.66 and for the shared effects of changes in level and slope was 1.88).

A further criticism regarding Huitema's opinion was that the test for autocorrelation lacks power because of the brief time series (Matyas & Greenwood, 1996). In other words, it is untenable to accept the null hypothesis of no serial dependency in single-subject time series data. For this reason, we may conclude that the use of a parametric statistical procedure such as a regression analysis is not applicable to evaluate treatment effects within single-subject designs.

As comment on the use of a regression model on estimation of effect sizes, first, capricious estimates may be produced from a regression model dealing with a time series with few points. In addition, the regression approach assumes an equal interval of time passage across data points. But individual data points are often not equally spaced (Scruggs & Mastropieri, 1998).

In addition to the methods mentioned above, other major means for interpreting intervention effect in single-subject designs are reviewed as below.

Standardized Mean Difference Approaches

An index of effect size using mean difference of the treatment and baseline phase divided by the standard deviation of the baseline phase or the pooled standard deviation was employed to describe the magnitude of treatment effects in certain single-subject researches (Busk & Serlin, 1992; Busse, Kratochwill, & Elliott, 1995; Kromrey & Foster-Johnson, 1996; White, Rusch, Kazdin, & Hartmann, 1989).

The formula is as follows:

$$d = \frac{\overline{X}_{treatment} - \overline{X}_{baseline}}{SD_{baseline}}$$

$$\text{or} \quad d = \frac{\overline{X}_{treatment} - \overline{X}_{baseline}}{SD_{pooled}}$$

Advantages of standardized mean difference approaches are that they can generate a value representing the strength of the treatment more than the conclusion if the treatment is significant (Kromrey & Foster-Johnson, 1996) and their procedures are similar to those used in multiple-subjects researches so calculation procedures are relatively straight-forward (Faith, Allison, & Gorman, 1996). Swanson and Sachse-Lee (2000) slightly modified this approach by exerting difference between the averaged score of the last three sessions in baseline phase and in treatment phase divided by the pooled standard deviation comprised of both last three sessions of baseline and treatment as an index of effect size and conducted a meta-analysis of single-subject studies on students with learning disabilities.

However, there are several drawbacks to standardized mean difference approaches. First, it is difficult to generate a meaningful and comparable estimate of effect size while there is no variability (Idleman, 1993; Scruggs, Mastropieri, & Casto, 1987; White, Rusch, Kazdin, & Hartmann, 1989). Then, the effect size is also unreliable while the scarcity of data points often appears in each phase (Idleman, 1993; Mastropieri, & Casto, 1987; Scruggs & Mastropieri, 1998). Third, Faith, et al. (1996) have mentioned that standardized mean difference approaches only focus on levels but neglect trends. A more fundamental issue is that non-independence of effect sizes may discredit the inferential results (White, Rusch, Kazdin, & Hartmann, 1989). Finally, more effort is needed to compute effect sizes owing to accuracy in data gained by recovering data points displayed in a graph (White, Rusch, Kazdin, & Hartmann, 1989).

Autoregressive Integrated Moving Average (ARIMA) Approach

According to given ample time-series data, the Box-Jenkins ARIMA approach can build up equations that include the components of autoregressive terms, moving average terms, and differencing values. Further, one can base on the number of differencing and the patterns of autocorrelation and partial autocorrelation functions to gain appropriate parameters for identifying an ARIMA model fitting data. The formula for the ARIMA model is (Box & Jenkins, 1976):

$$\phi(B)\nabla^{d}Z_{t} = \theta(B)a_{t}$$

In the formula, $\phi(B)$ signifies autoregressive procedures, the ”AR” model for short, $\nabla^{d}Z_{t}$ signifies differencing values of a series of observations, and $\theta(B)a_{t}$ signifies moving average procedures, namely, the “MA” model.

Applying the ARIMA approach to assess the intervention effects in single-subject studies can account for changes in level and trend as well as autocorrelation. And through computing and controlling the autocorrelation, time-series analysis techniques such as interrupted time-series analysis (ITSA) procedures were developed to evaluate whether there are significant differences between pre-intervention and post-intervention with single-subject data (Crosbie, 1993).

However, this approach is constrained by the requirement of a large number of data points, a minimum of 50 observations per phase being necessary to obtain accurate estimates for model identification (Box & Jenkins, 1976). According to Huitema, there are typically four data points per baseline phase (Methot, 1995). This implies that applying the ARIMA approach to single-subject behavioral data is not appropriate.

Simplified Time Series Analysis (The C Statistic)

In contrast to the ARIMA approach, the requirement of data points in the use of the C Statistic is more flexible. This method can be used on small data sets to determine whether or not the time series contains any trends (Tryon, 1982).

The formula is as follows:

$$C = 1 - \frac{\sum_{i=1}^{N-1} (X_i - X_{i+1})^2}{2\sum_{i=1}^{N} (X_i - \overline{X})^2}$$

And the standard error of the C statistic is:

$$S_c = \sqrt{\frac{N+2}{(N-1)(N+1)}}$$

While evaluating the treatment effects, the C statistic of the first baseline phase is generated and divided by its standard error initially to judge whether or not the baseline data contain any statistically significant trends. It is expected that no significant trends appear in the baseline phase. Then, the data of the treatment phase is appended to the baseline data. And the data from baseline plus treatment phases is tested for any trends by using the C statistic again. And the C Statistic is also divided by its standard error to test the significance. A statistical significant shift in the trend indicates that an intervention is effective.

Although the C statistic is a simple way to evaluate intervention effects, it can not be calculated directly from graphic displays. In other words, the calculation of the C statistic must be based on exact values.

However, the C Statistic may not reflect the intervention effects while there is no obvious trend in the treatment phase, e.g., ceiling effects or floor effects and only an abrupt trend change between the final data point in the baseline phase and the first data point in the treatment phase.

Randomization Test

Randomization tests were employed for investigating individual data by Levin, Marascuilo, and Hubert (1978), Wampold and Worsham (1986), and Edgington (1995). Each study utilized an original or ranked difference between means for each of the phases within a single-subject design as a statistic to test various hypotheses.

In contrast with parametric statistical tests, randomization tests dispense with assumptions of normality and homogeneity of variance (Marascuilo & Busk, 1988). According Levin, et al. (1978) and Edgington (1995), random assignment is an essential component for randomization tests. In studies, random assignment of subjects to treatments was replaced by treatment sequence.

But even under the foregoing scheme, time-related effects still confounded treatment effects and carryover effects prevailed. For these issues, Levin, et al. (1978) noted that increasing AB phases (e.g., the ABAB design) and systematic assignment could reduce criticisms of internal validity. However, reduction in the degree of severity of these problems does not provide a resolution for the preceding challenges. Also, systematic assignment violates the uncorrelated errors assumption. This problem not only occurs in a reversal design, in a multiple baseline design because it is impossible to assign subjects at random to various treatment orders in a multiple baseline design.

Consequently, randomization tests are neither applicable for analyzing and summarizing intervention effects of single-subject researches.

Percentage of Nonoverlapping Data (PND)

Scrugg, Mastropieri, and Casto (1987) recommended a nonparametric method based on computation of percentage of nonoverlapping data (PND) between baseline and treatment phase in single-subject data displays to aggregate relevance researches.

The PND score is figured as follows: first, draw a line that is parallel to the abscissa through the highest or the lowest data point in a baseline phase contingent on the expected direction of treatment effect. Second, extend the line through the subsequent treatment phase. Third, calculate the number of data points in an adjacent treatment phase above the highest or below the lowest data point in the preceding baseline phase, divide it by total number of data points in the treatment phases and then multiply it by 100%.

As the PND score is a percent score between 0 and 100, some interpretative criteria have been proposed (Scruggs & Mastropieri, 1994) as Table 1.

TABLE 1

Interpretative Criteria for the PND Score

Range of PND scores	Interpretation of Treatment Effectiveness
Above 90%	Extreme
70%-90%	Moderate
50%-70%	Slight
Below 50%	Ineffective

Thus, an advantage of the PND method is its ease of interpretation. Also, the PND technique is simple to execute without complex computations based on precise raw values for data points. A further feature of the PND method is that it is exempted from meeting the prerequisites of parametric statistical approaches such as normality and heterogeneity of variance.

There are several studies utilizing the PND score as a metric of effect size to synthesize single-subject researches. For example, Mastropieri and Scruggs (1985-86) explored the effectiveness of early intervention with socially withdrawn preschoolers, Scruggs, Mastropieri, Cook, and Escobar (1986) reviewed sixteen investigations to evaluate the treatment for children with conduct disorders, and Scruggs, Forness, and Kavale (1988) synthesized twenty studies on early language interventions. Further, more recently Mathur, Kavale, Quinn, Forness, and Rutherford (1998) summarized findings of sixty-four researches to examine the influence of social skills interventions with students with diagnosed emotional and behavioral disorders and Browder and Xin (1998) extended the application of the PND approach to assess the effectiveness of sight word instruction for students with moderate and severe disabilities.

However, some weaknesses of the use of the PND statistic have been raised (Scruggs, Mastropieri, & Casto, 1987; Strain, Kohler, & Gresham, 1998; White, 1987). These comments are summarized as follows:

1. Trend is not taken into account in the calculation of the PND.
2. The PND method may lower power for discriminating the different performance patterns that produce the equivalent PND score. As demonstrated in Figure 1, as created for this study in example, the PND scores of the left and the right are both

100%, but the performance patterns of the interventions are very different.

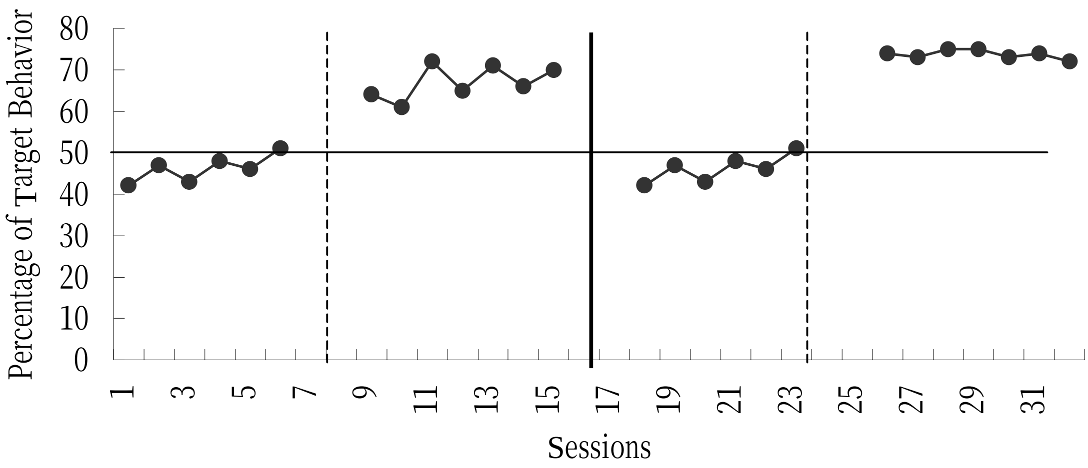

Figure 1. Example of divergent performance patterns that produce an equivalent PND score.

3. The PND may be unduly sensitive to outliers as it refers only to the most extreme data points on the distribution of a baseline phase. As shown in Figure 2, a treatment has a positive effects on target behaviors visually, but the PND score is 0% owing to the influence of an outlier.
4. In a reversal design, the PND might be inappropriate if the second baseline phase had an extinction trend and the second treatment phase had arising trend. In this case, it would constitute a shape of orthogonal slopes, then the PND scores would be underestimated (see Figure 3).

White (1987) suggested a nonparametric method of modifying and integrating Shewart's control chart techniques and the PND procedures to avoid the influence of outliers on the PND statistic. In order to calculate effect sizes, the highest 5% of data points in the baseline phase need to be removed, and the next highest point is obtained as a point of reference. However, the limitation of this method is that at least 20 data points in the baseline phase are required to calculate effect sizes. Besides, the highest 5% of data points are not necessary outliers.

Percentage of Data Points Exceeding the Median of Preceding Baseline Phase (PEM)

Ma (2002) proposes a method, namely the PEM (percentage of data points exceeding the median of preceding baseline phase) technique, to deal with some of the difficulties that are encountered in using the PND method. The first procedure of the PEM method is to drag a line horizontally though the median of the baseline phase and across the treatment phase. Then, there is a calculation of the proportion of data points in the treatment over or under the median of the baseline hinged in an anticipated direction. The gained PEM score lies in between 0% to 100%.

The PEM technique not only has advantages similar to that of the PND method, but also overcomes some limitations of the PND method. First, the PND may not reflect the treatment effect while outliers appear in the baseline. However, the PEM method is not influenced by outliers. As shown in Figure 2, a treatment has a positive effect on a target behavior. The PND score is 0% owing to the influence of an outlier (see the solid horizontal line) while the PEM score is 100% (see the dotted horizontal line).

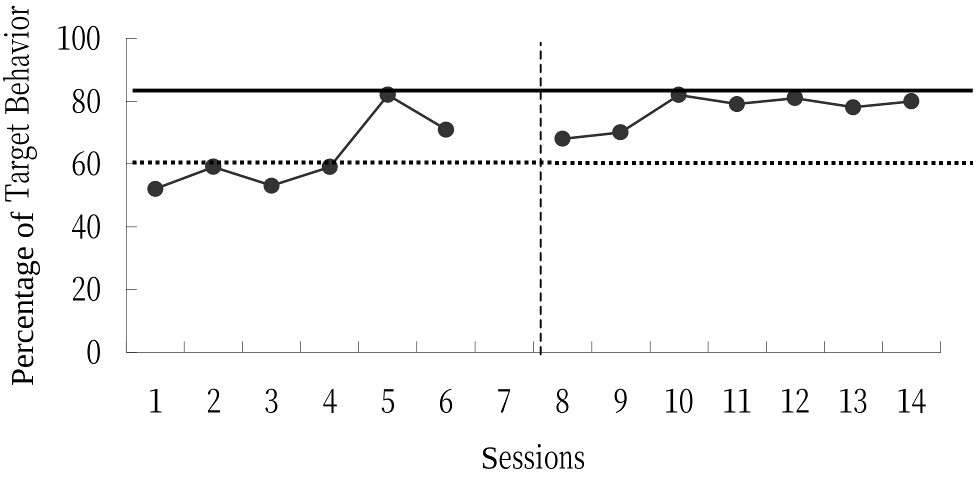

Figure 2. Example of outliers.

Second, when using the PND method, an orthogonal slope change in a second pair of baseline-treatment phases would cause an underestimation of the effect in the second treatment phase. To seek to overcome this shortcoming, Ma (2002) claims that the PEM method can improve by halves. As illustrated in Figure 3, an orthogonal slope change appears in the second pair of baseline-treatment phases. The PND score of the second treatment phase is 28.57% (2 out of 7 data points above the solid horizontal line). And the PEM score of the second treatment phase is 71.43% (5 out of 7 data points above the dotted horizontal line). In other words, the PEM method can decrease the influence of an orthogonal slope change more than the PND method to show the effect in a second treatment phase.

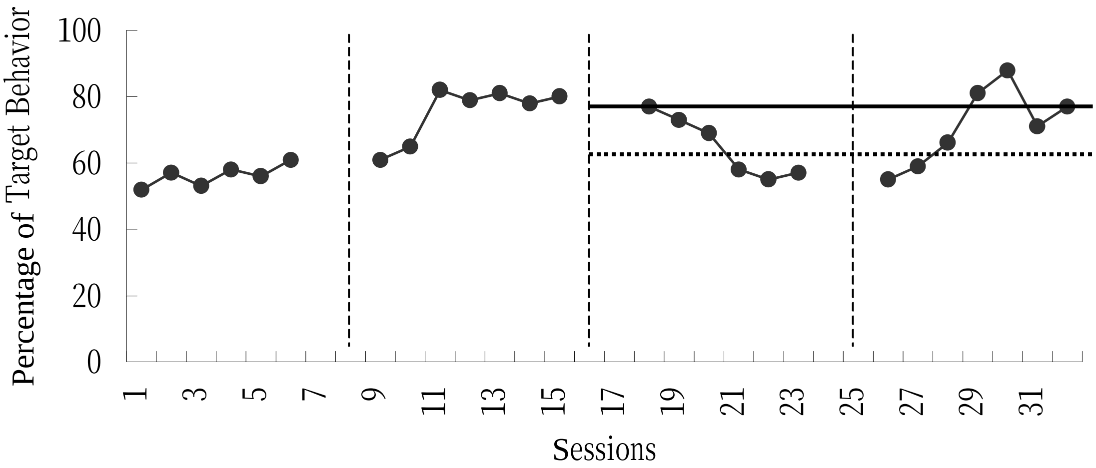

Figure 3. Example of orthogonal slope change.

Summary of 2.2

In the review mentioned previously, the nonparametric procedures of the PEM and the PND approach are regarded as optimal ways for synthesizing the results of single-subject researches. In addition, the PEM method overcomes some limitations of the PND method, and there should be an evaluation of whether or not the PEM method is practicable.