The Cost of Increased Validity: Combining a Multiple Baseline Design with an ABAB Design

Geoffrey Whitehurst

***Abstract*—A previous feasibility study, degradation of pilot flight performance when transitioning from digital instrumentation to analog instrumentation for the first time, was replicated using two single case research designs (SCDs). The validity of SCDs has often been questioned by researchers who use between-group designs. However, within the research fields using single case research, it is suggested that the validity of SCDs can be improved by systematic replication of single case experiments. This study investigated whether validity could be improved by systematic replication of single case experiments by comparing two SCDs: the multiple baseline design (MBD), which increases replications across subjects to improve validity, and the combined design, which increases replications both across and within subjects and may provide greater improvement in validity. The two designs were compared in terms of results, validity, and cost. The results of the data analyses did not provide any significant advantages or disadvantages for either design, and the improvement in validity of the combined design came at a considerable increase in cost.**

***Index Terms*—Analog Instrumentation, Combined Design, Digital Instrumentation, Multiple Baseline Design, Single Case Research**

I. Literature Review

Since its development in the early 20th century, across many fields, including aviation, null hypothesis significance testing (NHST) has been the most common methodology used in experimental and quasi-experimental studies to observe intervention effects. Garson [1] described experimental studies as characterized by the ability to randomly assign subjects into treatment and control groups, and quasi-experimental studies as those in which comparison groups are not true randomized groups. In either case, a researcher rejects the null hypothesis in favor of the alternative hypothesis if the *p*-value of the calculated test statistic is sufficiently small (less than the α-value) [2].

However, an important issue that has been a concern for researchers for many years, but is now becoming more prominent, is statistical power. Design, sample size, effect size, significance level, and the statistical test are all factors that determine statistical power, but sample size is often the only factor that the researcher may have control over. In aviation research, this control is often very limited due to a lack of resources, specifically, low numbers of participants meeting study criteria and the overwhelming cost and availability of flight simulators or aircraft. One solution to the problem of sample size is the single case design (SCD).

For more than a century, single case research has been used in the field of psychology. However, in aviation research, SCDs have rarely been used. A SCD normally begins with collecting baseline data, a series of observations referred to as the A phase. These data provide information about the participant prior to the introduction of the intervention. Baseline data can provide descriptive information; the participant's current performance based on the value and variability of the dependent variable; and predictive information, the participant's future performance based on the projected value of the dependent variable from the data trendline. Data collected during the intervention (B phase) can then be compared to the predicted performance based on the A phase to demonstrate intervention effects.

A. Threats to Internal Validity in SCDs

Threats to internal validity are confounding variables, such as history, maturation, and testing, within the study itself. As with any experimental design, threats to internal validity are a potential problem in SCDs and require the designs to be structured to address these threats. Replication of the A and B phases, to produce an "effect replication," has been the main mechanism for controlling threats to internal validity in SCD research. Acceptable evidence standards for showing intervention effects suggested by Kratochwill et al. [3] state that a minimum of three different phase repetitions are required to meet evidence standards. These phase repetitions can be either solely within participants (ABAB design) or both within and between participants (MBD). Fig. 1 displays both designs. The ABAB design can be conceived as a horizontal design in which the effect replication is produced by one person undergoing four phases. The MBD is a vertical design in which an AB design is conducted simultaneously with three or more participants. The introduction of the B phase is staggered in time across the participants to improve internal validity. Note that the replications in the MBD design are produced by having more than one participant.

The MBD allows the researcher to make both within-series and between-series comparisons to draw valid inferences from the data. The within-series comparison is the horizontal AB component, where the comparison is between the two phases for each individual participant. The between-series

This paper was submitted on October 07, 2013.

Geoffrey Whitehurst is with the College of Aviation, Western Michigan University, Kalamazoo, MI 49008, USA (e-mail: geoff.whitehurst@wmich.edu).

DOI: 10.5176/0000-0004_1.1.4

comparison is the vertical component, where the comparisons are between all the participants. That is, the A phase of each participant can be compared with the A phase of the other participants, and each B phase can be compared with the other B phases.

To increase systematic replication of single case experiments in order to try to improve both internal and external validity, a combination of the ABAB design and the MBD could be constructed, hereafter referred to as a combined design. This combined design would provide three phase changes for each of the three or more participants, and provide a minimum (for three participants) of nine phase changes across all participants. One problem with using a combined design in applied aviation research is the possibility of "Testing," one of the threats to internal validity. Testing is defined by Kratochwill et al. [3]: "Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with an intervention effect" (p. 10). For example, continuous exposure of participants to some new instrumentation might reduce the negative effect on their performance over time. Although some testing was expected to occur in this study, it was not expected to be sufficient to prevent the intervention effect from being observed in the second intervention phase.

B. *Threats to External Validity in SCDs*

External validity refers to how readily a study allows its findings to generalize to the population at large. With SCDs requiring only small sample sizes, often n = 1, the external validity is often questioned.

To improve external validity, systematic replication of single case experiments are needed [4]. The most common form of the design that meets the replication criteria advanced by Horner et al. [5] is the MBD, which includes an alternating baseline and intervention phases for each of three or more participants and provides the minimum requirement of three phase changes across three participants. The comparison across the participants strengthens the design's external validity by providing the between-series comparisons required for generalizability.

SCDs, as the name suggests, originated with the psychological study of one individual and was not concerned with external validity, only the internal validity of the study. A review of the literature shows that in most fields that use single case research, this is still the case. However, with research in other fields of research now using single case designs, external validity has steadily become an issue when using single case designs. The introduction of the MBD improved external validity by having both between-series as well as within-series comparisons. The combined design has the advantage of the MBD's between-series comparisons together with the systematic replication suggested by Hayes [4]. Applied aviation research is a field in which generalization is a necessity, so comparing the combined design to the MBD in terms of validity with regard to results and cost is very important.

II. PURPOSE

In this study, a combined design was used to replicate a study that examined the flight performance of student pilots transitioning from flying digital flight instrumentation equipped aircraft to flying analog flight instrumentation equipped aircraft [6].

The purpose of this study is to compare the combined design to the MBD, used in a previous study [7], when used in the same replicated applied aviation experiment, in terms of:

1. Internal validity,
2. External validity, and
3. Results of visual and statistical data analyses.

A brief cost analysis was also conducted to determine what increase in cost can be expected when the combined design is used instead of the MBD.

III. RESEARCH METHODOLOGY

A. *Research Design*

In a previous study [7], an MBD was used to examine the flight performance of student pilots transitioning from flying digital flight instrumentation (DFI) equipped aircraft to flying analog flight instrumentation (AFI) equipped aircraft. The DFI aircraft is fitted with the type of instrumentation the participants were learning to fly with, and the AFI aircraft is fitted with the type of instrumentation the participants have no experience flying with. In this study, the combined design was used to replicate the same study to enable a comparison with the MBD.

A Personal Computer Aviation Training Device (PCATD), which is capable of emulating both digital and analog flight instrumentation, was used as the platform for assessment. Each session required the participant to fly a radar vectored instrument flight profile, consisting of take-off, climb, cruise, and an Instrument Landing System (ILS) approach to a visual landing. The participant's flight performance was assessed using the FAA's Instrument Certification Practical Test Standard (PTS). Any deviation outside the limits set in the PTS was recorded as an error and the total number of errors per flight was used to assess overall flight performance. Each session was recorded electronically, by the computer flight software, and visually, by a video camera, to enable appropriate analysis.

The combined design, like the MBD, requires that data are collected on all participants prior to any intervention to provide baseline data for each participant. In this study, the baseline data are the flight performances of the participants flying the PCATD configured to emulate a Cessna 182 Glass, a DFI aircraft (see Fig. 2). The intervention data are the flight performances of the participants flying the PCATD configured to emulate the Cessna 182 Skylane RG, an AFI aircraft (see Fig. 3). Ideally, the baseline is expected to have no trend and no variability, thus giving "stable" data. *Trend* refers to a continuous increase or decrease in mean flight performance, and *variability* refers the difference between the actual flight performance each session and the mean flight performance. However, in this study, the participants were flying unfamiliar equipment but were continuing to “learn” during the study, and therefore both the baseline phase data and intervention phase data was expected to have a “downward” trend and reducing variability due to learning. The expected “downward” trend was to show an improvement in performance, a reduction in errors committed, and the reduction in the variability of the data was expected to be the result of the participant becoming more familiar with the equipment and environment.

There is no specified limit to the variability of the data for it to be considered “stable.” Kratochwill et al. [3] stated, “If the effect of the intervention is expected to be large and demonstrates a data pattern that far exceeds the baseline variance, a shorter baseline with some instability may be sufficient to move forward with intervention implementation” (p. 19). This puts the onus on the researcher to have some prior knowledge of the expected size of the intervention effect, from either previous research or review of relevant literature.

For this study, the acceptable variability of the data for introducing the intervention is based on the data from the original study [6] and the expected error rate of flight students at this stage of their flight training. The acceptable variability was set to an error rate within plus or minus 2 PTS errors of the trend line for two continuous sessions. Therefore, for this study, data are defined as “stable” when a level or downward trend and an error rate within plus or minus 2 PTS errors of the trend line for two continuous sessions has been achieved.

Each participant is randomly assigned to his or her order of participation (1, 2, 3, or 4) and begins by flying the DFI aircraft (baseline [A] phase). When all participants achieve “stability” in the A phase, participant 1 begins flying the AFI aircraft (intervention [B] phase). The other participants continue flying the A phase until participant 1 achieves stability in the B phase. Participant 2 then begins flying the B phase. Participant 1 continues flying the B phase and the other participants continue flying the A phase. This procedure is repeated until all participants are flying the B phase. For the combined design, the procedure is then repeated for a second A phase and again for a second B phase. The study is completed when all participants have achieved stability in the second B phase. Each phase requires a minimum of five data points, even if “stability” is achieved earlier.

B. *Participants*

Participants were recruited from flight students in a four-year university flight science degree program who met the following criteria: (a) within 15 flights of completion of instrument certification, and (b) no experience flying an aircraft equipped with analog flight instrumentation. These criteria were confirmed during an initial interview with each participant.

Criterion (a), 15 hours to instrument certification, was selected to ensure proficiency in instrument flying, but also to provide sufficient time to complete the research project before participants completed the instrument certification. This is important because once student pilots complete their instrument certification, they can begin their multi-engine course, the next stage of their training, and the multi-engine aircraft are a mixed fleet of both digital and analog flight instrumentation. This would present the possibility that a participant could fly aircraft equipped with analog flight instrumentation, thus compromising the other criterion for participation. Criterion (b) was to ensure that the introduction of the intervention, changing to an AFI aircraft, was the first time the participant had ever flown using this type of instrumentation.

Four student pilots were accepted as volunteers to participate in the study. Because of unrelated commitments, one participant withdrew during the intervention phase. The three remaining participants completed the study. Only the data from the three participants who completed the study were used in the data analysis.

C. *Method*

A PCATD was set up to emulate the Cessna 182 Skylane Glass for the digital flight instrumentation (DFI) equipped aircraft, and the Cessna 182 Skylane RG for the traditional analog flight instrumentation (AFI) equipped aircraft.

Each participant flew the PCATD emulating the DFI aircraft for the A phase and the AFI aircraft for the B phase. During each simulated flight, participants were asked to fly a radar vectored flight pattern and to complete an instrument approach. Each flight was recorded for later analysis of the participant’s flight performance.

D. *Dependent Variable*

The dependent variable for measuring participant flight performance consisted of the total number of times the aircraft deviated from the criteria listed in the FAA’s PTS for instrument flight check rides. The criteria are: (a) turn onto and/or maintain heading within ±10º; (b) level off and/or maintain altitude within ±100 feet; and (c) for all stages of flight, maintain required speed within ±10 knots. A deviation beyond any one of the three limits was recorded as one PTS error and the total number of PTS errors was recorded for each session. To enable an accurate assessment of the participant’s performance a Contour Nflightcam video camera was positioned with the flight controls in front of the participant. The wide angle 170º lens captured all information displayed on the flight instrumentation, as seen by the participant. The flights were initially recorded on an internal 16 GB Micro SD video card and later downloaded to the same external Seagate 1.0 terabyte hard drive used for recording the simulation technical parameters. The videos were replayed at a later time for analysis, data collection, and interrater reliability checks.

E. *Apparatus*

The PCATD equipment consisted of a Dell Optiplex SX260® computer with a Pentium® 2.40 gigahertz processor, and 1.0 gigabytes of SDRAM memory. Operating software was Microsoft Windows XP and simulation software was On-Top version 9.5. Flight support equipment for the PCATD included a Cirrus yoke, a throttle quadrant, an avionics panel, and rudder pedals. The On-Top software simulated the two aircraft types used in this study, the Cessna 182 Skylane Glass and the Cessna 182 Skylane RG. The technical flight parameters, which depicted how well participants flew the designated flight patterns, vertically and horizontally, were recorded for each flight on an external Seagate 1.0 terabyte hard drive. The On-Top simulation software automatically recorded these technical parameters and enabled them to be replayed at a later time for analysis, data collection, and interrater reliability checks.

F. Flight Patterns

In an effort to minimize any practice effects, four different flight patterns were used on a random basis. Participants were told that the PCATD aircraft was not programmed for any system failures and that the flight pattern would be a radar-vectored instrument flight with an instrument landing system (ILS) approach to a full-stop landing. By using vectored instrument approaches and not having system faults, the flight environment should have allowed for consistent flight performance. The approach patterns used should not have provided the participant with any adverse stress or pressure to perform, as these patterns were typical of their existing training environment. All flight patterns included a take-off and climb to an initial altitude; a radar vectored flight pattern, including one descending turn and an initial heading for localizer interception; and then an ILS approach to decision height for a visual landing.

Data were collected during instrument flight conditions, which began on cloud penetration at 300 feet on climb out and ceased at decision height (200 feet above the ground) on the ILS, when the participant switched to visual references for landing. Each flight pattern took approximately 20 minutes to complete. To realistically simulate an actual flight pattern and ensure that it was flown in a consistent way across trials and participants, the experimenter provided typical air traffic control instructions throughout the flight pattern. The experimenter, located in an adjacent room, communicated with the participant using a commercially available intercom system.

G. Data Collection

Data were collected from the participants over a period of 8 weeks. Participants would each fly one flight pattern per session, two or three times per week, based on their academic and flying schedules. Participants were randomly assigned to their order of participation, and this order was then maintained during the study.

Each participant's flight pattern was visually recorded in order to capture the exact information displayed on the flight instruments seen by the participant flying the PCATD. The advantages of reviewing the video recording for data collection were (a) each recording could be assessed by more than one rater, and (b) recordings could be stopped and/or rewound to confirm accuracy of assessment.

H. Interrater Reliability

A number of statistics can be used to determine interrater reliability. The intraclass correlation coefficient (ICC) is an index of the reliability of the ratings for a typical single judge. It is employed when most of the data are collected using only one judge, but two or more judges are used on a subset of the data for purposes of estimating interrater reliability [8]. Different guidelines exist for the interpretation of ICC, but one reasonable scale is that an ICC value of less than 0.40 indicates poor reproducibility; ICC values in the range 0.40 to 0.75 indicate fair to good reproducibility, and an ICC value of greater than 0.75 shows excellent reproducibility [9].

IV. Data Analysis

SCD designs are found predominantly in the social sciences, where intervention effects are expected to be large and could easily be detected by visual analysis. With the expansion of this methodology into other fields of research, where intervention effects may not be large, visual analysis is no longer considered sufficient. Therefore, statistical analyses have been and continue to be developed. In this study, both visual and statistical analyses were used to analyze the data from the combined design. Comparisons were also made between results from the visual and statistical analyses of the data from the MBD.

A. Visual Analysis

In this nonstatistical method of data analysis, data are plotted on a graph, in which the *y*-axis represents the dependent variable and the *x*-axis represents units of time [10]. The data for each participant are plotted on separate graphs, which are then arranged above each other for visual comparison of the intervention effect (see Fig. 4). On the basis of these graphs, a judgment is reached about the reliability or consistency of intervention effects [11].

In fields of research where single subject designs are common, such as psychology and special education, guidelines for visual assessment are being established. These guidelines suggest that to assess the effects within single subject designs, six features should be considered to examine within- and between-phase data patterns: (1) level, (2) trend, (3) variability, (4) immediacy of the effect, (5) overlap, and (6) consistency of data patterns across similar phases [12] [13] [14] [15] [16] [17] (see Fig. 5). The six features are defined as follows: "level" refers to the mean score for the data within a phase; "trend" refers to the slope of the best-fitting straight line for the data within a phase; "variability" refers to the range or standard deviation of data about the best-fitting straight line. "Immediacy of the effect" refers to the change in level between the last three data points in one phase and the first three data points of the next. The more rapid (or immediate) the effect, the more convincing the inference that change in the outcome measure was due to manipulation of the independent variable. "Overlap" refers to the values of the data points in the intervention phase approaching the values of the data points in the baseline phase. "Consistency of data in similar phases" involves looking at data from all phases within the same condition (e.g., all "baseline" phases; all "intervention" phases) and examining the extent to which there is consistency in the data patterns from phases with the same conditions. The greater the consistency, the more likely the data represent a causal relation.

Examination of the data within a phase is used (a) to describe the observed pattern of a unit's performance; and (b) to extrapolate the expected performance forward in time, assuming no changes in the independent variable were to occur [18], that is, extend the trend line into the next phase. The six visual analysis features are used collectively to compare the observed and projected patterns for each phase with the actual pattern observed after manipulation of the independent variable. This comparison of observed and projected patterns is conducted across all phases of the design [3].

All six features may not be relevant in all fields. Whitehurst (under review) found in his study using the MBD that of the six standards for visual analysis of data, only four were suitable for most types of applied aviation research. These four features were "level," "variability," "immediacy of the effect," and "consistency of data in similar phases." These four were considered suitable for the following reasons: "Level" would seem to apply to all fields of research, as it gives an indication of any change in the dependent variable that could be attributed to the introduction of the intervention; "Variability" will depend on the participants and would be an important feature to analysis in all fields; "Immediacy of the effect" is essential if the effect of the intervention by chance is to be discounted; and "Consistency of data patterns across similar phases" is an essential feature for fields of research if the effect of the intervention by chance is to be discounted.

The other two features, "trend" and "overlap," were considered unsuitable for the following reasons: "Trend" would be suitable for fields of research in which the intervention is expected to have a distinct effect, or even a reversal of the slope; and "Overlap" is not useful for analysis of this study as "overlap" is expected because of "learning" and could be expected for similar reasons in other research studies in aviation.

To infer a causal relationship between the dependent and independent variables by visual analysis, the researcher/rater is looking for a "consistency of data patterns across similar phases" but can see an "immediacy of effect" at the introduction of the intervention that shows a change in the "level" and is observable outside the "variability" of the data.

B. Statistical Analysis

Although statistical analyses are used extensively in between-group experimental designs, it was not until the 1970s that "statistical analyses for single case data began to receive increased attention" and "statistical analyses were proposed as a supplement to or replacement of visual inspection to permit inferences about reliability or consistency of changes" [14] (p. 241). Morley and Adams [19] recommended complementing visual analysis with a statistical analysis of the data, whenever possible.

Several statistical methods have been developed for the analysis of data from some SCDs, including the AB and ABAB. However, fewer methods are available for the analysis of data from a combined design or a MBD. Meta-analysis is one method that has been considered for these designs. Van den Noortgate and Onghena [20] also suggested the use of hierarchical linear models (HLM) for single case data. In this study, I used HLM to analyze data from both the combined design and the MBD.

HLM is commonly used in many research fields where data are multilevel or hierarchical, for example, students nested within classrooms and classrooms nested within schools. SCDs can also be considered as hierarchical, with measurements nested within individuals. Van den Noortgate and Onghena [20] suggested that data from a combined design or MBD can be modeled using a two-level HLM. The overall phase effect for the combined design was calculated using two baseline and two intervention phases, whereas the MBD overall phase effect was calculated using only one baseline and one intervention phase. The regression equations for the unconditional model, or the model with no treatment indicator, for both designs are:

For level 1

$$Y_{ij} = \beta_{0j} + e_{ij},\quad e_{ij} \sim N(0,\sigma^2) \qquad (1)$$

For level 2

$$\beta_{0j} = \gamma_{00} + u_{0j}\quad u_{0j} \sim N(0,\tau_{00}{}^2) \qquad (2)$$

where

Y_{ij} is the response score of participant j (j = 1, 2, 3 for both designs) for occasion i
(i = 1….20 for the MBD and i = 1…..50 for the combined design);

β_{0j} is the mean response for participant j;

γ_{00} is the mean across participants;

u_{0j} is the random error associated with participant means, var $(u_{0j}) = \tau_{00}$;

e_{ij} is the random error associated with occasion i for participant j, var $(e_{ij}) = \sigma^2$.

The regression equations for the conditional model, or the model with the treatment indicator, are:

Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j}(phase)_{ij} + r_{ij},\quad r_{ij} \sim N(0,\sigma^2) \qquad (3)$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad \text{and} \qquad (4)$$

$$\beta_{1j} = \gamma_{10},\quad u_{0j} \sim N(0,\tau_{00}{}^2) \qquad (5)$$

Where

Y_{ij} is the response score of participant j (j = 1, 2, 3 for both designs) for occasion i
(i = 1….20 for the MBD and i = 1…..50 for the combined design);

$(phase)_{ij}$ is an indicator that equals 1 if occasion i for participant j is part of the intervention phase, 0 otherwise;

β_{0j} is the mean response for participant j in the baseline phase;

β_{1j} is the magnitude of the effect of the intervention on participant j;

γ_{00} is the mean baseline level;

γ_{10} is the mean intervention effect;

u_{0j} is the random error associated with participant means, var $(u_{0j}) = \tau_{00}$;

r_{ij} is the random error associated with occasion i for participant j controlling for (phase) and is a conditional or residual variance, var $(r_{ij}) = \sigma^2$.

In the conditional model, the parameters of interest are the fixed effects γ_{00} and γ_{10} and the variance parameters σ^2 and τ_{00} . The parameters of interest can be calculated using the Scientific Software International (SSI, Inc.) HLM7 software. An estimate of the effect size can also be computed by dividing the overall between phase effect (γ_{10}) by the square root of the residual between-person variance (σ^2) [20].

V. Results

In this section, I present the interrater reliability followed by the results from visual analysis and the statistical analysis.

A. *Interrater Reliability*

All videos were reviewed and rated by the principal investigator (PI). A random selection of 20% of the videos, from each phase of each participant, was reviewed and rated by a Certified Flight Instructor Instrument-Aircraft (CFII-A) to provide interrater reliability data. The second rater is a CFII-A with 13,500 flight hours who has been instructing student pilots on instrument flying for 30 years and has been a company check pilot for 20 years. The ICC was calculated using SPSS one-way random effect model and the single measure ICC = .948, 95% CI = (.894, .975) shows excellent reproducibility [9].

B. *Visual Analysis*

For visual analysis, the data were plotted for each of the three participants. Fig. 6 shows the graphed data for the three participants for the combined design. The dotted trend-lines in the first Phase A show that "stability," a "downward" trend, and variability about the trend line, within the plus or minus 2 PTS errors, is achieved for all three participants within the first five sessions. Participant 1 began the B phase at session 6. After an initial increase in the number of PTS errors (from 0 to 6 errors), which marks the intervention effect, "stability" was achieved by session 10, after 5 sessions of the intervention phase. Therefore, Participant 2 began the B phase at session 11. Also, after a marked intervention effect (from 6 to 19 errors), Participant 2 achieved "stability" by session 15. The B phase for Participant 3 therefore began at session 16. Participant 3 also had a marked intervention effect (from 9 to 19 errors), before achieving "stability" by session 20.

The return to A phase for Participant 1 began at session 21 with no withdrawal effect and an almost error-free phase. Participant 2 returned to A phase at session 26 with a withdrawal effect (from 0 to 5 errors), before achieving "stability" by session 30. Participant 3 returned to A phase at session 31 without withdrawal effect and achieved "stability" by session 35. The second B phase was introduced for Participant 1 at session 36, and there was an intervention effect (from 0 to 3 errors), but a smaller increase than at the introduction of the first B phase. Participant 1 quickly achieved "stability," so the second B phase for Participant 2 was introduced at session 41. Again a smaller intervention effect (from 1 to 5 errors) was observed with a quick return to "stability" for Participant 2. Participant 3 began the second B phase at session 46 with another marked intervention effect (from 0 to 14 errors). The study was concluded after Participant 3 quickly returned to "stability" in the second B phase at session 50.

It can be seen that for each participant there was a marked intervention effect at the introduction of the two intervention phases. Although clearly observable, the intervention effect experienced by Participants 1 and 2 at the introduction of the second intervention phase was smaller than that experienced by Participant 3. The fact that the intervention was introduced at different times for each of the participants suggests that the degradation in flight performance (the intervention effect) experienced by each participant is directly related to the change from digital flight instruments to analog flight instruments (the intervention) and not a chance event. The means and standard deviations (*SD*) of the number of PTS errors for the combined design are presented in Table I. For comparison, the means and *SD*s from the MBD are given in Table II.

C. *Statistical Analysis – HLM*

The two-level HLM models in Equations 1 and 2 can be used for both the MBD and combined designs. The results are presented in Table III and Table IV, respectively.

For the combined design, the estimated overall baseline mean (γ_{00}) is 3.35 and the estimated coefficient of the overall phase effect (γ_{10}) is 2.59, which is statistically significant, $p <$ .001. An estimate of the overall effect size is calculated by dividing the overall between-phase effect (2.59) by the square root of the residual between-person variance (3.85) and is 0.67, a large effect size.

For the MBD, the estimated overall intercept (γ_{00}) is 5.43 and the estimated coefficient of the phase-indicator (γ_{10}) is 3.50, which is statistically significant, $p = .001$. An estimate of the overall effect size is calculated by dividing the overall between-phase effect (3.50) by the square root of the residual between variance (3.69) and is 0.95, a very large effect size.

The results from both designs show there was an effect at the introduction of the intervention. However, the overall effect size of the combined design was smaller than the overall effect size of the MBD. The effect size is calculated by dividing the overall between phase effect by the square root of the residual between variance. The overall between phase effect reduced, whereas the square root residual between variance increased slightly. The reduction in effect size is mainly due to a reduction in the between phase effect, confirming the "learning" the participants were expected to make in flying the PCATD and assimilating the new form of information.

VI. Discussion

Hayes [4], Horner et al. [5], and Kratochwill et al. [3] all argue that both internal and external validity can be improved by systematic replication of single case experiments. To increase replications, extra phases can be added to the design; for example, an ABAB design becomes ABABAB design; or, more participants can be added to an MBD—a three participant AB, AB, AB design becomes a four participant AB, AB, AB, AB. A third option is the combined design, which combines the ABAB with the MBD, which was used in this study. The problem with increasing the number of replications, either through phases or participants, is the inevitable increase in time and associated costs, especially in today's economic climate. Thus, it is important to compare the combined design to the MBD to see if the advantages of the combined design outweigh the additional costs.

I compared the designs with respect to the internal and external validity and the results of the analyses from the two designs. I also compared the cost for the combined design to the MBD to determine what increase in cost is associated with increasing replications.

A. Internal Validity

Kratochwill et al. [3] list the following nine threats to internal validity in their *Standards* for SCDs: Ambiguous Temporal Precedence, Selection, History, Maturation, Statistical Regression, Attrition, Testing, Instrumentation, and Additive and Interactive Threats to Internal Validity. The combined design and the MBD deal with these threats as follows:

Ambiguous Temporal Precedence – Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect. In both designs, the dependent variable is observed for several measurements before actively manipulating the independent variable at different time points for different participants. The effect of the independent variable on the dependent variable is then observed for several measurements. In this way, both the combined design and the MBD negate this threat.

Selection – Systematic differences between/among conditions in participant characteristics could cause the observed effect. Both the combined design and the MBD negate this threat by exposing each participant to both conditions of the experiment.

History – Events occurring concurrently with the intervention could cause the observed effect. Both the combined design and the MBD negate this threat by the replication of the intervention phase at different points in time.

Maturation – Naturally occurring changes over time could be confused with an intervention effect. Both the combined design and the MBD negate this threat by the replication of the intervention phase at different points in time.

Statistical Regression – When cases are selected on the basis of their extreme scores, their scores on other measured variables typically will be less extreme, a psychometric occurrence that can be confused with an intervention effect. This is unlikely to be a threat for applied aviation research, and was no threat to this study, as participants are not normally selected on their individual flying ability, but on their flying ability required at a specified point in their flight training.

Attrition – Loss of respondents during a single-case time-series intervention study can produce artificial effects if that loss is systematically related to the experimental conditions. In this study, attrition occurred, but the effect was negated by the fact that more than the minimum number of participants were recruited to begin the study. Attrition would be a problem regardless of the design used if the number of participants fell below the minimum of three recommended by Kratochwill et al. [3].

Testing – Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with an intervention effect. In this study, "testing" (or learning) had the effect of reducing the intervention effect on the second introduction of the intervention in the combined design. This would suggest that there is a potential problem that "testing" may reduce the intervention effect to a level that is not clearly observable and/or statistically significant for the combined design.

Instrumentation – The conditions or nature of a measure might change over time in a way that could be confused with an intervention effect. For both the combined design and the MBD, the flight sessions were of a short duration to prevent other factors, such as fatigue, from being confused with the intervention effect. Confounding factors would also have been observed during the baseline measurements.

Additive and Interactive Threats to Internal Validity – The impact of a threat can be added to that of another threat or may be moderated by levels of another threat. Both the combined design and the MBD negate this threat by the replication of the intervention phase at different points in time.

All of the above threats to internal validity were negated by both the combined design and the MBD, so there was no advantage in using the combined design in this study.

B. External Validity

Single-subject designs are frequently criticized for their limited external validity, but this is usually aimed at studies involving single participants. In both the combined design and the MBD, the intervention is introduced to more than one individual, which improves the external validity. In this study, the intervention has an effect across several diverse participants from a particular flight training program. The student participants were not selectively chosen and could therefore be considered to be typical of any collegiate flight training program training students on technically advanced aircraft. The results of this study could therefore be generalized to students in similar flight training programs.

The combined design replicates the intervention effect across the participants at the second B phase. The two AB phases can be looked at as the SCD's equivalent to the between-group randomized block design. A correlation between the two AB phases would provide an improvement in external validity over the MBD. To determine the correlation between the replicated intervention effects, an ICC was calculated. Using SPSS one-way random effect model, the single measure ICC = .573, with 95% CI = (–1.170, –.140). These results show fair to good reproducibility [9]. This suggests that the combined design has an advantage over the MBD with respect to external validity.

C. Data Analysis

The data from the combined design and the MBD were analyzed both visually and statistically:

Visual Analysis: The combined design and the MBD were compared using the four visual features suggested by Whitehurst (under review): "level" refers to the mean score for the data within a phase; "trend" refers to the slope of the best-fitting straight line for the data within a phase; "variability" refers to the range or standard deviation of data about the best-fitting straight line; "immediacy of the effect" refers to the change in level between the last three data points in one phase and the first three data points of the next (see Fig. 4 and 7).

Level: Even though some differences are very small, both designs showed an increase in the overall mean between each phase A and phase B for all participants.

Trend: The overall trend for all participants in all phases for both designs is "downward," showing the expected "learning" effect.

Variability: The overall variability for each participant in both designs reduces as the phases progress, again showing the expected "learning" effect. The variability does not prevent the intervention effect being easily observable at the start of each B phase.

Immediacy of Effect: Both designs clearly showed immediacy of effect at each introduction of the intervention.

The four visual features show that the results of the visual analyses of the two designs both show evidence that would infer a causal relationship between the dependent and independent variables.

Statistical Analysis: The results of the HLM analyses for both designs are statistically significant. However, each of the estimated coefficients and the effect size for the combined design are smaller for the combined design than those of the MBD, which would suggest the expected "learning" occurred.

The data analysis from the two designs produced similar results, with both designs showing a significant degradation in flight performance for all participants at the introduction of the analog flight instrumentation.

D. Cost Analysis

This study used a PCATD to simulate flight conditions. Although PCATDs have been approved for use in flight training, they do not simulate the real airplane to the same degree as an advanced aviation training device (AATD) such as the Redbird FMX. The Redbird FMX is a full-motion AATD with wrap-around visuals and a fully enclosed cockpit. If funding had been available, an AATD or a flight simulator would have provided a more realistic environment for the research study. For the purposes of this cost analysis, I used the costs associated with the Redbird FMX, since this is probably an appropriate AATD to use in aviation studies. A basic cost calculation can be made to compare the cost of the MBD and combined MBD and ABAB design. The cost calculation is kept simple by basing it on the cost of the AATD Redbird FMX, the largest single cost item, and does not include any other costs, such as principal investigator (PI), co-PI, and/or assistant's time, which is required for the simulated flights, reviewing the videos, and data analysis.

Both the MBD and the combined design required only three participants. For the MBD, each participant flew 20 flight profiles. At 30 minutes per flight profile, the study required a total time of 3 × 20 × 0.5 = 30 hours; at a cost of $75 per hour for the Redbird FMX, this cost would be $2,250. For the combined design, each participant flew 50 flight profiles. At 30 minutes per flight profile, the study required a total time of 3 × 50 × 0.5 = 75 hours; at a cost of $75 per hour for the flight simulator, this cost would be $5,625, a 250% increase in cost.

VII. Conclusion

Both the combined design and the MBD have strong internal validity. The external validity of the combined design is superior to the MBD because of the replication of the AB phases. The results from both designs show that there is a significant degradation of flight performance for student pilots trained on aircraft equipped with digital flight instrumentation when they encounter analog flight instruments for the first time. However, the combined design also showed that although "learning" occurred during their first encounter with the different instrumentation, it was insufficient to prevent degradation of flight performance at a subsequent exposure to the analog instrumentation.

Although the study would suggest that the combined design improved the internal and external validity, quantifying this improvement is very difficult. Without a method of quantifying the improvement, it would prove very difficult justifying the very large increase in cost associated with using the combined design in the current economic climate. Further research is therefore required to determine a method of quantifying improvement in internal and external validity, to provide researchers with sufficient information to make a decision on which design is appropriate for their study.

References

[1] G. D. Garson, *Research Design. Overview*. Available: http://faculty.chass.ncsu.edu/garson/PA765/design.htm

[2] J. Ferrin, M. Bishop, T. Tansey, M. Frain, E. Swett, and F. Lane, "Conceptual and practical implications for rehabilitation research: Effect size estimates, confidence intervals, and power," *Rehab. Educ.*, vol. 21, no. 2, pp. 87–100, 2007.

[3] T. R. Kratochwill, J. Hitchcock, R. H. Horner, J. R. Levin, S. L. Odom, D. M. Rindskopf, and W. R. Shadish. (2010). Single-case designs technical documentation. What Works Clearinghouse, Washington, DC. [Online]. Available: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

[4] S. C. Hayes, "Single case experimental design and empirical clinical practice," *J. Consult. Clin. Psychol.*, vol. 49, no. 2, pp. 193–211, Apr. 1981.

[5] R. H. Horner, E. G. Carr, J. Halle, G. McGee, S. Odom, and M. Wolery, "The use of single subject research to identify evidence-based practice in special education," Exceptional Children, vol. 71, no. 2, pp. 165–179, 2005.

[6] G. Whitehurst and W. Rantz, "Digital training to analog flying: Should we teach new dogs old tricks?" *J. Aviation / Aerospace Education and Research*, vol. 21, no. *3*, pp. 17–22, *2012*.

[7] G. Whitehurst, "A comparison of two designs: Results from multiple baseline across subjects single case research and a between-group experiment" *Aviation Psychology and Applied Human Factors.* Under review.

[8] K. L. Wuensch. (2010). Inter-rater agreement. East Carolina University, NC [Online]. Available: http://core.ecu.edu/psyc/wuenschk/StatsLessons.html

[9] B. Rosner, *Fundamentals of Biostatistics*. Belmont, CA: Duxbury Press, 2005.

[10] S. Zhan and K. J. Ottenbacher, "Single subject research designs for disability research," *Disabil. Rehab.*, vol. 23, pp. 1–8, Jan. 15, 2001. Available: http://www.socialresearchmethods.net/kb/expblock.php

[11] C. G. Long and C. R. Hollin, "Single case design: A critique of methodology and analysis of recent trends," *Clin. Psychol. Psychother.*, vol. 2, no. 3, pp. 177–191, Oct. 1995.

[12] W. W. Fisher, M. E. Kelley, and J. E. Lomas, "Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs," *J. Appl. Behav. Anal.*, vol. 36, no. 3, pp. 387–406, Fall 2003.

[13] M. Hersen and D. H. Barlow, Single-case Experimental Designs: Strategies for Studying Behavior Change. New York: Pergamon, 1976.

[14] A. E. Kazdin, *Single-case Research Designs: Methods for Clinical and Applied Settings.* New York: Oxford University Press, 1982.

[15] C. H. Kennedy, *Single-case Designs for Educational Research*. Boston: Allyn and Bacon, 2005.

[16] D. Morgan and R. Morgan, *Single-case Research Methods for the Behavioral and Health Sciences*. Los Angeles: Sage, 2009.

[17] B. Parsonson and D. Baer, "The analysis and presentation of graphic data," in *Single Subject Research*, T. Kratchowill, Ed. New York: Academic Press, 1978, pp. 101–166.

[18] M. Furlong and B. Wampold, "Visual analysis of single-subject studies by school psychologists," *Psych. Sch.*, vol. 18, 80–86, Jan. 1981.

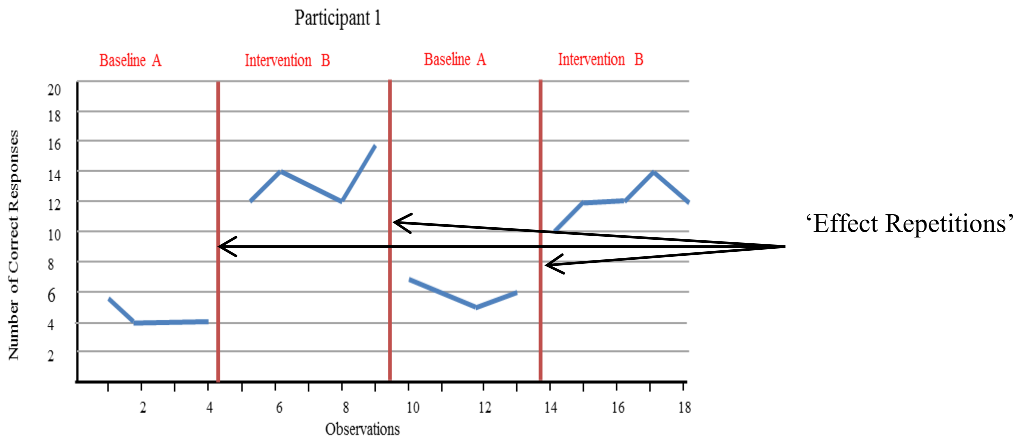
[19] S. Morley and M. Adams, "Graphical analysis of single-case time series data," *Br. J. Clin. Psychol.*, vol. 30, pp. 97–115, May 1991.

[20] W. Van den Noortgate and P. Onghena, "Combining single-case experimental data using hierarchical linear modeling," *Sch. Psychol. Q.*, vol. 18, no. 3, pp. 325–346, Fall 2003.

Geoffrey R. Whitehurst was born in Birmingham, England, in 1953. He received the B.S. degree in mathematics from London University, in 1974 and M.A. and Ph.D. degrees in Evaluation, Measurement and Research, from Western Michigan University, Kalamazoo, MI, in 2013.

From 2000 to 2004, he was the Chief Ground Instructor at the International Pilot Training Center at Western Michigan University. From 2004 to 2007 he instructed at the Qatar Aeronautical College, Doha, Qatar. Since 2008, he has been an Assistant Professor with the College of Aviation, Western Michigan University, Kalamazoo. He is the author of one book, and several articles. His research interests include human factors and how cockpit technology is impacting flight training.

SCD - ABAB (Horizontal)



A-B-A-B Design

Basic MBD - A B
A B (Horizontal/Vertical)
A B

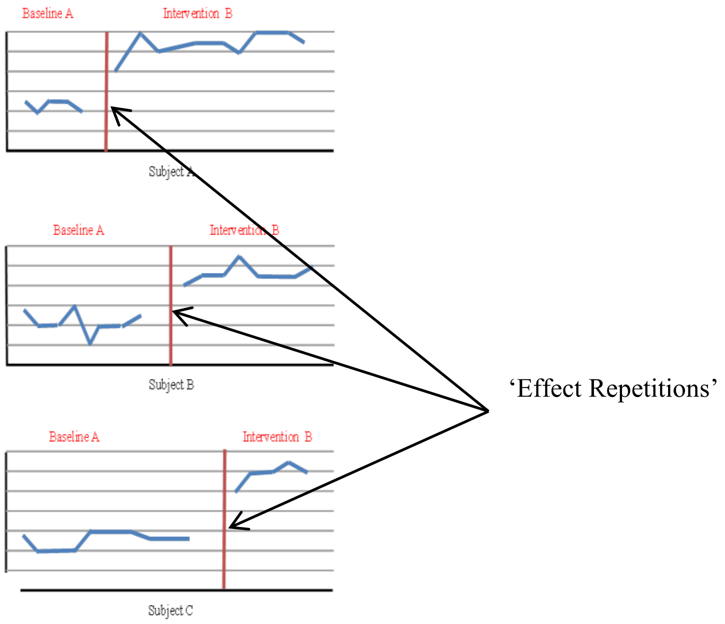


Figure 4. Multiple-baseline Across Three Subjects, basic A-B Design

Fig. 1. Visual representation of designs.

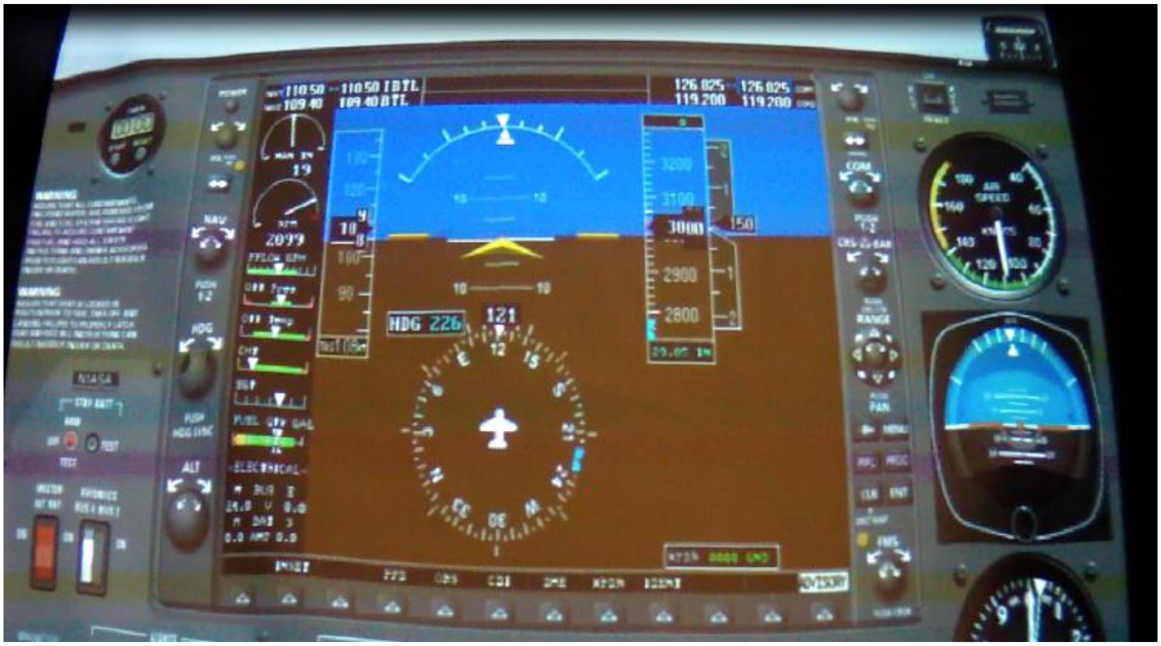


Fig. 2. *Digital flight instrumentation.*

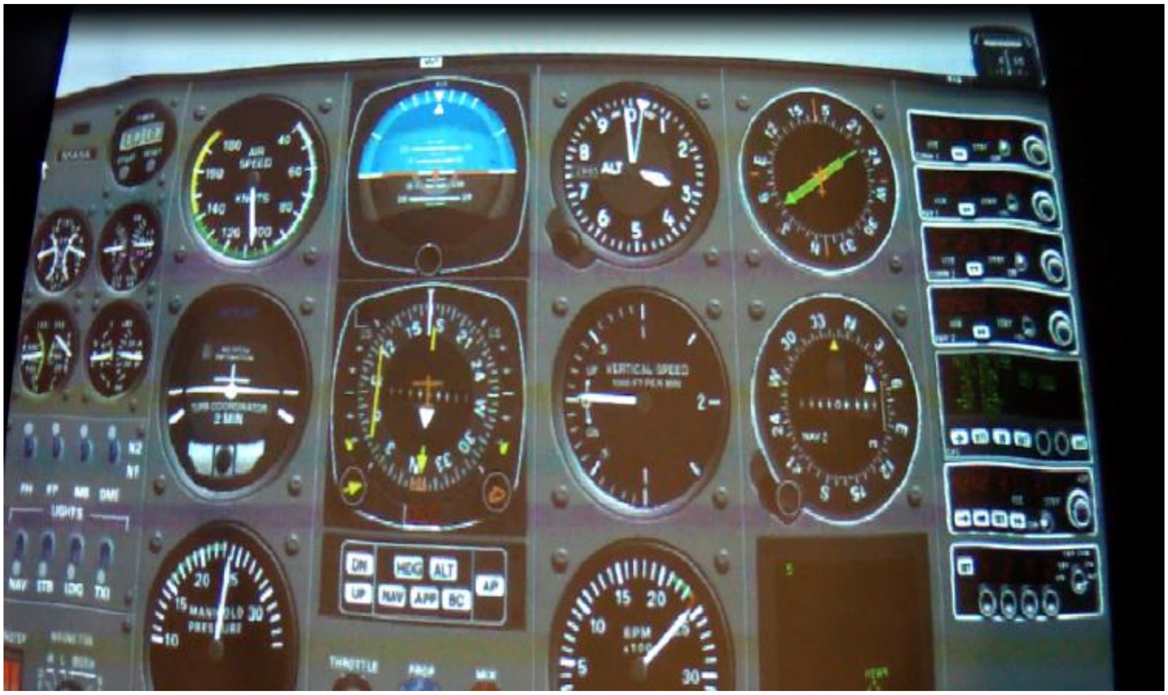


Fig. 3. Analog flight instrumentation.

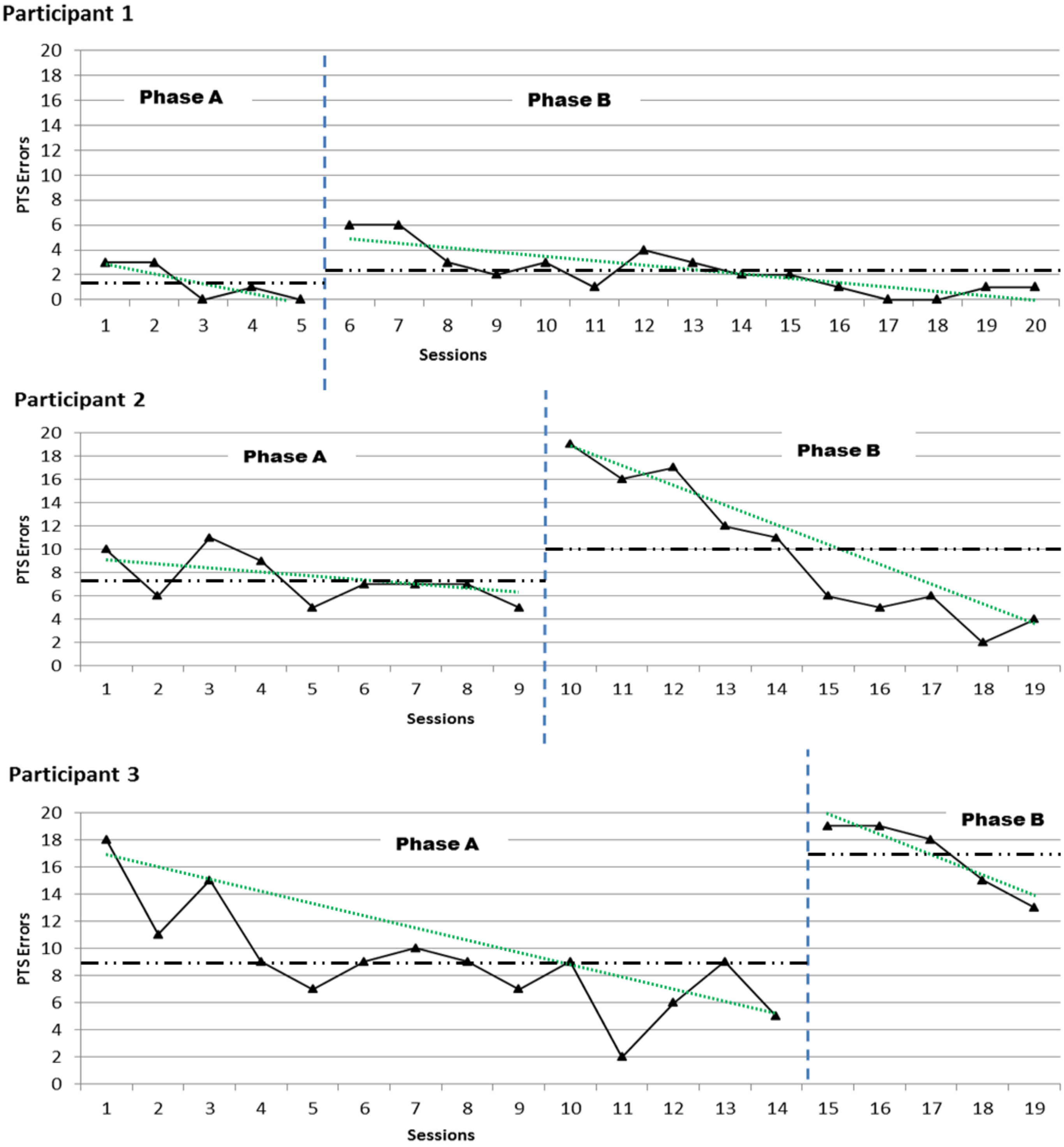
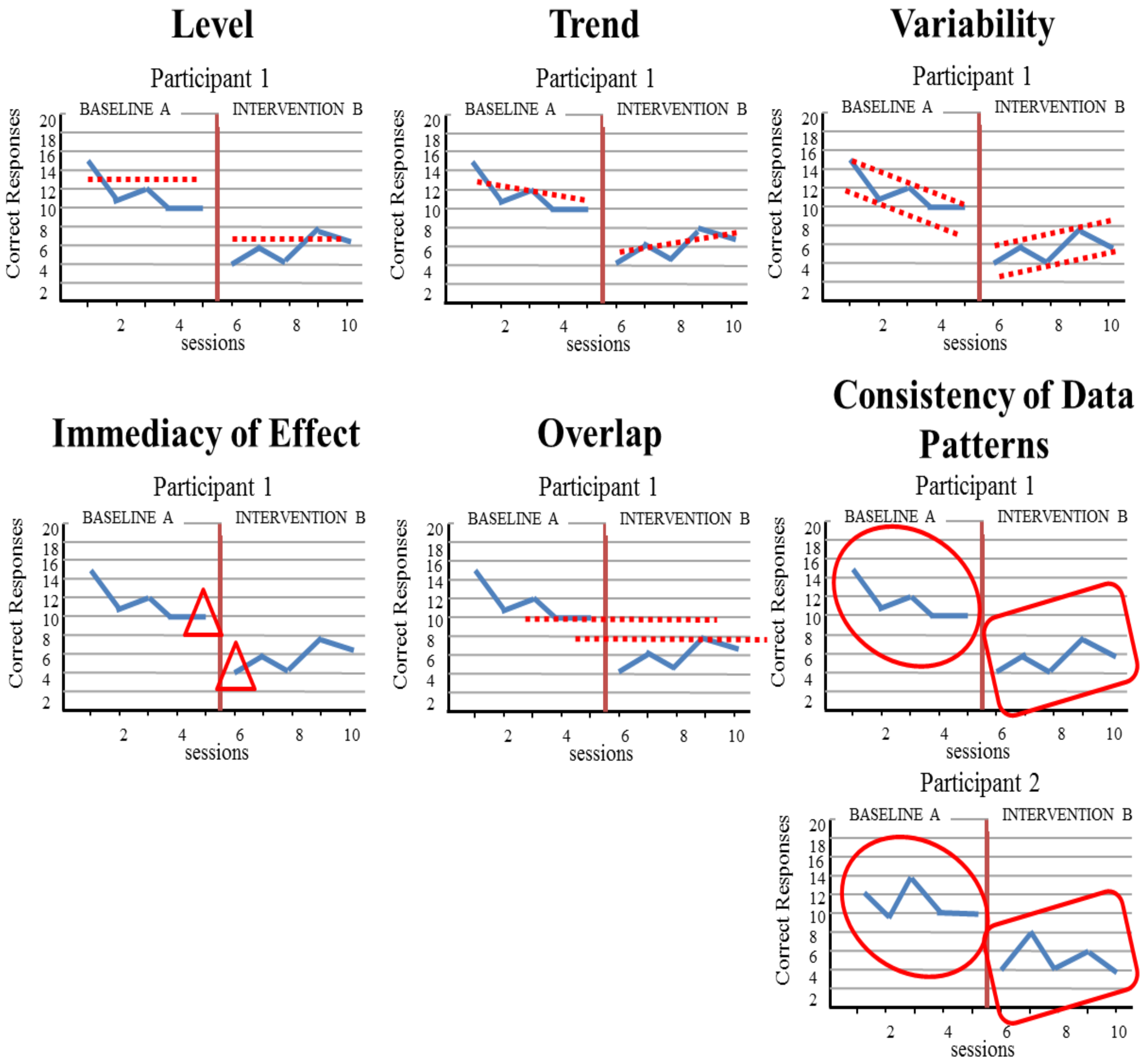


Fig. 4. *Multiple baseline across subjects research design with overall mean and trend.*



Level

Trend

Variability

Immediacy of Effect

Overlap

Consistency of Data Patterns

Fig. 5. *Visual analysis features.*

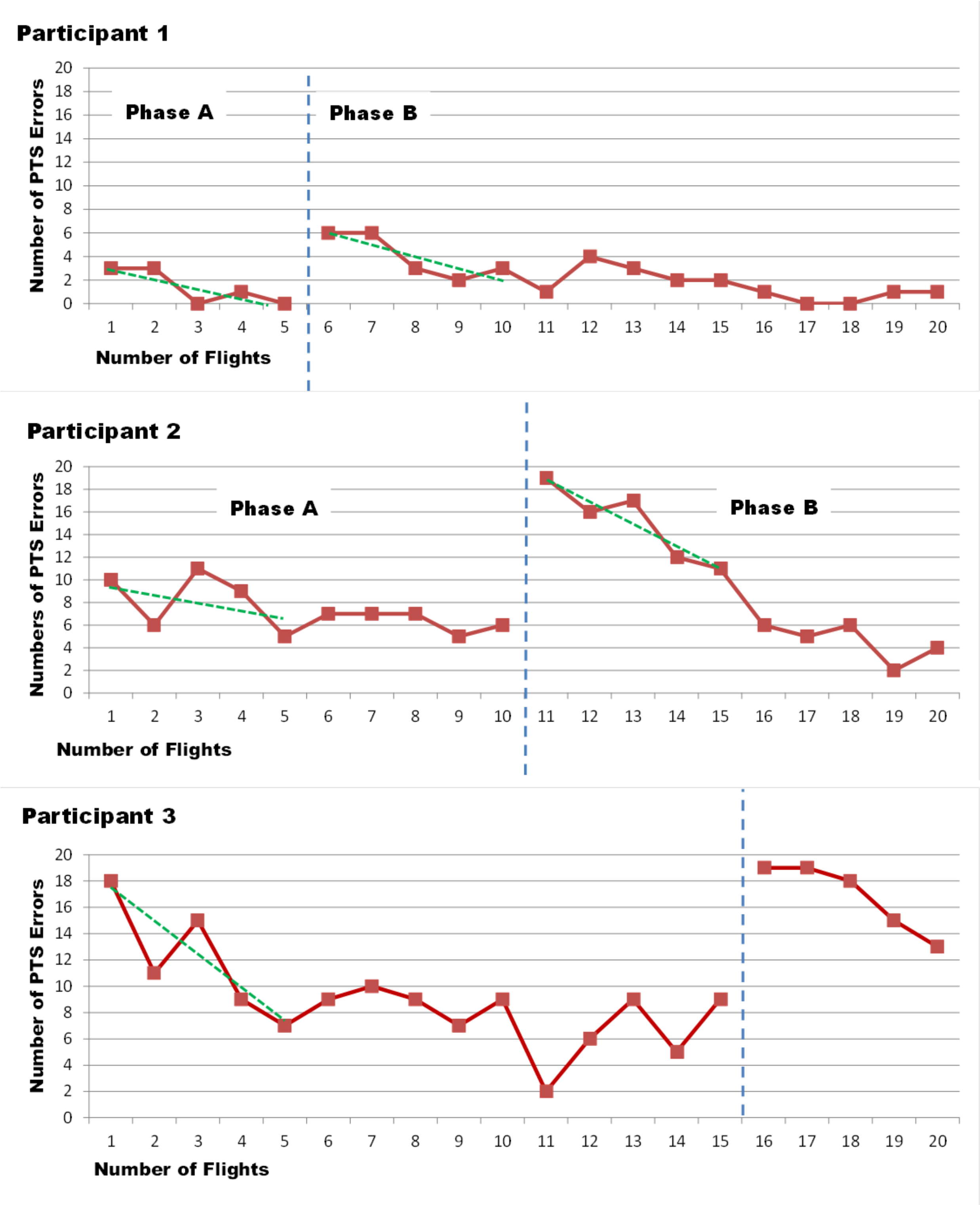


Fig. 6. *Graphed data for all participants with trend lines leading to intervention.*

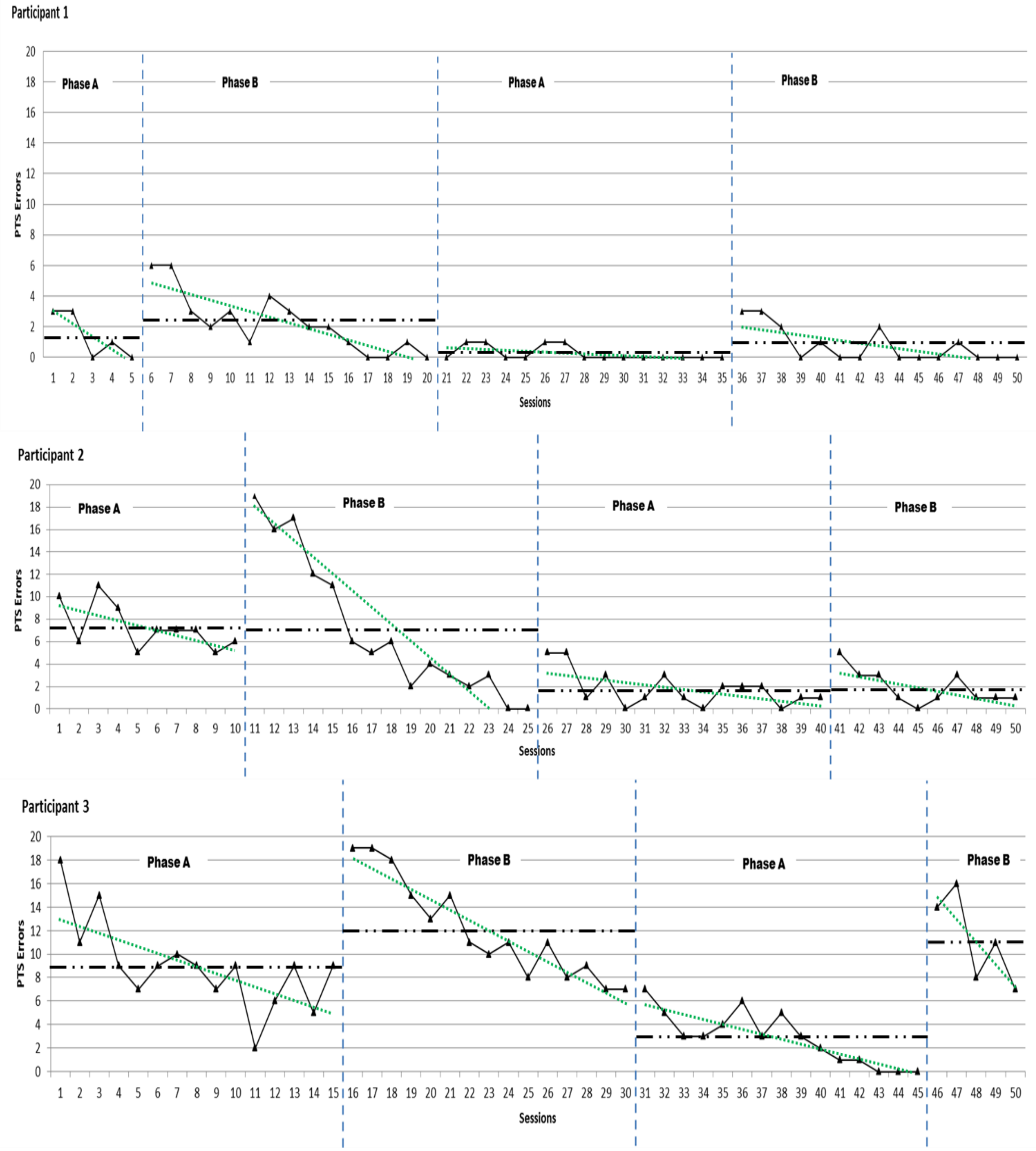


Fig. 7. *Graphed data for combined design with overall mean and trend lines.*

Table I
Visual Analysis Results for the Combined Design

Phase	Participant	Mean	SD
A1	1	1.40	1.52
	2	7.30	2.06
	3	2.87	2.23
B1	1	2.27	1.94
	2	7.07	6.03
	3	12.07	4.23
A2	1	0.27	0.46
	2	1.80	1.61
	3	9.00	3.82
B2	1	0.80	1.15
	2	1.90	1.52
	3	11.20	3.83

Table II
Visual Analysis Results for the MBD

Phase	Participant	Mean	SD
A1	1	1.40	1.52
	2	7.30	2.06
	3	2.87	2.23
B1	1	2.27	1.94
	2	7.07	6.03
	3	12.07	4.23

Table III
Fixed and Random Effects for Combined Design

Effects	Estimate	Standard Error	*df*	*t*	χ^2	Pr > |*t*|	Var. Comp.
				Value			
Fixed							
Intercept (γ_{00})	3.351765	2.240330	2	1.496		0.273	
Condition (γ_{10})	2.589803	0.637393	146	4.063		<0.001	
Random							
Level 1 (u_{0j})		3.80254	2		99.439	<0.001	14.459 (τ_{00})
Level 2 (r_{ij})		3.85188					14.837 (σ^2)

Table IV
Fixed and Random Effects for MBD

Effects	Estimate	Standard Error	*df*	*t*	χ^2	Pr > |*t*|	Var. Comp.
				Value			
Fixed							
Intercept (γ_{00})	5.431867	3.195408	2	1.700		0.231	
Condition (γ_{10})	3.502933	1.040230	56	3.367		0.001	
Random							
Level 1 (u_{0j})		5.39892	2		87.622	<0.001	29.148 (τ_{00})
Level 2 (r_{ij})		5.68614					13.588 (σ^2)