School Psychology Quarterly
2010, Vol. 25, No. 1, 28–44

© 2010 American Psychological Association
1045-3830/10/$12.00 DOI: 10.1037/a0018600

Visual Analysis of Multiple Baseline Across Participants Graphs When Change Is Delayed

Rebecca G. Lieberman, Paul J. Yoder, Brian Reichow, and Mark Wolery
Vanderbilt University

A within-subjects group experimental design was used to test whether three manipulated characteristics of multiple baseline across participants (MBL-P) data showing at least a month delayed change in slope affected experts' inference of a functional relation and agreement on this judgment. Thirty-six experts completed a survey composed of 16 MBL-P graphs. Graphs with steep slopes, once change began, were far more likely to be judged as showing a functional relation. Generally, experts disagreed with each other regarding functional relation judgments. Implications for the types of dependent variables that fit the requirements of MBL-P are discussed.

Keywords: multiple baseline across participants, visual analysis

Single-subject research designs comprise a variety of experimental methods that play a key role in establishing evidence-based practice in education (Horner et al., 2005; Shernoff, Kratochwill, & Stoiber, 2002). Through the use of single-subject methods, researchers can systematically and rigorously build evidence of the efficacy of treatments at the individual or small-group level. The identification of functional relations between independent and dependent variables in single-subject studies with high internal validity contributes to the growing body of knowledge regarding educational practices that are efficacious for individuals with a variety of developmental needs. Indeed, the ability to distinguish between change during treatment and change due to treatment is a key strength of this experimental methodology.

It has been proposed that single-subject methods could be an integral part of the Response to Intervention framework (RTI) (Barnett, Daly, Jones, & Lentz, 2004). Single-subject studies may be informative as follow-up to larger group studies or to implementation of school- or classroom-wide tier 1 interventions. In such a case, single-subject methods may be used to examine on a small scale whether tier 2 or tier 3 interventions are effective for students who do not change when exposed to a tier 1 intervention. The use of single-subject methodology would allow a school psychologist to determine whether potential change in the dependent variable was actually because of implementation of the tier 2 or tier 3 intervention. In addition, the single-case design provides an empirical basis for adjusting interventions at each tier of the RTI framework, including more specific interventions at tier 2, and particularly more intensive interventions at tier 3 (Daly et al., 2007; Riley-Tillman & Burns, 2009).

Because interventions become more focused and intensive at tiers 2 and 3, assessment data must be collected more frequently and precisely. At tier 2, data are collected weekly to monitor students' response to intervention, whereas at tier 3, data are collected several times a week (Riley-Tillman & Burns, 2009). The experimental control of the single-case de-

Rebecca G. Lieberman, Paul J. Yoder, Brian Reichow, and Mark Wolery, Department of Special Education, Vanderbilt University, Nashville, Tennessee.

Brian Reichow is now at the Yale Child Study Center, Yale University School of Medicine, New Haven, Connecticut.

This research was supported in part by grant T32HD07226 from the National Institute of Child Health and Human Development to Vanderbilt University (to R.G.L.).

We acknowledge Jon Tapp for assistance with agreement estimation and Jordan Yoder for creating early versions of the visual stimuli used in our survey.

Correspondence concerning this article should be addressed to Rebecca G. Lieberman, Department of Special, Education, Peabody College, Box 228, Vanderbilt University, Nashville, TN 37203. E-mail: rebecca.g.lieberman@vanderbilt.edu (until 2011). After 2011, correspondence may be sent to Paul Yoder, Department of Special, Education, Peabody College, Box 228, Vanderbilt University, Nashville, TN 37203. E-mail: paul.yoder@vanderbilt.edu

sign allows practitioners to establish through assessment whether students' behavior is changing in response to the treatment as opposed to some extraneous factor (Riley-Tillman & Burns, 2009). If it can be demonstrated through repeated measurement of the individual performance that change in behavior has not occurred as a result of treatment, informed decisions can then be made to alter intervention in some way, that is, increase intensity, adjust some component of treatment, or move the student to the next tier. Because change in behavior is examined at the individual level in single-subject experiments, use of this methodology to make programmatic decisions within RTI fits well with the original logic of single-subject design and may be particularly relevant to the classroom setting (Daly et al., 2007). The multiple baseline design across participants is a choice for such studies, because it does not require a withdrawal or reversal of the intervention. Integral to both establishment of evidence-based practices and structuring of RTI decisions is the appropriate visual analysis of single-subject data.

Past empirical work has examined the analysis of visual data, predominately in withdrawal or reversal designs (A-B-A-B). General findings of these studies indicate that agreement between raters tends to be quite low. A study by Park, Marascuilo, and Gaylord-Ross (1990) examined reliability of judgments of "significant intervention effects" in published A-B design visual data. Graphs contained titles providing some contextual information to expert judges. Results showed percentage agreement between all pairs of raters was slightly more than chance levels. In addition, 16% of the graphs were categorized as nonsignificant, and 35% as "unclear," despite the fact that these graphs were previously published and judged to show significant effects by journal reviewers (Park et al., 1990). However, it is probable that some types of data yield more agreement than others. For example, DeProspero and Cohen (1979) found that immediate shifts in level were a key factor in determining the presence of a functional relation among participants. In addition, Ottenbacher (1986) found that agreement among raters of a "clinically significant" change in behavior increased when there was a large level shift in the data. Gibson and Ottenbacher (1988) reported similar findings. Clinically significant change does not necessarily mean change because of treatment (i.e., a functional relation). In this article, we requested experts to judge whether a functional relation is present not simply that clinically significant change occurs in the treatment phase.

As described earlier, several studies have been conducted examining the visual analysis of data using variants of the A-B design. However, relatively little is known about visual analysis of data generated in the multiple baseline design. The multiple baseline across participants (MBL-P) design is often used in intervention research focused on promoting developmental or academic skills. The multiple baseline design is one of the most widely used single-subject experimental designs when dependent variables are not easily reversed (Kennedy, 2005).

The degree of internal validity refers to the extent to which the design and data support an inference that the behavior change is due to the independent variable and nothing else (i.e., a functional relation; Kennedy, 2005). In a MBL-P design, a functional relation can be inferred (a) when data are stable during the baseline phase, (b) when change in the dependent variable only takes place once introduction of the independent variable has occurred, (c) when baselines in the other tiers remain stable even as the data path changes in the tier where the independent variable has been introduced, and (d) when change in behavior during the treatment phase replicates across participants. Of note, most texts do not say *when* during the treatment phase the dependent variable needs to change for a functional relation to be inferred. In addition, there is no consensus on what constitutes sufficient consistency across participants to infer a functional relation.

We have almost no empirical information on factors that influence functional relation inferences when the dependent variable changes long after the onset of treatment (Parsonson & Baer, 1992). At least some believe that the longer the latency between the onset of the treatment and the onset of change in the dependent variable, the more likely events outside of the treatment context can influence the dependent variable (Kazdin, 1981). Because the MBL-P design is sometimes used with variables that do not show immediate or rapid changes, it is necessary to understand the factors that influence judgments of functional relations when establishing an evidence base for treatments or when making program decisions within an RTI framework.

Three aspects of graphed data may influence identification of functional relations in the multiple baseline design when a delay in trend change in the dependent variable is present: (a) steepness of slope, (b) consistency in the latency of change, and (c) expectancy of delayed change. In an unpublished pilot study, graduate students used steepness of slope more than any other factor when identifying functional relations in the presence of delayed changes in the dependent variables (Yoder, Reichow, Wolery, & Tapp, 2007). Precise replication across participants of the latency between treatment phase onset and the beginning of change in the dependent variable shows a lawfulness that is generally not explainable by maturation or other environmental influences outside of treatment (Parsonson & Baer, 1992). Finally, contextual variables that lead to an expectancy that the change will be delayed can support experts in agreement regarding functional relation inferences, assuming other evidence supporting a functional relation is present (Parsonson & Baer, 1992).

The current study examined three questions regarding graph sets displaying delayed (i.e., at least 1 month after onset of the treatment) changes in the dependent variable. (a) Does steepness of slope, consistency of latency of change, and expectancy of delayed change affect experts' judgment of a functional relation? (b) Which of these three factors characterize graph set(s) that experts judge to demonstrate a functional relation with a high level of confidence? (c) To what extent do experts agree when making functional relation judgments?

This study adds to the published data on visual analysis in several ways. First, rater agreement in the MBL design is not known (Franklin, Gorman, Beasley, & Allison, 1996). Past studies have examined single A-B graphs. The stimuli of the current study were graph sets composed of 3 individual multiple baseline graphs, allowing for detecting replication of effects across participants. Second, contextual information regarding the general treatment approach, past efficacy data, participant description, measurement context, and dependent variable were provided to raters in the current study. Previous studies have provided limited contextual information that would typically accompany visual data. Third, the effects of changes in slope of the trend line on identification of a functional relation were examined (as opposed to immediate level shifts). Many academic and developmental skills are learned gradually, resulting in slope, not immediate level changes. Finally, because all graphs in this study showed change in the dependent variable during the treatment phase, the expert's task was to decide whether a functional relation, not just a replicated change, was present. Asking expert participants to identify change because of treatment (and not solely clinically significant change) expands on previous studies of the analysis of visual data.

Method

Design

A within-subjects group experimental design was used. We manipulated steepness of slope and consistency of latency of change by systematically varying these characteristics in the experimental stimuli. We attempted to manipulate expectation of whether the change would be delayed by varying participant characteristics, the dependent variable, and measurement context in graphs and contextual information given to expert judges.

Participants

An email to identify participants for the current study was distributed to the review boards of the *Journal of Behavioral Education*, the *Education and Treatment of Children*, and the *Journal of Applied Behavior Analysis*, three journals that regularly publish single-subject research data. Email addresses that were not available in the journal were found through professional affiliation information provided by the journal. If a review board member could not be located through the published affiliation, a name search was conducted through the search engine Google to find their current affiliation. If an email address could not be obtained after this final step, the reviewer was dropped from our original sample. All available reviewers were sent an email asking them to define themselves as (a) an applied scientist (someone who uses

MBL to address questions), and/or (b) a reviewer (someone who decides whether MBL studies are published). Our initial sample of expert reviewers included those who indicated they (a) used MBL-P design in their research, or (b) determined which MBL-P studies are published. The reviewers who indicated they were neither an applied scientist nor reviewed MBL studies specifically were excluded from our sample. A total of 66 reviewers were identified as experts in MBL-P design based on the earlier criteria. A second email with a link to an online survey was sent to the 66 reviewers requesting their participation in the study. Of the 66 reviewers sent this second email, 36 completed the survey, a 55% return rate. Table 1 presents information on the expert judges. All aspects of the study were approved by the investigators' Institutional Review Board.

Materials

An online survey was developed using the Internet program Survey Monkey. The survey consisted of three questionnaires designed to measure: (a) experts' rating of 16 graph sets; (b) the degree to which the dependent variable, population, and measurement context influenced each participant's expectation of a delayed change in the dependent variable; and (c) educational and professional characteristics of the experts.

The first section of the survey contained a general instruction page informing the experts that the independent variable in the graph sets was a treatment using prompts and reinforcement that had been found to be effective and efficient in other populations and contexts, and that the observational sessions illustrated in the graphs were conducted weekly. Experts were asked to rate the graphs by indicating whether they were (a) confident the data demonstrated a functional relation between the independent and dependent variables, (b) not confident the data demonstrated a functional relation between the independent and dependent variables, or (c) confident the data did not demonstrate a functional relation between the independent and dependent variables.

Table 1
Current Professional Roles and Degrees Earned

Variable	Percentage of participants
Current profession	
Full professor	45
Associate professor	14
Assistant professor	20
Program coordinator	3
Program director	6
Licensed psychologist	6
Behavior consultant	3
Lecturer/consultant	3
Highest degree earned	
EdD	17
PhD	57
PhD/MD	23
PsyD	3

A page indicating the population, measurement context, and dependent variable was placed before each figure to alert participants to key contextual information. Each graph set contained three individual graphs presented in a vertical stack with a typical MBL-P format (i.e., onset of the treatment phase was staggered across graphs and occurred only after change was demonstrated in the preceding graph [i.e., tier]). The first 4 figures were sequenced to alert raters that population and vertical axis labels would change among figures. The final 12 graph sets were randomly ordered. Each participant received the same version of the survey (i.e., the order of the graph sets was the same for each expert reviewer).

Fictitious data were used in the graphs for several reasons. First, using fictitious data created 16 graph sets showing sufficiently long and stable baselines and treatment phases, with treatment phases beginning at least four sessions after the preceding tier showed a change in the dependent variable. Second, using fictitious data allowed creation of graphs in which variability and level (the median of the last 3 points in baseline and the median of the first 3 points in the treatment phase) were constant across graphs, but the trend of the dependent variable increased during the treatment phase once the dependent variable began to change. Finally, fictitious data enabled creation of graphs in which all "participants" showed change beginning at least 4 sessions (i.e., 1 month) after the onset of the treatment.

Two characteristics of the dependent variable in the graphs varied among the 16 graph sets: (a) steepness of the trend, once change in the dependent variable occurred; and (b) consistency in the latency of change across “participants” within a graph set. A steep trend change was a two-point-per session gain. A shallow trend was a one-point-per session gain. In consistent latency of change graphs, all 3 “participants” data changed the same number of sessions from the onset of the treatment phase. In inconsistent latency of change graphs, the number of sessions between onset of the treatment phase and onset of the change varied by at least 4 sessions across tiers within the graph set (range of latency difference within inconsistent graph set was 8–20 data points).

Four graph sets were created for each of the 4 design cells that resulted from crossing the 2 levels of steepness with 2 levels of consistency of latency. The four graph sets within a design cell varied on metric of the dependent variable, the numerical pattern used to generate the nonchanging aspects of the data, and the position on the ordinate of the first data point in baseline condition. These three variables were considered irrelevant variables when making inferences of functional relations. Figures 1–4 provide an example graph set for the four design cells.

To attempt to manipulate judges’ expectancy of whether a dependent variable would change many sessions after the onset of the treatment phase, we altered the dependent variable, population, and measurement context concurrently among graph sets. These were indicated on the *y*-axis and on the page preceding each figure. To lead the raters to expect the dependent variable would show a rapid change after the onset of the treatment, the dependent variable was in-seat behavior, the population was typically developing 6-year-olds, and the measurement context was the treatment setting. To lead the raters to expect the dependent variable would show a delayed change after the onset of the treatment, the dependent variable was new words read, the population was 6-year-olds with severe intellectual disabilities, and the measurement context was a generalization setting. Within the four design cells, there were two graph sets designed to lead experts to expect an immediate change in the dependent variable and two graph sets designed to lead experts to expect a delayed change in the dependent variable.

The second section of the survey comprised a questionnaire to identify whether experts used the principles we manipulated to formulate their expectations of whether there would be an immediate or delayed change in the dependent variable. In the questionnaire, participants indicated their expectation of an immediate or delayed change in the dependent variable once onset of treatment occurred based on the described population, behavior, treatment setting, and measurement setting. The participants could indicate “don’t know” if they were not sure whether an immediate or delayed change would occur. The populations described in the questionnaire were 6-year-old children with typical development or with severe intellectual disabilities. Behaviors described were simple behavior targets or complex academic tasks. Treatment and measurement settings were either the same or different. The questionnaire could only be completed after rating the 16 graph sets to ensure the questions in this second section of the survey would not influence how the experts rated the 16 graph sets.

In the final section of the survey, participants completed a brief descriptive questionnaire to indicate their highest degree earned and current professional role. This was used to describe the experts (see Table 1). Reviewers were informed that only responses to the first section containing graph sets were required for their participation. Returned surveys without completed responses to all graph sets were excluded from all analyses. Some participants chose not to respond to the expectancy questionnaire (Section 2). These reviewers were, therefore, not included in any analyses examining the expectancy condition.

Procedure

Instructions on how to access the link to the survey to participate in the study were provided in an email, as well as contact information if problems occurred. The participants were informed they could end participation in the study at any time, and were given a deadline of 2 weeks to submit their surveys. This deadline was extended for several participants who contacted the first author directly. The participants were told that the estimated completion time for the survey was 20 minutes. There were no time constraints placed on participants to complete the survey. Judges were instructed to log on to the survey only once. Investigators remained

(*text continues on page 37*)

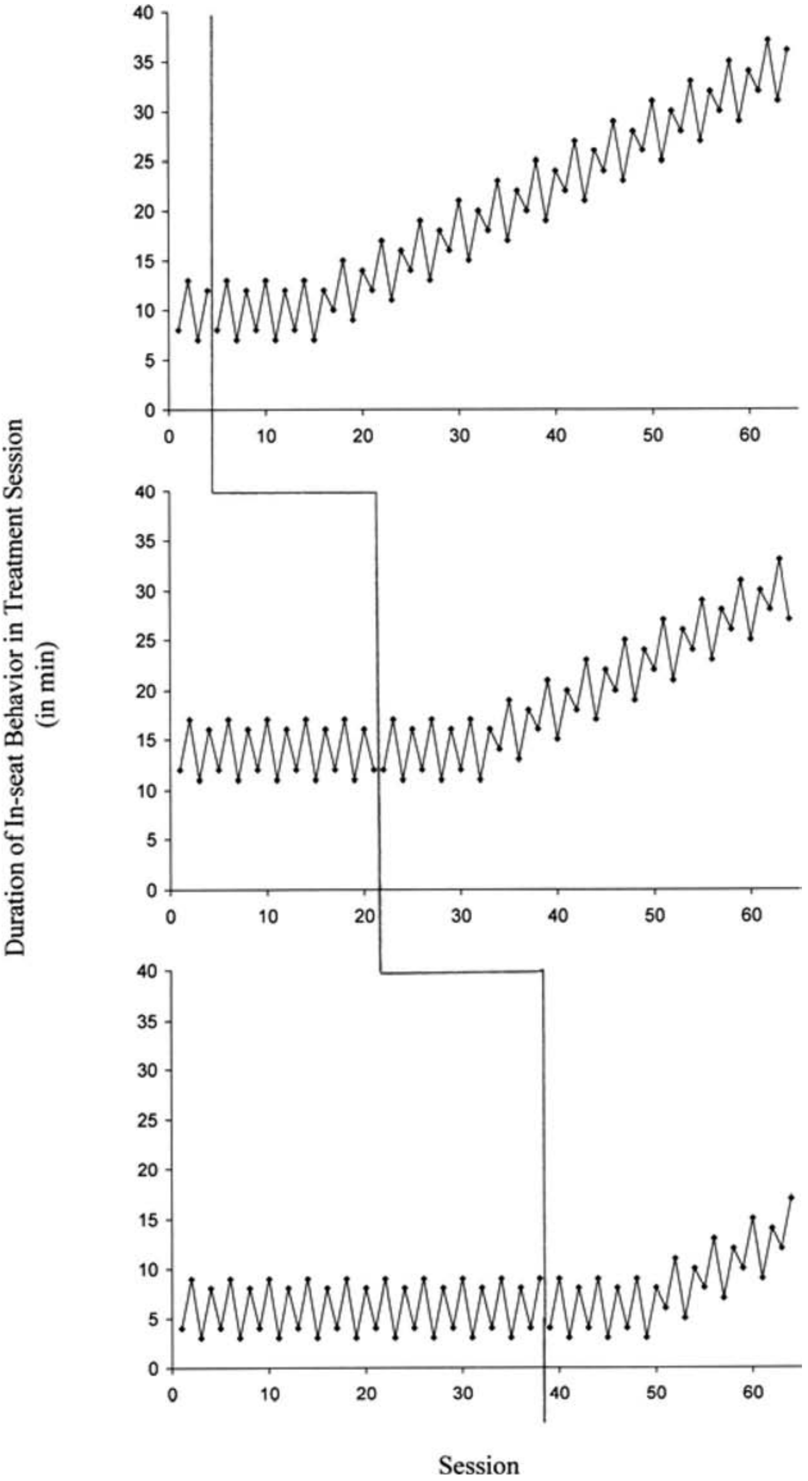

Figure 1. Example of a graph set exhibiting steep slope with consistent latency of change.

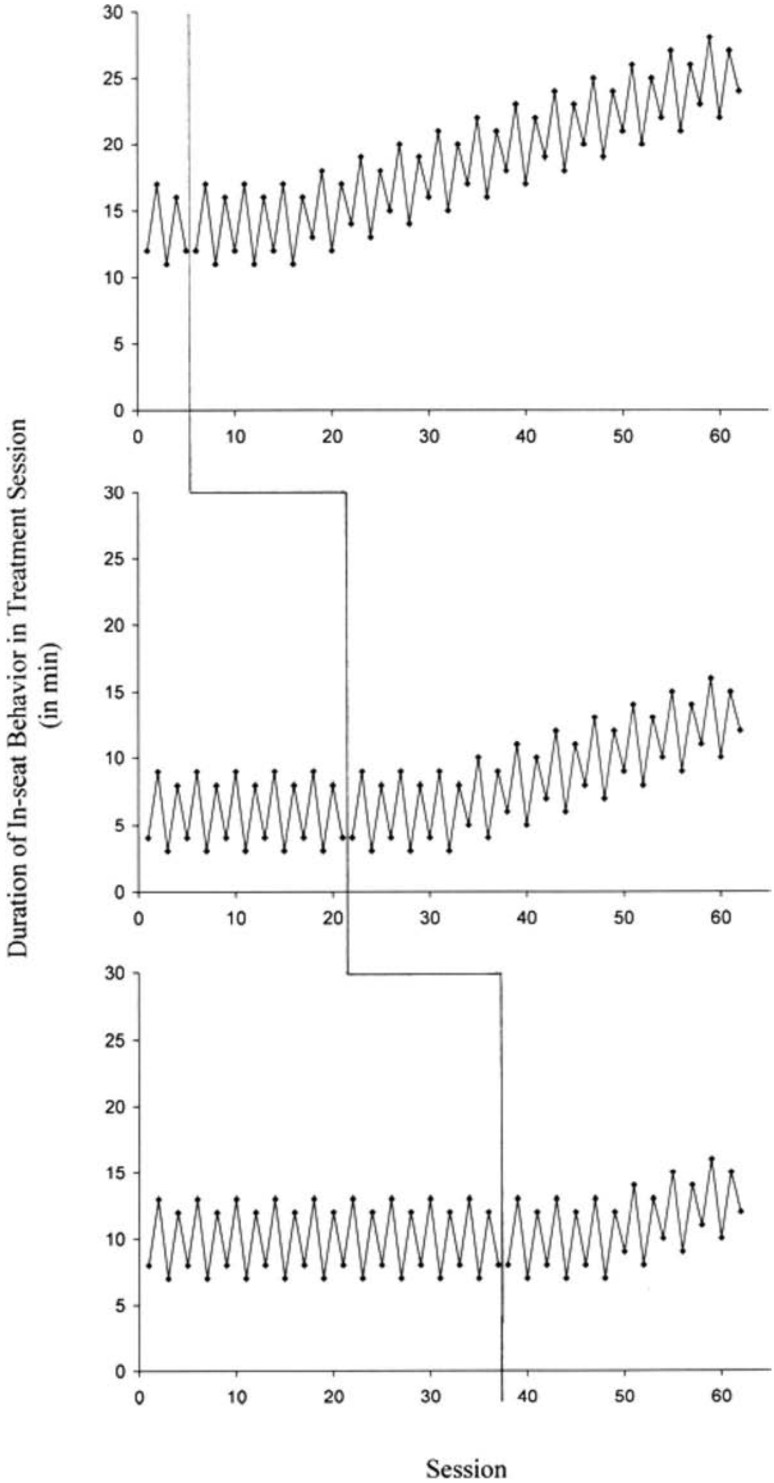

Figure 2. Example of a graph set exhibiting shallow slope with consistent latency of change.

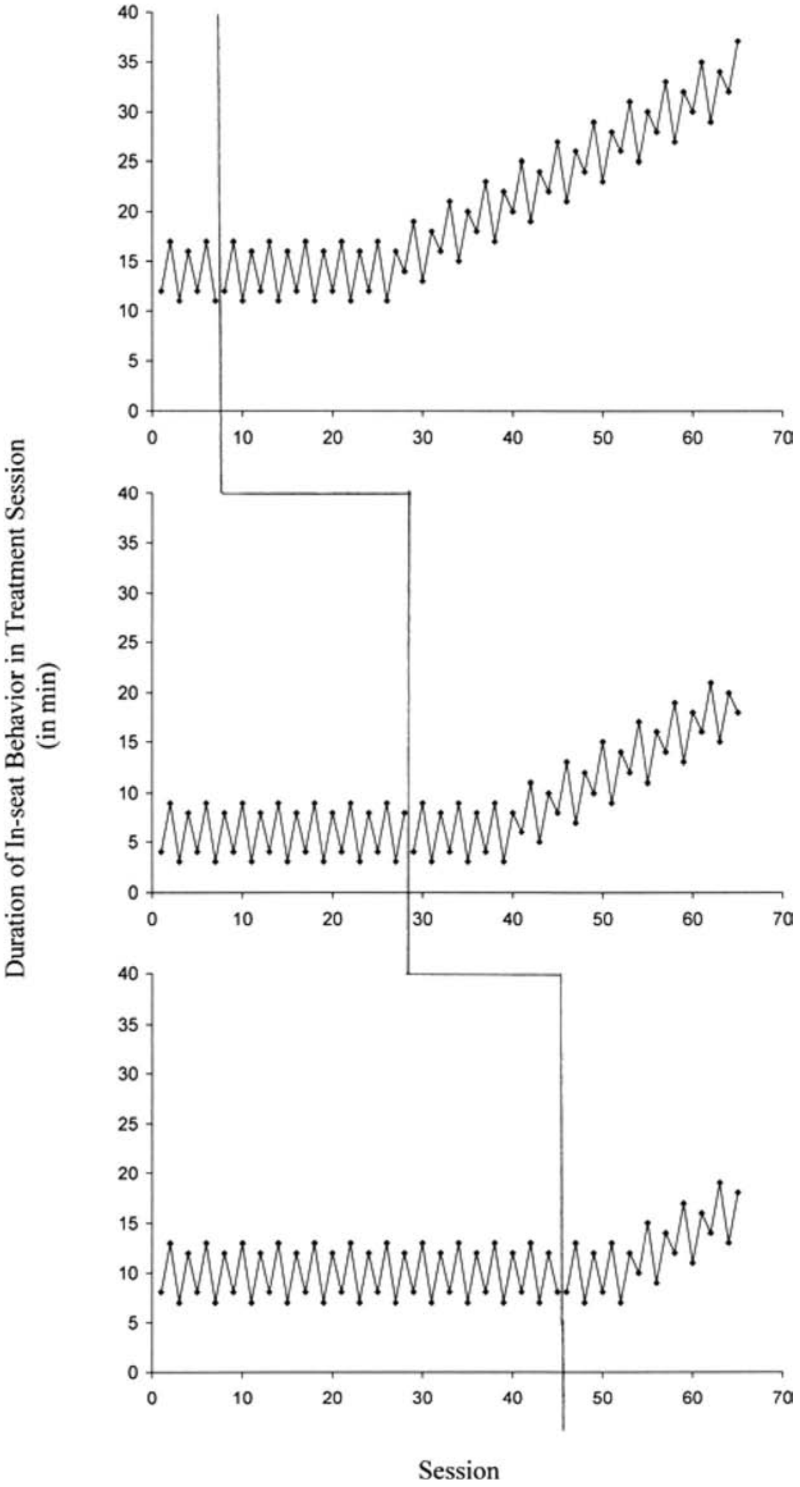

Figure 3. Example of a graph set exhibiting steep slope with inconsistent latency of change.

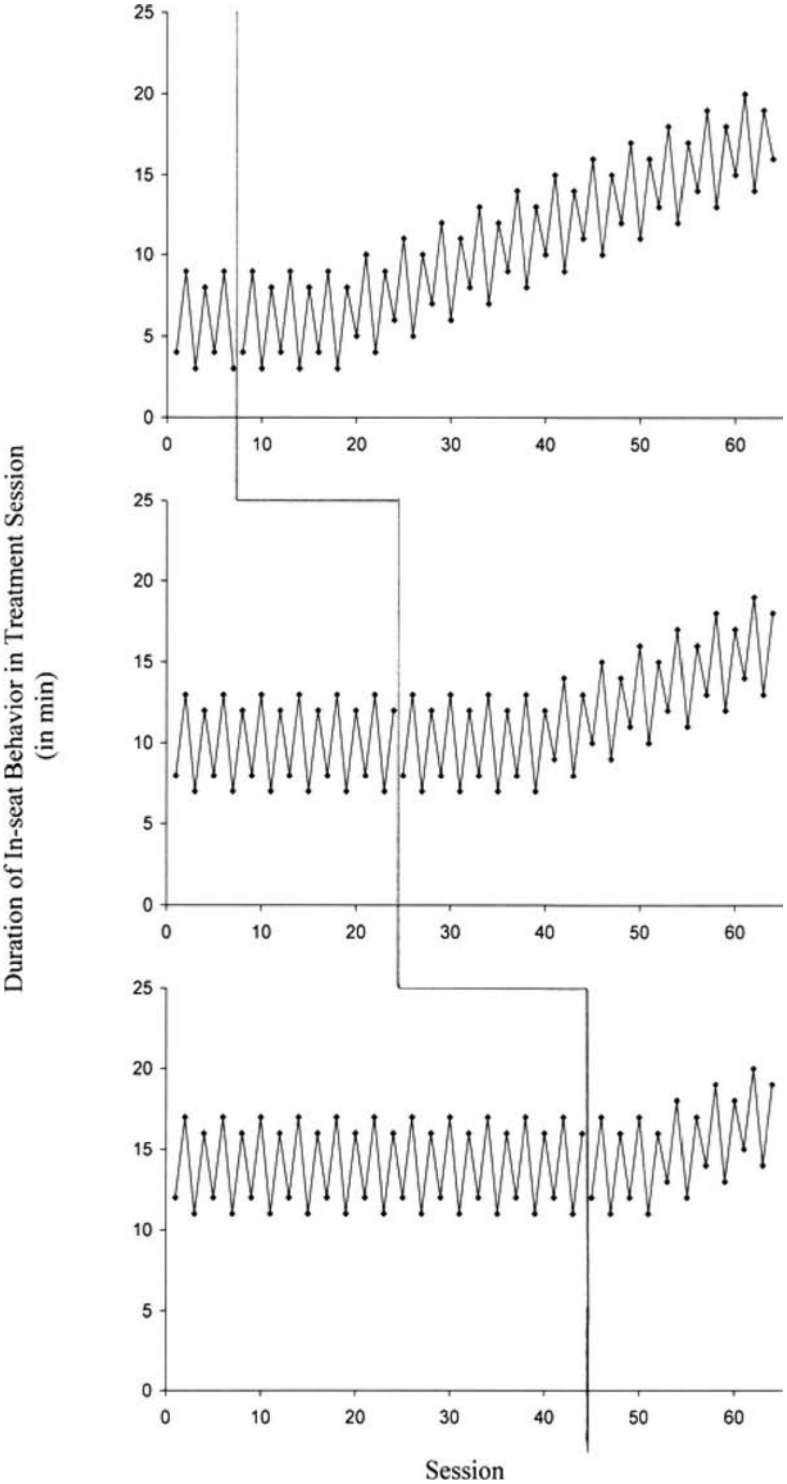

Figure 4. Example of a graph set exhibiting shallow slope and inconsistent latency of change.

blind to which participants submitted which responses. If respondents indicated a desire to enter, they were included in a raffle for a $25 gift certificate to a popular online store. A total of 57% chose to be entered into the raffle.

Results

Manipulation Check on Expectancy of Delayed Change

Thirty-two participants responded to the manipulation check questionnaire. Responses from these participants were recoded into two categories (a) responded as expected or (b) do not know or responded opposite than expected. The average proportion of items responded to in the expected direction was .70 ($SD = .27$, 95% CI $= .60$, .79). Only experts who scored at least 7 of 8 responses as expected on the manipulation check questionnaire were included in analyses involving the expectancy factor (i.e., $n = 15$).

Effects of Steepness, Consistency, and Expectancy of Delay

Judges' responses were recoded to provide an index of the extent to which they judged the data supportive of a functional relation. Graph sets judged to support a confident inference of a functional relation were given a value of "3." Graph sets evoking an unconfident inference regarding presence or absence of a functional relation were given a value of "2." Graph sets judged to support a confident inference that there was not a functional relation were given a value of "1." The sum of these numerical scores across the graph sets assigned to the 8 design cells defined by the cross of the three 2-level within subject factors was the dependent variable for the test of the first research question. Because the data violated the assumption of normality of residuals for repeated measures ANOVA, the first research question was tested using a series of permutation tests, which do not make the normality assumption (Good, 2000).

The permutation test is a nonparametric test analogous to the paired *t* test. The permutation test creates an empirically derived distribution of test statistics through multiple random shufflings of the sign of the difference score for the tested between-condition difference in the observed data. After shuffling, a postshuffle *t* score is computed to quantify the standardized difference between conditions. The observed *t* score is then compared with this distribution of postshuffle *t* scores. If fewer than 2.5% of the postshuffle *t* scores are more extreme than the observed *t* score, the observed *t* score is "statistically significant." When interactions are tested, we are simply asking whether the difference in difference scores could have occurred through a random shuffling of algebraic signs of the difference score, given the observed data patterns. The following analyses were conducted using the computer software program StatXact 4 (StatXact 4 for Windows, 2000).

There were no 3- or 2-way interactions in the data ($p > .07$). Using all 36 participants, the main effect for steepness was significant (Permutation test $T = 4.25$, $p = .00003$, $d = .8$). The mean sum of ratings for steep graph sets was 17.1 ($SD = 4.6$); the mean sum of ratings for shallow graph sets was 13.7 ($SD = 4.0$). Using all 36 experts, the effect of consistency was significant, but trivial in size (Permutation test $T = 2.47$, $p = .02$, $d = .12$). The mean sum or ratings for consistent graph sets was 15.6 ($SD = 4$), while the mean for inconsistent graph sets was similar ($M = 15.1$, $SD = 4$). Among the 15 experts with good evidence that manipulations of dependent variable, population, and measurement context affected their expectancy that the change in the dependent variable would be delayed, there was no expectancy effect (Permutation test $T = 1.29$, $p = .27$). Table 2 presents the means for the cells resulting from crossing the steepness and consistency factors.

Results were also analyzed at the participant level. In the steep versus shallow condition, .72 of the participants judged the steep graphs to indicate stronger evidence of a functional relation than shallow graphs, with a 95% CI (.57, .88). In the consistent versus inconsistent condition, .53 of the participants judged the consis-

Table 2
Means and Standard Deviations for Sum of Steep and Shallow Conditions

Condition	*M*	*SD*
Steep slope with consistent latency	8.80	2.44
Steep slope with inconsistent latency	8.40	2.24
Shallow slope with consistent latency	6.96	1.96
Shallow slope with inconsistent latency	6.92	2.08

tent graphs to demonstrate a functional relation over the inconsistent graphs with a 95% CI (.36, .70). In addition, only .47 of the reduced sample testing the expectancy effect showed higher ratings for graph sets with characteristics thought to create expectations of a delayed change than graph sets with characteristics thought to create expectations of immediate change after the onset of the treatment phase. Therefore, only the steepness manipulation affected a significant majority of experts' functional relation judgments.

Graphs Eliciting Agreement of a Functional Relation

We averaged the within-figure ratings across all raters as an estimate of the "true score" for each graph set (Shavelson & Webb, 1991). Figures with an across-rater average rating of 1 to 1.66 were considered rated as "certain that a functional relation was not present." Figures with an average rating of 1.67 to 2.33 were considered rated as "uncertain as to the presence of a functional relation." Figures with an average rating of 2.34 to 3 were considered rated as "certain that a functional relation was present."

Of 16 graph sets, 11 were judged to be "uncertain." Four graph sets were judged as certain to not demonstrate a functional relation. All four graph sets exhibited shallow slopes. Only one item was judged as "certain" by the raters to exhibit a functional relation (see Figure 5). This item had an average rating of 2.41, and contained graphs depicting steep slopes, consistent latency of change, and variables that were designed to create an expectation that the change in the dependent variable would be delayed. Fifty-four percent of raters confidently judged this item to demonstrate a functional relation (95% CI, .38, .70). The confidence interval does not include .33, indicating this agreement occurred at greater than chance levels. However, the confidence interval does include .50, indicating that a nonsignificant majority rated this graph as a functional relation.

Estimate of Between-Rater Agreement on Functional Relation Judgments

To address between–rater agreement, mean agreement was measured across all possible pairings of experts. The average percentage agreement (number of items on which the pair of raters agreed in their rating/16 items) was .40 (*SD* = .21). Four percent of the pairs of experts scored >.80 agreement. To determine whether agreement was greater for the types of graph sets most likely to yield judgments of a functional relation, mean agreement was computed for steep versus shallow sloped graphs. Agreement for graph sets with steep slopes (.40, *SD* = .23) was about the same as for graph sets with shallow slope (.39; *SD* = .23). In addition, agreement was calculated for the four graph sets that were determined as certain to not demonstrate a functional relation. Agreement among raters across these four items was .42. Only 25% of pairings of raters had agreement of .75 or more across these four graph sets.

Generalizability and Decision Study Analysis

Generalizability studies were conducted because (a) the pairwise agreement was low and (b) we needed to know whether the across-expert average of ratings was sufficiently reliable to interpret the group level findings. Two studies were conducted: (a) one for the sum of ratings for the graph sets in the steep condition versus the sum of ratings for the graph sets in the shallow condition and (b) another for each graph set. A generalizability study allowed us to estimate reliability of among-graph-set variance compared with (i.e., divided by) total variance, which included among-judge, within-graph-set variance (Shavelson & Webb, 1991). One wants there to be much more variance between graph sets or types of graph sets than variance among experts. The intraclass correlation coefficient (g coefficient) for the steep versus shallow distinction was only .50, indicating 50% of the variance was due to factors other than graph sets (e.g., expert raters). The intraclass correlation coefficient for the distinction between each of the 16 graph sets was only .40.

Decision studies were conducted to determine how many experts' judgments need to be averaged to derive reliable estimates of the extent to which graph sets' data support an inference of a functional relation (Shavelson & Webb, 1991). The decision study for the ratings of steep versus shallow graph sets indicates that it would take the average of only 4 randomly selected experts to distinguish the steep from

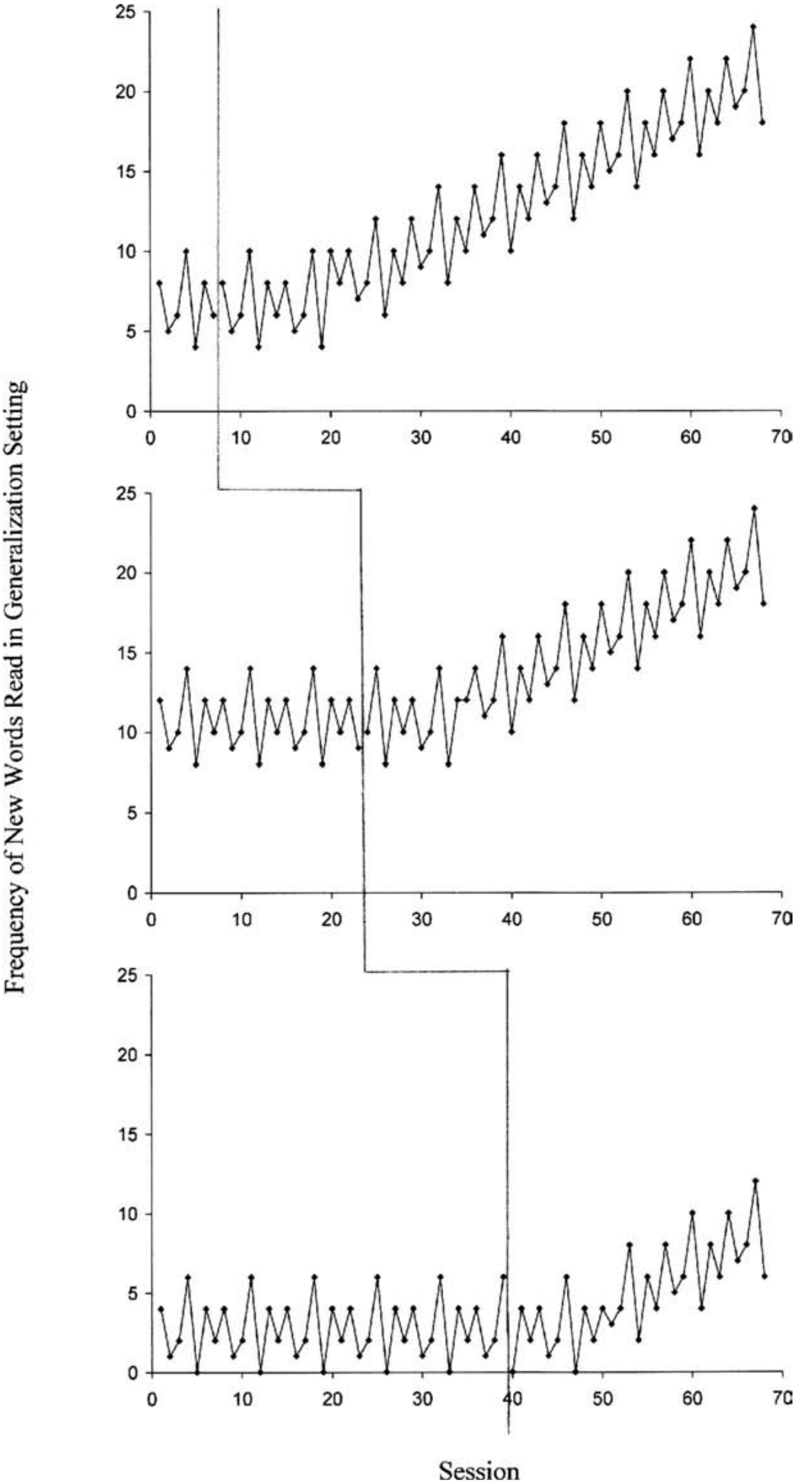

Figure 5. Graph set judged as "certain" to demonstrate a functional relation by participants.

the shallow graph sets with an intraclass correlation of over .8. The decision study for separate graph sets indicated that it would take the average of only 6 randomly selected experts to distinguish among the graph sets with an intraclass correlation of over .8. Clearly, averaging across our sample of experts was sufficient to yield reliable distinctions among graph sets (i.e., our sample size exceeded 4 and 6).

Discussion

The purpose of this study was to test whether steepness of slope in the trend line, consistency of latency of change, and expectancy that the dependent variable would show a delayed change would affect experts' judgment of functional relations in MBL-P data showing at least 1 month delay between onset of the treatment phase and the onset of change in slope of the trend line for the dependent variable. In addition, we described the attributes of the graph set that the majority of experts judged to show a functional relation. Then, we quantified the extent to which all pairs of experts agreed at a point-by-point level in their judgments. Finally, we conducted generalizability and decision studies to determine how many experts' judgments would need to be averaged to reliably distinguish the degree to which types of graphs support an inference of a functional relation.

Experts rated graph sets with steep slopes as showing more evidence of functional relations than graph sets with shallow slopes. However, the average rating for 15 of 16 figures indicated no or equivocal evidence of a functional relation. A slight majority of the experts agreed that only one figure illustrated data that supported a confident inference of a functional relation. This figure contained steep trend lines, consistent latency of change, and contextual information designed to create an expectation that the change in the dependent variable would be delayed. At a graph set by graph set level, the average total percentage agreement between pairs of experts was low (i.e., 40%, even for the steep graph sets). This level of agreement is even lower than the 60% pairwise agreement found by Park et al. (1990). As Park et al. (1990) examined A-B design graphs that most likely exhibited immediate level shifts, comparison of these agreement levels supports the notion that agreement on inferences of a functional relation is even lower in MBL-P with delayed change of the dependent variable. In addition, the generalizability studies described in this article showed that any one expert's judgment is not reliable even when summarizing across graph sets of a similar type (e.g., steep). However, the average across all 36 experts is a reliable discriminator among individual graph sets and between sets of graph sets with steep versus shallow slopes. Indeed, the ratings of only 6 judges were needed to reliably distinguish the level at which the data support a functional relation among graph sets.

Many studies have demonstrated that one can have poor point by point agreement for most pairs of raters yet interpretable data patterns at the mean level if one has "enough" raters. In a review designed to show the value of aggregating across fallible estimates, five separate studies showed that the average of many judges approximates the correct number more closely than any single judgment (Rushton, Brainerd, & Pressley, 1983). Classical measurement theory seeks to explain this phenomenon by stating that the sum (or average) of a set of multiple measurements results in cancelling out over- and underestimates of the true score (Rushton, et al., 1983).

Potential Weakness in the Current Study

Like most laboratory experiments, one can argue that our attempts to control aspects of the data not under investigation may restrict the generalizability of the findings of this study. Like many experimental psychologists, we justified using controlled stimuli because carefully constructed graphs sets allowed us to increase the precision of our inferences. For example, we ensured that the treatment phase in any second or third tier began at least 4 sessions after the dependent variable began to change in the previous tier to reduce the probability that experts would be distracted and influenced by aspects of the graphs we were not studying.

One way to characterize the unusually clean graphed data in this study's stimuli is that the data were highly autocorrelated. A change in the data pattern in strongly autocorrelated data should be easier to detect than a change in data with autocorrelation closer to zero. Therefore, we can assume that raters had an easier time detecting a change (but not necessarily a functional relation) in the graph sets for this study than they would if the graphs displayed less autocorrelation. Detecting change is a necessary, but insufficient, condition for inferring a functional relation. Thus the low agreement and low average ratings occurred despite the strong autocorrelation in the data, not because of it.

Some may argue that we could have controlled for aspects of the data not under study by using a formula to generate the data, as others have done (e.g., Fisher, Kelley, & Lomas, 2003; Matyas & Greenwood, 1990). Although there is value to such an approach for some questions, we did not use it because we were studying delayed changes in slope only. In past studies using a formula to generate the data, standardized differences in phase means were being studied. Large standardized differences in phase means and allowing variability to differ between phases would not allow us to control for level (as defined by the last 3 data points in the baseline and the first 3 data points in the treatment phase) and variability shifts. Controlling for level and variability is useful because it allows precise specification of the type of dependent variable change we are studying. In addition, using standardized phase mean differences as a way to quantify and manipulate "effect size" does not necessarily result in manipulating a functional relation inference when the dependent variables change in slope long after the onset of the treatment.

Methodological issues aside, it must be acknowledged that experts may have responded more conservatively in this study than they would have done in judging their own graphs or in reviewing articles. Thus, they were not likely to make Type I errors (indicating a functional relation was present when it was not). Regardless, the results support the assertion that visual analysts are likely to accept large (in this case steep) changes as compared with small (shallow) changes as evidence of a functional relation (Baer, 1977). Even if the mean expert rating is shifted down (i.e., away from a confident judgment of a functional relation), the relative influence of factors on expert judgment of functional relations, the across-expert ratings on individual graph sets, and the degree of agreement is probably generalizable.

Implications of Findings

The average expert rating for 15 of 16 graph sets in the current study indicated that a functional relation was either "uncertain" or "certain not a functional relation." These data indicate that it is difficult to convince experts that a functional relation exists when the dependent variable's trend changes more than a month after the onset of treatment. However, there was one graph set that the majority of experts judged to show a functional relation. The characteristics of this graph set and the results of the analysis testing factors that influence experts' functional relation judgments suggest that one is more likely to convince experts that a functional relation was demonstrated if the slope of the change, once it occurs, is steep and the latency of the delay in change is consistent across participants. The fact that the graph set for which the majority of experts indicated as evidence of a functional relation also had contextual variables thought to create an expectation that the change would be delayed may be a chance finding. That is, the results of this study do not support the hypothesis that expectancy of delayed change affects experts' judgment of a functional relation.

Given that steepness of slope and consistent latency of change are not within the control of the experimenter, it is risky to use a MBL-P design when dependent variables are expected to show a delayed change in slope. When designing a study in which it is reasonable to expect the dependent variable to change long after the onset of the treatment, one can use a different research design to address the causal question. When one has already collected the data in the context of a MBL-P design and finds that the dependent variable changes more than a month after the onset of the treatment, one should indicate that confidence that a functional relation occurred is weakened unless that change is steep and shows a consistent latency of change across tiers.

Similarly, using the MBL-P design to collect data as part of an RTI framework may be equally problematic. Delayed changes in the dependent variable are most likely to occur for developmental or academic variables, and for populations of learners who have more significant needs. In a school setting implementing an RTI framework, these are exactly the types of dependent variables and learners we might expect to see at tier 2 and tier 3 (i.e., academic variables for students who do not respond to tier 1 interventions). The results of this study suggest judgments by individual practitioners of whether a tier 2 or tier 3 intervention is functionally related to change in behavior would be unreliable in these cases, and may confound programmatic decisions for certain students. Selection of the appropriate single-subject design should therefore be based upon the student, the dependent variables of interest, and the reason for continuous measurement. If the goal of the practitioner is to document change in behavior for a student at tier 1, an A-B design will suffice. If the goal of the practitioner is to detect a functional relation between treatment and changes in behavior to make policy decisions for students at tier 3, an experimental design is needed (Riley-Tillman & Burns, 2009). It is important to select a design where functional relations can be detected reliably, with careful consideration of students and dependent variables. In cases where the dependent variable may show delayed changes, the MBL-P design may not be the most appropriate design.

One "solution" to showing functional relations on variables expected to show changes long after the onset of the treatment phase is to break the skill into component steps and demonstrate functional relations on each of these steps. However, there are skills for which we have insufficient knowledge to decompose the skill completely (e.g., pragmatically appropriate conversational skill, comprehension of connected text). Therefore, it is likely that we will continue to have a need to test functional relations on dependent variables that show delayed slope changes.

The present data indicate the need to show all, not just part, of the data from the treatment phase. In addition, when attempting to show whether a functional relation occurs, the dependent variable needs to be measured temporally close to, and regularly after, the onset of the treatment phase. However, information allowing the reader to determine time between sessions is essential to determining the length of the latency from treatment onset to change in the dependent variable, which according to our data may impact judgment of functional relations. Some authors have used the practice of presenting baseline data, initiating an intervention with a dependent variable likely to produce delayed changes, not collecting or not showing data for those early treatment phase sessions, and then showing postintervention or treatment phase data once the dependent variable changes (Case, Harris, & Graham, 1992; Koegel, Camarata, Valdez-Menchaca, & Koegel, 1998; Whalen & Schreibman, 2003). It can be argued that there are ethical reasons not to test children at times when we have no reason to expect they have yet acquired the taught skills. However, the interval between onset of the treatment phase and the beginning of change during the treatment phase can be represented clearly on graphed data even when testing does not occur temporally close to the onset of the treatment phase. Showing treatment phase data only after the dependent variable has begun to change without a graphically illustrated indication of how long after the onset of the treatment phase such change occurs makes it appear that the change is immediate after the onset of the treatment phase. This study's findings make it clear that these practices should be avoided.

It is probable that some investigators do not understand the need for showing early treatment phase data. After all, before this study, the importance of the latency between treatment onset and change onset on experts' functional relation inferences had not been demonstrated. The results of this study indicate that it is clearly important to note the latency between the onset of the treatment phase and onset of change in the dependent variable in the Results sections because such information is critical to inferring a functional relation in MBL-P designs. The absence of attention to latency information may result in some readers incorrectly equating change during the treatment phase with a functional relation between the treatment and the dependent variable.

Our informal review of textbooks and thorough search of empirical articles examining the factors that influence experts' inference of a functional relation indicates little attention is given to teaching experts or students to interpret MBL-P data in which the dependent variable's slope change is delayed after the onset of the treatment phase. Future studies are needed to determine whether direct instruction with performance feedback affects agreement among visual analysts when making functional relation judgments about graph sets with delayed change. Such training may increase the agreement among experts that a functional relation cannot be judged confidently from MBL-P data when the dependent variable (a) changes long after the onset of the treatment phase and (b) has inconsistent latency of change across tiers. In addition, such training also may increase agreement among experts that a functional relation can be confidently judged from MBL-P data even when the latency of change is long after the onset of the treatment phase if the slope of the change is steep and latency of change is consistent across tiers. The question of whether steep slopes should be such a strong influence on whether experts judge that the data support an inference of a functional relation when the change in the dependent variable is delayed will have to be left to future debate.

Just as group experiments differentiate between time effects (i.e., significant change over time) and treatment effects (i.e., significant differences in gain between control and treatment groups), one purpose of single-subject treatment research is to differentiate between change during a treatment phase from change because of a treatment. Teachers, insurance companies, policymakers, and parents need to know whether treatment is worth investing in because it reliably causes desired changes in children. It is noteworthy that all of the graphs in this study showed a change in the dependent variable in a therapeutic direction during the treatment phase. In this study, we added to what is presently understood regarding characteristics of visual data that influence raters' inferences of functional relations and their agreement when making this judgment. The systematic examination of specific graph characteristics leading to judgments of functional relations may aid in more accurate interpretation of data relevant to establishing evidence-based practice and to making programmatic decisions within the RTI framework.

References

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10,* 167–172.

Barnett, D. W., Daly, E. J., Jones, K. M., & Lentz, F. E., Jr. (2004). Response to intervention: Empirically based special service decisions from single-case designs of increasing and decreasing intensity. *Journal of Special Education, 38,* 66–79.

Case, L. P., Harris, S. R., & Graham, S. (1992). Improving mathematical problem-solving skills of students with learning disabilities: Self-regulated strategy development. *Journal of Special Education, 26,* 1–19.

Daly, E. J., Martens, B. K., Barnett, D., Witt, J. C., & Olson, S. C. (2007). Varying intervention delivery in response to intervention: Confronting and resolving challenges with measurement, instruction, and intensity. *School Psychology Review, 36,* 562–581.

DeProspero, A., & Cohen, C. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36,* 387–406.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119–158). Hillsdale, NJ: Erlbaum.

Gibson, G., & Ottenbacher, K. J. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *Journal of Applied Behavioral Science, 24,* 298–314.

Good, P. I. (2000). *Permutation tests : A practical guide to resampling methods for testing hypotheses* (2nd ed.). New York: Springer.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165–179.

Kazdin, A. (1981). Drawing valid inferences from case studies. *Journal of Consulting and Clinical Psychology, 49,* 183–192.

Kennedy, C. H. (2005). *Single-case designs for educational research.* Boston, MA: Allyn & Bacon.

Koegel, L. K., Camarata, S., Valdez-Menchaca, M., & Koegel, R. (1998). Setting generalization of question-asking by children with autism. *American Journal on Mental Retardation, 102,* 346–357.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341–351.

Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy, 40,* 464–469.

Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *Journal of Experimental Education, 58,* 311–320.

Parsonson, B., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. Levin (Eds.), *Single-case research design and analysis* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: Single-case designs for measuring response to intervention.* New York: The Guilford Press.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

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

Shernoff, E. S., Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: An illustration of task force coding criteria using single-participant research design. *School Psychology Quarterly, 17,* 390–422.

StatXact 4 for Windows [Computer software]. (2000). Cambridge, MA: CYTEL Software Corporation.

Whalen, C., & Schreibman, L. (2003). Joint attention training for children with autism using behavior modification procedures. *Journal of Child Psychology and Psychiatry, 44,* 456–468.

Yoder, P., Reichow, B., Wolery, M., & Tapp, J. (2007). *Visual analysis of delayed trend changes in multiple baseline across participants designs.* Unpublished manuscript, Vanderbilt University, Nashville, TN.

Correction to Lieberman et al. (2010)

In the article, “Visual analysis of multiple baseline across participants graphs when change is delayed,” by Rebecca G. Lieberman, Paul J. Yoder, Brian Reichow, and Mark Wolery (*School Psychology Quarterly*, Vol. 25, No. 1, pp. 28-44. doi: 10.1037/a0018600), there were several errors in the text. The corrected text is shown below.

On page 38, column 2 the sentence starting “The intraclass correlation (g coefficient) for” should have read “The intraclass correlation coefficient (g coefficient) for the steep versus shallow distinction was only .12, indicating 88% of the variance was due to factors other than graph sets (e.g., expert raters).” The sentence immediately following that starts “The intraclass correlation coefficient for the distinction” should have read “The intraclass correlation coefficient for the distinction between each of the 16 graph sets was only .13.”

The paragraph immediately following on the same page, beginning with “Decision studies were conducted”, should have read:

“Decision studies were conducted to determine how many experts’ judgments need to be averaged to derive reliable estimates of the extent to which graph sets’ data support an inference of a functional relation (Shavelson & Webb, 1991).The decision study for the ratings of steep versus shallow graph sets indicates that it would take the average of 26 randomly selected experts to distinguish the steep from the shallow graphs sets with an intraclass correlation of over .8. The decision study for separate graph” should have read “The decision study for separate graph sets indicated that it would take the average of 26 randomly selected experts to distinguish among the graph sets with an intraclass correlation of over .8. Clearly, averaging across our sample of experts was sufficient to yield reliable distinctions among graph sets (i.e., our sample size exceeded 26).”

On page 40, column 2, line 11, the sentence that starts “Indeed, the ratings of only 6” should have read “Indeed, the ratings of 26 judges were needed to reliably distinguish the level at which the data support a functional relation among graph sets.”

DOI: 10.1037/a0025527