12-2013

Single Case Research Designs: An Effective and Efficient Methodology for Applied Aviation Research

Geoffrey R. Whitehurst

*Western Michigan University, geoff.whitehurst@hotmail.com*

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the [Aerospace Engineering Commons](https://scholarworks.wmich.edu/aerospace), and the [Education Commons](https://scholarworks.wmich.edu/education)

**Recommended Citation**


https://scholarworks.wmich.edu/dissertations/225

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
SINGLE CASE RESEARCH DESIGNS: AN EFFECTIVE AND EFFICIENT METHODOLOGY FOR APPLIED AVIATION RESEARCH

by

Geoffrey R Whitehurst

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the degree of Doctor of Philosophy Educational Leadership, Research and Technology Western Michigan University December 2013

Doctoral Committee:

Jessaca K. Spybrook, Ph.D., Chair
Edward B. Applegate, Ph.D.
William G. Rantz, Ph.D
The purpose of this three-paper dissertation is to examine the potential of single case research designs (SCDs) as an appropriate and efficient experimental design for use in applied aviation research. In the current environment of dwindling resources, funding for experiments requiring large sample sizes, a normal requirement for between-group designs, is becoming difficult to find. However, the need to improve safety within the aviation industry is an ongoing requirement, especially as advances in technology continue have an impact on how the industry operates. SCDs are experimental designs that require very few participants and therefore have the potential to save time and associated costs.

The first paper reviews published articles in three prominent journals to determine the types of experimental and quasi-experimental designs commonly used in applied aviation research and to compute a post hoc statistical power analysis for each experiment. The review shows that between-group experimental and quasi-experimental designs dominated applied aviation research and most designs lacked statistical power to detect medium and small effect sizes. SCDs were introduced as efficient alternative designs for many applied aviation studies.
The objective of the second paper is to examine if the results from an SCD produced similar findings to those in the between group designs. To do this, a between-group experiment was replicated using a SCD. The results from the SCD and between-group experiment were similar. However, a cost analysis suggests the multiple baseline design (MBD), the specific type of SCD used, is more cost effective.

The purpose of paper three is to compare two different types of SCDs, the MBD and a combined design, which combined the MBD with a standard SCD, known as the ABAB design. An applied aviation experiment was replicated to compare the two designs in terms of, internal and external validity, visual and statistical results, and cost. The comparison suggests that the results were similar, and that internal and external validity may be improved by the replication of phases in the combined design. However, the improvements came with a considerable increase in cost of resources and time.
ACKNOWLEDGMENTS

I am extremely grateful for my family. To my wife Cheryl—none of this would possible without her love, support, encouragement, and patience while I spent endless hours on the computer. To my daughters Samantha and Lauren who provided support and encouragement in their own individual ways.

I would like to express my sincere gratitude to the chair of my dissertation, Dr. Jessaca Spybrook. I will always be grateful for your time, support, and enthusiasm about the dissertation. I would also like to acknowledge the other members of my dissertation committee: Dr. Edward Applegate and Dr. William Rantz. I appreciated your insights and was grateful for the way in which the entire committee worked together to support and assist me during the dissertation process.

I am also glad to have this formal opportunity to thank my colleagues who have encouraged me in my professional development.

Geoffrey R Whitehurst
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................ ii

LIST OF TABLES .................................................................................................................. ix

LIST OF FIGURES ................................................................................................................. x

CHAPTER

I. INTRODUCTION .................................................................................................................. 1

   Background of the Problem ............................................................................................... 1

   Statement of the Problem ................................................................................................. 2

   Single Case Research Designs ......................................................................................... 3

      Types of Design .............................................................................................................. 3

      Types of Analysis .......................................................................................................... 5

   Dissertation Format and Related Purposes of the Three Studies ................................ 8

   Paper 1 ............................................................................................................................ 8

      Research Objectives ..................................................................................................... 8

   Paper 2 ............................................................................................................................ 9

      Research Objectives ..................................................................................................... 9

   Paper 3 ............................................................................................................................ 10

      Research Objectives ..................................................................................................... 10

   Significance of the Research ......................................................................................... 10

   References ....................................................................................................................... 11
II. DWINDLING RESOURCES: THE USE OF SINGLE CASE RESEARCH DESIGNS AS A PARSIMONIOUS ALTERNATIVE FOR APPLIED AVIATION RESEARCH ................................................................. 14

  Background ........................................................................................................ 14

  Purpose ............................................................................................................. 18

  Aviation Journals ............................................................................................. 18

    Article Classification ....................................................................................... 19

    Journal Reviews ............................................................................................ 20

    Power Analysis .............................................................................................. 23

    Single Case Research Designs ...................................................................... 25

    Data Analysis ................................................................................................. 31

    Single Case Multiple Baseline on Subjects Example .................................... 32

    Quasi-Experimental: Two Interventions One-Group
    Pretest/Posttest Design: Original Study ....................................................... 33

    Single Case Multiple Baseline on Subjects Design ....................................... 34

  Discussion ......................................................................................................... 35

  References ......................................................................................................... 37

III. A COMPARISON OF TWO DESIGNS: THE RESULTS FROM MULTIPLE BASELINE ACROSS SUBJECTS SINGLE CASE RESEARCH AND A BETWEEN-GROUP EXPERIMENT ......................... 40

  Literature Review ............................................................................................ 40

  Purpose ............................................................................................................. 42

  Research Methodology .................................................................................... 42
Table of Contents—Continued

CHAPTER

Research Design.................................................................................................................. 42
Participants............................................................................................................................. 45
Method.................................................................................................................................. 46
Apparatus ............................................................................................................................... 47
Flight Patterns ...................................................................................................................... 48
Data Collection .................................................................................................................... 48

Data Analysis ....................................................................................................................... 49
Visual Analysis ..................................................................................................................... 50
Interrater Reliability.............................................................................................................. 53
Statistical Analysis ................................................................................................................ 54
Model Selection .................................................................................................................... 56
Calculation of Overall Level Change (LCO\textsubscript{Overall}) ...................................... 57
LCO\textsubscript{Overall} Test Statistic ...................................................................................... 58
Overall Standardized Effect Size ......................................................................................... 59

Results .................................................................................................................................. 60
Interrater Reliability ............................................................................................................. 60
Visual Analysis ..................................................................................................................... 60
Statistical Analysis ................................................................................................................ 63
Discussion .............................................................................................................................. 64
Visual Analysis ..................................................................................................................... 64
<table>
<thead>
<tr>
<th>CHAPTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Analysis – Comparison of Findings ........................................ 67</td>
</tr>
<tr>
<td>Advantages of Single Case Designs in This Context .................................... 67</td>
</tr>
<tr>
<td>Cost Analysis .................................................. ........................................ 68</td>
</tr>
<tr>
<td>Disadvantages of Multiple Baseline Single Case Designs in This Context ............ 69</td>
</tr>
<tr>
<td>Conclusion .......................................................... ........................................ 70</td>
</tr>
<tr>
<td>References .......................................................... ........................................ 71</td>
</tr>
<tr>
<td>IV. THE COST OF INCREASED VALIDITY: COMBINING A MULTIPLE BASELINE DESIGN WITH AN ABAB DESIGN ......................................................... 74</td>
</tr>
<tr>
<td>Literature Review .......................................................... ........................................ 74</td>
</tr>
<tr>
<td>Threats to Internal Validity in SCDs .......................................................... ........................................ 75</td>
</tr>
<tr>
<td>Threats to External Validity in SCDs .......................................................... ........................................ 78</td>
</tr>
<tr>
<td>Purpose .......................................................... ........................................ 79</td>
</tr>
<tr>
<td>Research Methodology .......................................................... ........................................ 79</td>
</tr>
<tr>
<td>Research Design .......................................................... ........................................ 79</td>
</tr>
<tr>
<td>Participants .......................................................... ........................................ 82</td>
</tr>
<tr>
<td>Method .......................................................... ........................................ 83</td>
</tr>
<tr>
<td>Dependent Variable .......................................................... ........................................ 84</td>
</tr>
<tr>
<td>Apparatus .......................................................... ........................................ 85</td>
</tr>
<tr>
<td>Flight Patterns .......................................................... ........................................ 85</td>
</tr>
<tr>
<td>Data Collection .......................................................... ........................................ 86</td>
</tr>
</tbody>
</table>
### Table of Contents—Continued

#### CHAPTER

**Interrater Reliability** ................................................................. 86

**Data Analysis** ............................................................................. 87

**Visual Analysis** ........................................................................ 87

**Statistical Analysis** ................................................................. 91

**Results** ..................................................................................... 94

**Interrater Reliability** ................................................................. 94

**Visual Analysis** ........................................................................ 95

**Statistical Analysis – HLM** ..................................................... 98

**Discussion** ................................................................................. 100

**Internal Validity** ....................................................................... 101

**External Validity** ....................................................................... 103

**Data Analysis** ........................................................................... 104

**Cost Analysis** .......................................................................... 106

**Conclusion** ................................................................................ 107

**References** ................................................................................ 108

#### V. CONCLUSION ......................................................................... 111

**Review and Summary of Chapters II, III, and IV** ......................... 111

**Summary of Chapter II** ........................................................... 111

**Summary of Chapter III** ........................................................... 113

**Summary of Chapter IV** ........................................................... 114
Table of Contents—Continued

CHAPTER

Significance of Findings ................................................................. 116
Recommendations ........................................................................... 117
Future Research .............................................................................. 119
References ....................................................................................... 120

APPENDICES

A. PTS Observation Form ................................................................. 122
B. Pre-Flight Instructional Script ...................................................... 125
C. PCATD Engine Power and Throttle Settings for Airspeed .......... 128
D. Flight Pattern Narration ............................................................... 130
E. Human Subjects Institutional Review Board Letter of Approval .... 139
LIST OF TABLES

2.1 Journal Publication Dates ........................................................................................................ 19
2.2 Article Classification .................................................................................................................. 21
2.3 Categories of Designs for Single Intervention.......................................................................... 22
2.4 Categories of Designs for Multiple Interventions ................................................................. 23
2.5 Type of Statistical Test and Associated Effect Sizes ............................................................... 25
3.1 Summary of Values From Regression Required for Computation of Weighted Overall Level Change ......................................................................................................................... 64
4.1 Visual Analysis Results for the Combined Design ................................................................. 97
4.2 Visual Analysis Results for the MBD ....................................................................................... 98
4.3 Fixed and Random Effects for Combined Design ................................................................. 99
4.4 Fixed and Random Effects for MBD ...................................................................................... 100
LIST OF FIGURES

2.1 AB design................................................................................................................. 27
2.2 ABA design............................................................................................................... 29
2.3 ABAB design............................................................................................................ 29
2.4 Multiple baseline across three subjects, basic AB design....................................... 30
3.1 Digital flight instrumentation.................................................................................... 43
3.2 Analog flight instrumentation.................................................................................... 43
3.3 Graphed multiple baseline design data for three participants................................. 51
3.4 Visual analysis features............................................................................................ 53
3.5 Durbin-Watson Test statistic $d$................................................................................ 57
3.6 Graphed data for all participants with trend lines leading to intervention............... 61
4.1 Visual representation of designs................................................................................ 76
4.2 Digital flight instrumentation.................................................................................... 81
4.3 Analog flight instrumentation.................................................................................... 81
4.4 MBD with overall mean and trend lines.................................................................... 88
4.5 Visual analysis features............................................................................................ 89
4.6 Graphed data for all participants with trend lines leading to intervention............... 95
4.7 Graphed data for combined design with overall mean and trend lines............... 105
CHAPTER I
INTRODUCTION

Background of the Problem

Applied aviation research is extremely expensive and time-consuming. Generally applied aviation researchers use null hypothesis significance testing (NHST) with experimental and quasi-experimental designs to observe intervention effects. NHST examines which of two hypotheses—the first that there is no difference between groups (termed the null hypothesis or $H_0$), or the second, that there is a difference between groups (termed the alternative hypothesis or $H_1$)—is apparently true (Jones & Sommerlund, 2007). Ferrin et al. (2007) described this model as one in which a researcher “calculates the test statistic, and if it is sufficiently large and the $p$-value is sufficiently small, the null hypothesis is rejected and the corresponding alternative hypothesis is accepted” (p. 87).

These designs typically require the researcher to recruit large numbers of participants to ensure adequate statistical power to detect the primary effect of interest. The power of a statistical test refers to the chance of a test detecting a difference, or effect of a particular magnitude, between the groups being analyzed, assuming there is a true difference between the groups. Investigations into studies in many research fields found a large percentage of the studies had low statistical power to detect an effect of a specific magnitude or neglected a power analysis entirely (Aguinis, Beaty, Boik, & Pierce, 2005;
Borkowski, Welsh, & Zhang, 2001; Ferrin et al., 2007; Ison, 2011; Jones & Sommerlund, 2007; Osborne, 2008).

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. However, in aviation, this control is often very limited due to resource availability and cost of equipment. Resource availability refers to the pool of candidates meeting the criteria of the study. In applied aviation research, the candidates typically are pilots and the criteria normally include such items as flight hours, qualifications/certifications, type ratings, and aircraft flown. Finding pilots that meet all the criteria can be challenging, especially since the researcher is often limited to a small geographic area because of budget constraints. Cost of equipment refers to the high cost of using the specialized equipment, such as flight simulators and/or aircraft, normally required for studies involving flight safety, an important area of continuous research in the field of aviation.

**Statement of the Problem**

The between-group methodology generally used by aviation researchers usually requires sample sizes, for acceptable statistical power, that are often not available due to budgetary constraints. Ison (2011), in his analysis of statistical power in aviation research, stated that in the body of aviation research he had reviewed, power was infrequently discussed or calculated, and aviation research studies are often underpowered. He went on to say that improvements in research design and data analysis
would provide for more complete, easier to understand, replicable, and meaningful research. Thus, it is critical for aviation researchers, and possibly other fields of research suffering from similar problems, to consider alternative research designs that require smaller sample sizes but still produce valid results.

**Single Case Research Designs**

As the name suggests, single case designs (SCDs) are designs in which single units are studied, similar to a within-subject design in the experimental design domain. However, a unit can be a single individual, or a group of individuals. The advantages of SCDs include simplicity of design, ease of use in multiple settings, and the need for only a small number of participants. They provide a methodological approach well suited to the investigation of individuals, single groups, or multiple groups (Kazdin, 1982). In many research fields, such as psychology, psychiatry, education, rehabilitation, social work, and counseling, SCDs are commonly used to answer questions about the effect of interventions on human performance and human behavior.

**Types of Design**

SCD is a broad term and, in fact, there are several different designs that are considered SCDs. The following are outlines of some of the major SCDs.

**AB design.** The simplest type of SCD is a two-phase AB design. During the first phase, or A phase, data are collected from observing the unit, at fixed-time intervals, prior to the introduction of the intervention. These data are used to establish a baseline for
comparison with the data collected after the introduction of the intervention, during the second phase, or B phase.

**ABA and ABAB designs.** The ABA design includes a return or reversal back to baseline, where the intervention is removed, to provide stronger evidence of whether the intervention effect is a result of the introduction of the intervention and not due to chance or extraneous factors. The ABAB design goes one stage further by adding a second intervention phase. Systematic replication of the phases further strengthens the conclusion that the intervention effect occurred only after the introduction of the intervention and improves the internal validity of the design.

**Multiple baseline design (MBD).** The MBD uses the basic AB design for several units simultaneously. All units begin the A phase at the same time, but the B phase is staggered to improve internal validity. Kratochwill and Levin (2010) state, “The internal validity of the design is strengthened through a staggering or sequential introduction of the interventions across time, with desired changes in the outcome measure occurring repeatedly and selectively with the successive intervention introductions” (p. 129). This type of design can be thought of as a simultaneous replication design and also strengthens the external validity of the study (Hayes, 1981; Koehler & Levin, 1998, 2000). The extent to which the intervention effect is similar across recipients helps to enhance the population external validity of the intervention (Bracht & Glass, 1968).

**Combined design.** A combined design is the combination of two SCDs, the ABAB design and the MBD. In this design, the ABAB design is used for several units
simultaneously instead of the AB design. This replicates the baseline and intervention phase for each unit.

**Types of Analysis**

Historically, the most popular method for the analysis of data for all types of SCDs is visual inspection (Bulté & Onghena, 2009). However, the accuracy and reliability of visual analysis have been questioned (DeProspero & Cohen, 1979). Morley and Adams (1991) recommended complementing visual analysis with a statistical analysis of the data, whenever possible.

**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 (Zhan & Ottenbacher, 2001). Then six features are used 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 (Fisher, Kelley, & Lomas, 2003; Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; Morgan & Morgan, 2009; Parsonson & Baer, 1978) (details on these features are provided in Chapter III).

The main problems with visual analysis are the lack of concrete decision rules, the requirement of a particular pattern of the data (e.g., stable baselines without a trend in the direction of the expected change), and the overlooking of small but systematic effects (Kazdin, 1982). The rules governing the visual analysis of SCDs are rather subjective, and research on visual analysis procedures has raised questions concerning the interrater
agreement in decisions concerning between-phase changes. The requirement for a stable baseline without a trend in the direction of expected change is often not realistic. Typically, baseline data have some variability that can make visual judgment of the direction or existence of trends very difficult without the use of trend lines. The problem of overlooking small but systematic effects is also heightened when the baseline data are variable, because any small but continuous change in the data would be difficult to identify.

**Statistical analysis.** The parametric statistical tests that are mostly used in group research, such as t tests and ANOVAs, are often inappropriate in SCDs, because they require assumptions about the data that are not typically met with SCDs (Bulté & Onghena, 2009). These assumptions include independence—no correlation between error terms or between independent variables and error; homogeneity of variances—the variance within each of the populations is equal; and normally distributed—required for statistical inference. Violations of assumptions can seriously influence Type I error rates, rejecting the null hypothesis when it is in fact true, and Type II errors rates, not rejecting the null hypothesis when in fact the alternate hypothesis is true, and can result in overestimation or underestimation of the inferential measures and effect sizes (Osborne & Waters, 2002). Tests such as an ANOVA are also inappropriate because it only compares the means of the two phases, ignoring any systematic changes that take place within a phase (Hartmann, 1974). Trends that may occur within or across phases are not explicitly included in the simple ANOVA model and are thus treated as error (Center, Skiba, & Casey, 1985).
Several statistical methods have been and continue to be developed for use with SCDs. A linear regression model is one proposed method (Gorsuch, 1983; Kaestner & Ross, 1974; Kelly, McNeill & Newman, 1973). Analytical methods that use this model include Last treatment day (LTD) (White, Rusch, Kazdin, & Hartmann, 1989); Center’s mean-only (Center M) and Center’s mean and trend difference (Center’s MT) (Center et al., 1985); Allison’s mean-only difference (Allison’s M) and Allison’s mean plus trend difference (Allison MT) (Allison & Gorman, 1993; Faith, Allison & Gorman, 1996); and Huitema trend and level change (Huitema, 2011). LTD compares data points at the end of the intervention phase predicted from two different regression lines: one data point predicted by an extended Phase A regression line, and the other by the Phase B regression line. Center-M, and Center-MT both test for phase differences by controlling for overall data trend. Allison-M and Allison-MT parallel the two Center techniques, but control for Phase A trend only. Huitema also uses linear regression models for effect size calculation, but also provides a level change and trend calculation. The Huitema technique is used in Paper 2 (Chapter III) for an ABAB design.

Combined designs are used far more infrequently than the other types of SCDs, and statistical analysis techniques for these designs have been limited to meta-analysis techniques. A meta-analysis treats the results of each individual case as a separate effect and seeks to statistically combine these values. Van den Noortgate and Onghena (2003) suggested that another analytical approach to meta-analysis is hierarchical linear modeling (HLM), where the data are considered as a two-level model with measurements
nested within individuals. The HLM model is used to analyze the data in Paper 3 (Chapter IV).

**Dissertation Format and Related Purposes of the Three Studies**

This dissertation is comprised of three individual research studies (Chapters II, III, and IV) that investigate the suitability of a SCD, namely the MBD, as an appropriate and efficient alternative design to the commonly used between-group designs.

**Paper 1**

In study one, a hand search of three prominent U.S. aviation journals was conducted to identify articles using experimental or quasi-experimental designs. A post hoc power analysis of these studies was then carried out to determine if small sample sizes, commonly found in applied aviation research, were producing underpowered studies in the experimental research designs being used. Single case research designs, rarely used in aviation research, were then described and suggested as an efficient alternate methodology to the between-group methodology generally used by aviation researchers. Results for study one are reported in Chapter II. Implications of the findings are also discussed in Chapter II and further integrated with studies two and three in Chapter V.

**Research Objectives**

The first paper addresses the following research questions:
1. Are sample sizes used in published research articles producing underpowered studies to detect an effect of a reasonable magnitude?
2. Is the MBD an appropriate design to replicate between-group experimental and quasi-experimental designs?

**Paper 2**

Study two proposed the MBD as an efficient alternative methodology to the between-group methodology commonly used in applied aviation research. A between-group study by Whitehurst and Rantz (2012), which investigated performance degradation in student pilot performance when transferring from flying digital flight instrumentation to flying analog flight instrumentation, was replicated to determine if a MBD could (a) produce similar statistical results, and (b) be more efficient by saving on resources required. Results for study two are reported in Chapter III. Implications of the findings are also discussed in Chapter III and further integrated with studies one and three in Chapter V.

**Research Objectives**

The second paper addresses the following research question:
1. In a replicated study, is the MBD a more efficient design than the between-group experimental design originally used, in terms of time and cost of resources?
In study three, two SCDs, the MBD used in study two and the combined design, were compared to determine if the combined design produced improved results in terms of internal validity, external validity, visual analysis results, statistical analysis results, and cost. The MBD and the combined design were used to replicate the between-group study (Whitehurst & Rantz, 2012), which investigated performance degradation in student pilot performance when transferring from flying digital flight instrumentation to flying analog flight instrumentation. Results for study three are reported in Chapter IV. Implications of the findings are also discussed in Chapter IV and further integrated with studies one and two in Chapter V.

**Research Objectives**

The third paper addresses the following research questions:

1. Does the increase in the systematic replication of phases provided by the combined design improve internal validity, external validity, and data analysis results?

2. What is the cost of the increase in the systematic replication of phases provided by the combined design?

**Significance of the Research**

SCD methodology is used extensively in many research fields, especially psychology and education. However, I found only one article (Rantz, 2010) published in
the three U.S. aviation journals during the periods reviewed, that used a SCD. The most common research methodology being used is between-group designs, which by design requires large sample sizes. Many researchers often struggle to find the large sample sizes required because of small pools of suitably qualified candidates, limited time, and budgetary constraints. SCDs offer an efficient and cost-effective design for many aviation studies; hence, it is critical to introduce and demonstrate the use of SCDs to the field of aviation.

References


CHAPTER II
DWINDLING RESOURCES: THE USE OF SINGLE CASE RESEARCH DESIGNS AS A PARSIMONIOUS ALTERNATIVE FOR APPLIED AVIATION RESEARCH

Background

Generally aviation researchers use null hypothesis significance testing (NHST) with experimental and quasi-experimental models to observe intervention effects. Garson (2010) 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. This method of inquiry has been widely adopted since its development in the early 20th century (Cohen, 1992; Sedlmeier & Gigerenzer, 1989). NHST investigates research problems by determining which of two alternatives—the first that there is a difference between groups (termed the alternative hypothesis or H1), or the second, that there is no difference between groups (termed the null hypothesis or H0)—is apparently true (Jones & Sommerlund, 2007). Ferrin et al. (2007) described this model as one in which a researcher “calculates the test statistic, and if it is sufficiently large and the p-value is sufficiently small, the null hypothesis is rejected and the corresponding alternative hypothesis is accepted” (p. 87). The test statistic is a quantity calculated from the sample data. Its value is used to decide
whether or not the null hypothesis is rejected. The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone, if the null hypothesis, $H_0$, is true.

It is the probability of wrongly rejecting the null hypothesis if it is in fact true. It is equal to the significance level of the test for which we would only just reject the null hypothesis. The $p$-value is compared with the actual significance level of our test and, if it is smaller, the result is significant. That is, if the null hypothesis were to be rejected at the 5% significance level, this would be reported as “$p < 0.05$.” Small $p$-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. It indicates the strength of evidence for, say, rejecting the null hypothesis $H_0$, rather than simply concluding “Reject $H_0$” or “Do not reject $H_0$” (Easton & McColl, 1997).

However, there is a history of controversy among researchers about the use of this methodology. An important issue that has been talked about for many years, but is now becoming more prominent, is statistical power. The power of a statistical test is defined as “the probability, given that the null hypothesis is false, of obtaining sample results that will lead to the rejection of the null hypothesis” (Coladarci, Cobb, Minium, & Clarke, 2007, p. 403). In other words, power refers to the chance of a test detecting a difference, or effect, between the groups being analyzed, assuming there is a true difference between the groups. Therefore, if a test has low statistical power to detect an effect of a given magnitude, the chance of the test detecting an effect is lower, which could mean that even
though there is an effect, it may not be detected. Many investigations into studies conducted in areas such as psychology, medicine, behavioral accounting, business, education, and also aviation found a large percentage of the studies had low statistical power or neglected power entirely (Aguinis, Beaty, Boik, & Pierce, 2005; Borkowski, Welsh, & Zhang, 2001; Ferrin et al., 2007; Ison, 2011; Jones & Sommerlund, 2007; Osborne, 2008).

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. However, in aviation, this control is often very limited due to resource availability. The problem with the small sample sizes is that studies may not be powered to detect a meaningful treatment effect. Jones and Sommerlund (2007) stated, “NHST is vulnerable to sample size, or rather NHST is vulnerable to coincidences, when small sample sizes are used and the effect size is small” (p. 225). This is a very serious problem for the aviation industry, since many of the studies have safety implications, and this lack of power in the statistical analyses may be preventing some significant results from being seen and acted upon (Ison, 2011). In other words, if studies do not have the power to detect effect sizes that are common in aviation, then there may be consequences for the industry, as effect sizes need only be small to have a very big safety implication for aviation.

One such study (Whitehurst & Rantz, 2012) involved the safety implications of current pilot training when using only Technically Advanced Aircraft (TAA) equipped with digital flight instrumentation. Based on anecdotal information from flight instructors
and flight examiners, the investigators were concerned about performance degradation when TAA trained pilots first encounter analog flight instrumentation common in older aircraft. Although TAA is now the predominant type of aircraft being produced by manufacturers, analog aircraft still far outnumber their TAA counterparts, especially in General Aviation. Due to lack of resources and the lack of suitable participants required to power a full study, only a feasibility study could be carried out. Although the feasibility study suggested there may be performance degradation of the TAA trained pilots when they first encounter analog flight instrumentation, a study with sufficient statistical power would be needed to validate the results. Without this validation, the study does not provide the statistical significance to justify any action by aviation authorities, or training providers, to correct the problem.

In an article on power analysis in aviation research, D. C. Ison (2011) stated,

> It is readily apparent that aviation research studies are often underpowered. . . . Considering that small sample sizes are common in aviation research, lamenting the need to increase sample size is not practical and provides no solution to aviation researchers. (p. 79)

To paraphrase Ison, aviation studies are not likely to be adequately powered to detect small effect sizes with the small sample sizes commonly found in these studies. However, large sample sizes are often impractical because (a) it is costly to do these studies and more people mean more money, and (b) there are not often enough people that meet the criteria for a study.

This would suggest that research into alternative methodologies for use in the field of aviation needs to be initiated. One such alternative methodology that does not
require large numbers of participants, and would therefore be more cost-effective, is single case design research. This methodology has designs that may be suitable for some aviation research studies and may offer a solution to the problem of limited sample sizes.

**Purpose**

This purpose of this paper is to introduce the single case research design as an efficient alternative design for appropriate studies in applied aviation research. This paper begins by reviewing three prominent aviation journals to identify studies using experimental research designs. Next, a power analysis of these studies is conducted to investigate if small sample sizes, commonly found in applied aviation research, are producing underpowered studies in the experimental research designs being used. Single case designs are then described as an alternate methodology suitable for small sample sizes, and an example is given of how a single case multiple baseline on subjects design could have been a suitable alternative design for use in this study.

**Aviation Journals**

The three prominent aviation journals in this study include (a) *International Journal of Applied Aviation Studies* published by the Federal Aviation Administration Academy, (b) *Journal of Aviation/Aerospace Education and Research* published by Embry-Riddle Aeronautical University, and (c) *Collegiate Aviation Review* published by University Aviation Association. These journals were selected because they publish
research specific to the field of aviation. See Table 2.1 for the publication dates for each journal reviewed.

Table 2.1

*Journal Publication Dates*

<table>
<thead>
<tr>
<th>Publication</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Aviation/Aerospace Education and Research (JAAER)</td>
<td>1995–2011</td>
</tr>
<tr>
<td>Collegiate Aviation Review (CAR)</td>
<td>1985–2011</td>
</tr>
</tbody>
</table>

**Article Classification**

The review of the three journals first classified the articles into two general categories, research or nonresearch. The research articles were then further categorized into experimental or nonexperimental designs, and, finally, the experimental designs were then categorized into true (randomized) experimental designs or quasi-experimental designs. The categorization was based on the following definitions.

**Research and nonresearch.** According to the federal regulations definitions (U.S. Department of Health and Human Services, 2013),

(d) *Research* (46.102) means a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge. Activities which meet this definition constitute research for purposes of this policy, whether or not they are conducted or supported under a program which is considered research for other purposes. For example, some demonstration and service programs may include research activities.
And according to the U.S. Department of Justice (2013), systematic investigation is “a predetermined method for answering certain questions or studying a specific program or topic” (p. 1).

For classification purposes, articles containing one or more research questions or hypotheses were classified as research articles; all other articles were classified as nonresearch.

**Experimental and nonexperimental.** Thompson and Panacek (2007) described nonexperimental studies as “purely observational and the results intended to be purely descriptive” (p. 18) and list the common nonexperimental designs as cross-sectional, case-control, before and after (retrospective), historical controls (retrospective), surveys/questionnaires, case series, and case report.

**True experimental and quasi-experimental.** Gribbons and Herman (1997) stated,

Among the different types of experimental design, there are two general categories:

- true experimental design: This category of design includes more than one purposively created group, common measured outcome(s), and random assignment. Note that individual background variables such as sex and ethnicity do not satisfy this requirement since they cannot be purposively manipulated in this way.

- quasi-experimental design: This category of design is most frequently used when it is not feasible for the researcher to use random assignment.

**Journal Reviews**

A hand search of the three journals was carried out to categorize articles. The article abstracts were used to initially divide the articles into two main categories,
research and nonresearch. The method section of each of the research articles was then used to determine if the article was true experimental, quasi-experimental, or nonexperimental. Intra-rater reliability was established by randomly selecting 25% of the journals and conducting a second hand search and classification of the articles published in them.

During the review period, the *International Journal of Applied Aviation Studies* (IJAAS) published 151 articles, which included 25 nonresearch, 80 nonexperimental, and 46 experimental articles. The *Journal of Aviation/Aerospace Education and Research* (JAAER) published 128 articles, which included 77 nonresearch, 45 nonexperimental, and 6 experimental articles. The *Collegiate Aviation Review* (CAR) published 191 articles, which included 49 nonresearch, 130 nonexperimental, and 12 experimental articles. Of the 64 articles using experimental designs, 30 studies (18 IJAAS, 5 JAAER, and 7 CAR) were categorized as true experimental designs and 34 studies (28 IJAAS, 1 JAAER, and 5 CAR) were categorized as quasi-experimental designs. The classification of the articles is given in Table 2.2.

### Table 2.2

*Article Classification*

<table>
<thead>
<tr>
<th>Publication</th>
<th>Articles</th>
<th>Nonresearch</th>
<th>Research</th>
<th>Nonexperimental</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Nonexperimental</td>
<td>True Quasi</td>
</tr>
<tr>
<td>IJAAS</td>
<td>151</td>
<td>25 (16.6%)</td>
<td>80 (53.0%)</td>
<td>18 (11.9%)</td>
<td>28 (18.5%)</td>
</tr>
<tr>
<td>JAAER</td>
<td>128</td>
<td>79 (61.7%)</td>
<td>45 (35.2%)</td>
<td>5 (3.9%)</td>
<td>1 (0.8%)</td>
</tr>
<tr>
<td>CAR</td>
<td>191</td>
<td>49 (25.7%)</td>
<td>130 (68.1%)</td>
<td>7 (3.7%)</td>
<td>5 (2.6%)</td>
</tr>
</tbody>
</table>
All studies with experimental designs were first categorized by the number of interventions used, then subcategorized by number of groups of participants, and, finally, subcategorized into either pretest/posttest or posttest-only designs. For pretest/posttest designs, measurement of the dependent variable is taken for all of the comparison groups both before and after the treatment group(s) receive the intervention. Posttest-only designs measure the dependent variable of all the comparison groups only after the treatment group(s) have received the intervention. See Tables 2.3 and 2.4 for the categories of experimental designs.

Table 2.3

Categories of Designs for Single Intervention

<table>
<thead>
<tr>
<th>Category (number of articles)</th>
<th>Journal (number of articles)</th>
<th>Number of participants</th>
<th>Power analysis for medium effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Experimental</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Two-group</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (6)</td>
<td>IJAAS (3)</td>
<td>80/41/27</td>
<td>.36/.24/no test</td>
</tr>
<tr>
<td>JAAER (2)</td>
<td></td>
<td>64/57</td>
<td>.50/.45</td>
</tr>
<tr>
<td>CAR (1)</td>
<td></td>
<td>19</td>
<td>.25</td>
</tr>
<tr>
<td>Posttest only (12)</td>
<td>IJAAS (5)</td>
<td>46/20/16/16/16</td>
<td>.34/.19/.15/.24/.84</td>
</tr>
<tr>
<td>JAAER (3)</td>
<td></td>
<td>40/27/21</td>
<td>.43/.24/.28</td>
</tr>
<tr>
<td><strong>Quasi-Experimental</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single-group</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (8)</td>
<td>IJAAS (8)</td>
<td>375/199/50/32/30</td>
<td>1.0/1.0/1.0/.87/.47</td>
</tr>
<tr>
<td>Posttest only (3)</td>
<td>IJAAS (3)</td>
<td>24/8/8</td>
<td>.70/.35/no test</td>
</tr>
<tr>
<td><strong>Two-group</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (4)</td>
<td>IJAAS (3)</td>
<td>90/34/17</td>
<td>.63/.81/.30</td>
</tr>
<tr>
<td>CAR (1)</td>
<td></td>
<td>30</td>
<td>.45</td>
</tr>
<tr>
<td>Posttest only (3)</td>
<td>IJAAS (2)</td>
<td>120/14</td>
<td>.78/.49</td>
</tr>
<tr>
<td>CAR (1)</td>
<td></td>
<td>38</td>
<td>.32</td>
</tr>
</tbody>
</table>
Table 2.4

**Categories of Designs for Multiple Interventions**

<table>
<thead>
<tr>
<th>Category (number of articles)</th>
<th>Journal (number of articles)</th>
<th>Number of participants</th>
<th>Power analysis for medium effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Experimental</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (1)</td>
<td>IJAAS (1)</td>
<td>16</td>
<td>.15</td>
</tr>
<tr>
<td>Posttest only (1)</td>
<td>IJAAS (1)</td>
<td>51</td>
<td>.74</td>
</tr>
<tr>
<td>Multiple-group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (3)</td>
<td>IJAAS (2)</td>
<td>30/24</td>
<td>.20/.15</td>
</tr>
<tr>
<td></td>
<td>CAR (1)</td>
<td>96</td>
<td>.74</td>
</tr>
<tr>
<td>Posttest only (7)</td>
<td>IJAAS (6)</td>
<td>100/90/50/45/26/1</td>
<td>1.0/.54/.28/.29/.16/.1</td>
</tr>
<tr>
<td></td>
<td>CAR (1)</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>.23</td>
</tr>
<tr>
<td><strong>Quasi-Experimental</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest only (7)</td>
<td>IJAAS (6)</td>
<td>30/20/12/8/5/3</td>
<td>.90/.95/.41/.23/.14/.1</td>
</tr>
<tr>
<td></td>
<td>CAR (1)</td>
<td>26</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.81</td>
</tr>
<tr>
<td>Two-group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (1)</td>
<td>CAR (1)</td>
<td>30</td>
<td>.45</td>
</tr>
<tr>
<td>Posttest only (8)</td>
<td>IJAAS (6)</td>
<td>104/80/31/24/20/1</td>
<td>.84/.72/.91/.38/.37/.2</td>
</tr>
<tr>
<td></td>
<td>JAAER (1)</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>CAR (1)</td>
<td>64</td>
<td>.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38</td>
<td>.32</td>
</tr>
<tr>
<td>Multiple-group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest/Posttest (1)</td>
<td>CAR (1)</td>
<td>87</td>
<td>.42</td>
</tr>
</tbody>
</table>

**Power Analysis**

In his study of statistical power in aviation research, Ison (2011) found that aviation research studies appeared to “fall short of the minimum desirable statistical power levels” (p. 82). However, Ison’s study included all articles in which a statistical
test was used. In this study, the area of interest was the statistical power levels of only the experimental and quasi-experimental designs identified in the hand search of the three journals. A post hoc power analysis of the statistical tests in each of the identified studies was conducted using G*Power 3.1 software. Based on the statistical test used, the software requires input of information from the study on the alpha level, the total sample size, whether one- or two-tailed for t tests, or how many groups for F tests. The statistical power for the test is then computed for whatever Cohen’s $d$ effect size is selected. Although there are other measures of effect size being used in other fields, Cohen’s $d$ is the only measure used/mentioned in the articles reviewed, which would suggest it is currently the standard effect size being used in applied aviation research.

An example of the process would be a study that uses a one-way ANOVA to analyze the main effects of the study. Assume the researcher selected an alpha level of .05 for his study and had two groups of 15 participants. The alpha level (.05), the total sample size (30), and the number of groups (2) would be entered into the G*Power 3.1 software, together with the effect size of interest. For this example, the medium effect size (.25), described by Cohen (1988) as perceptible to the naked eye, was selected. The calculated power for this example was low at .26, meaning there would be only a 26% chance that the study would show significant results if the effect size $d = .25$.

The G*Power 3.1 calculations for the studies of interest were conducted at small, medium, and large effect sizes, as outlined by Cohen (1988, 1992) (see Table 2.5). As a medium effect size is generally used when no prior research provides an alternate, the statistical power of the studies for medium effect sizes is given in Tables 2.3 and 2.4.
Table 2.5

*Type of Statistical Test and Associated Effect Sizes*

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Effect sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
</tr>
<tr>
<td><em>t</em> test (independent means)</td>
<td>0.20</td>
</tr>
<tr>
<td>One-way ANOVA</td>
<td>0.10</td>
</tr>
<tr>
<td>Chi-square – goodness of fit</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Adapted from Cohen (1992)

**Single Case Research Designs**

The history of single-subject research can be traced back to the earliest days of psychological science, when researchers such as Wundt, Ebbinghaus, and Pavlov used single subjects or small groups of subjects to make scientific advances in their fields of study. Today single case research designs are currently used to answer questions about human performance and human behavior in many areas of research, including psychology, psychiatry, education, rehabilitation, social work, counseling, and other disciplines. They provide a methodological approach well suited to the investigation of individuals, single groups, or multiple groups (Kazdin, 1982). The advantages of single case research include simplicity of design, ease of use in multiple settings, and the need for only a small number of participants. Aviation is one field that has not yet recognized the advantages offered by single case research designs, as there have been only two studies published, both by Rantz (Rantz, Dickinson, Sinclair, & Van Houten, 2009; Rantz & Van Houten, 2011), that used this methodology.
In the classic between-group comparison designs, inferences about treatment effectiveness are typically drawn by observing changes in the target behavior(s) among those receiving treatment, compared to those receiving no treatment or a comparison treatment (Kazdin, 2003; Nock, Janis, & Wedig, 2008). By contrast, in single-subject research designs, inferences about treatment effectiveness typically are drawn by observing changes in the dependent variable over time within the individual(s) when treatment is present compared to when it is absent (Rizvi & Nock, 2008). The two-phase (AB) single case design is very similar to the interrupted time-series design, the major difference being the nature of the unit that provides the data for analysis (Huitema, 2011, p. 369). The unit for single case designs is usually a single participant, whereas for interrupted time-series designs the unit is usually a group of participants.

There are many types of single-subject research designs that have varying degrees of complexity, and a brief description of these designs follows.

**The AB design.** This design is the most basic of the single case designs. All other single case designs are, essentially, variations of the AB design (Richards, Taylor, Ramasamy, & Richards, 1999). In the AB design, the researcher collects baseline data (A) and then implements the intervention (B) to determine the effect on the dependent variable. See Figure 2.1.
**Figure 2.1. AB design.**

*A phase.* A baseline period is necessary to gather data on the dependent variable before any intervention is applied. Without a baseline phase, there is no way of knowing whether the intervention had any true effect or whether it is responsible for any changes in the subject. In this sense, the baseline period has not only a descriptive function, but also a predictive function in that it is presumed to predict how the dependent variable would continue in the absence of the intervention (Barlow, Nock, & Hersen, 2009; Kazdin, 2003). If the dependent variable is unstable and vacillates widely before the intervention is applied, then it becomes increasingly difficult to demonstrate that the intervention has any effect. Kazdin (2001) described a stable rate as one in which there is little variability as well as a lack of a trend (or slope). A flat line is ideal, but a small amount of variability is more realistic and acceptable. The A phase data are collected...
over several observations. For a phase to qualify as an attempt to demonstrate an effect, the phase must have a minimum of three data points (Kratochwill et al., 2010).

**B phase.** Once a baseline period has been established, the intervention must be applied in a systematic and conscientious manner. Ideally, the only difference between the A and B phase is the addition of the intervention (see Figure 2.1). This discrimination allows for the most valid conclusion to be drawn. Similar to the A phase, the B phase data are collected over several observations to ensure a “stable” data path for comparison with the “prediction” of the data path based on the baseline data.

Kazdin (1982, p. 106) described a stable rate of performance as being characterized by the absence of a trend (or slope) in the data and relatively little variability in performance, whereas Sidman (1960) suggested that the determination of stability is usually based on the judgment, intuition, and experience of the investigator. With the no clearly definition of “stable” data or stability of performance, it is incumbent on investigators to define their means of determining when “stability” has been achieved. If stability cannot be determined, then the design is not suitable for the study.

**The ABA and ABAB designs.** In the ABA design, also known as the withdrawal design, the A phase and B phase previous described are followed by a second A baseline phase. Experimental control is increased when the intervention is subsequently withdrawn, so the baseline conditions are again in effect; the probability that the effect on the dependent variable is caused by chance, or extraneous factors, and not by the addition or withdrawal of the intervention has been significantly reduced (see Figure 2.2).
Figure 2.2. ABA design.

The ABAB design not only adds a second A phase to the basic AB design, but adds a second intervention B phase after the withdrawal (second baseline)—a phase to strengthen or validate the relationship between the dependent and independent variables (see Figure 2.3).

Figure 2.3. ABAB design.
The multiple baseline design. The multiple baseline design is the basic AB design replicated within the same study. Multiple baseline designs are used to collect data over two or more subjects (and usually three or more; Barlow & Hersen, 1984). Kratochwill et al. (2010) stated, “To Meet Standards a multiple baseline design must have a minimum of six phases with at least 5 data points per phase” (p. 16). In other words, a minimum of three subjects is required when using the basic AB design, with five data points in each phase. The end result appears, visually, as a series of AB designs stacked on top of one another (Richards et al., 1999) (see Figure 2.4).

*Figure 2.4. Multiple baseline across three subjects, basic AB design.*
In the basic multiple baseline design on subjects, the researcher takes repeated measures of baseline performance (A phase) concurrently on two or more baselines (subjects). When stable, predictable baselines are obtained across the subjects, the intervention (independent variable) is introduced (B phase) to one of the subjects. Data continue to be collected across all subjects. When the performance of the subject in the B phase stabilizes, the intervention is introduced to a second subject. The procedure is repeated until the intervention has been introduced to all of the subjects. A change in performance for each subject at the introduction of the intervention provides a convincing argument of a relationship between the dependent and independent variables.

To strengthen or validate the study further, a withdrawal (back to baseline) A phase may also be included for each subject. A change in performance for each subject at the introduction and withdrawal of the intervention provides a stronger argument of a relationship between the dependent and independent variables.

**Data Analysis**

Single case research designs were originally used in fields such as psychology, in which interventions usually produced effect sizes that could easily be detected by visual analysis. With the expansion of this methodology into other fields of research, visual analysis is no longer considered sufficient and several statistical analysis methods have been and continue to be developed. Although there are now several statistical analyses available, no one type of analysis has been accepted as the “gold” standard for any particular single case design.
**Single Case Multiple Baseline on Subjects Example**

The hand search of the three journals categorized 64 of the 470 published articles as using an experimental design. The specific research designs of interest were experimental and quasi-experimental designs with a single intervention for which repeated measures post-intervention would not produce a “testing” effect. Testing is defined by Kratochwill et al. (2010): “Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with the intervention effect” (p. 10) A multiple baseline on subjects single case research design may be a suitable alternate methodology for these types of designs.

There are 18 true (two-group) experimental studies with 6 pretest/posttest designs and 12 posttest-only designs. There are 18 quasi-experimental studies with 11 within-subject (one group)—8 pretest/posttest designs and 3 posttest-only designs, and 7 two-group designs—4 pretest/posttest designs and 3 posttest-only designs.

An example of how a study could have been conducted using a multiple baseline on subjects single case research design, one of the studies were selected. The study was selected because the results were not statistically significant, and the post hoc power analysis had shown statistical power below the recommended minimum of .80 for all effect sizes.

There follows a brief description of the original methodology and then a description of how the alternate methodology could have been used.
Quasi-Experimental: Two Interventions One-Group Pretest/Posttest Design: Original Study

This study by Webb, Estrada, and Athy (2009) investigated the effect of stroboscopic light at two different frequencies on motion sickness. This study conducted two experiments on two different types of vehicular motion using one group for each experiment. To simulate the specific type of vehicular motion, a Multi Axis Ride Simulator (MARS) was programmed with motion data collected from the actual vehicle using a tri-axial accelerometer. Outside stabilizing visual references were removed by surrounding the simulator with a black curtain, and a 750W strobe light provided the stroboscopic effect.

Data were collected before and after each MARS session using three instruments: Psychomotor Vigilance Task (PVT), Motion Sickness Questionnaire (MSQ), and Postural Balance Assessment (PBA). For analysis, the difference scores were calculated by subtracting the scores from the preadministration from scores of the postadministration.

The study used a repeated measures design, with each participant attending three experimental sessions one day apart over a 5-day period (i.e., Monday, Wednesday, and Friday). The independent variable was frequency of the stroboscopic light (no-strobe, 4 Hz, and 8 Hz). The participants experienced one of the three lighting conditions per session, and the order of presentation was pseudo-randomized among three possible orders.
When participants first arrived at the testing facility, they completed the PVT, MSQ, and PBA. Next, participants experienced the 20-minute session on the MARS. During the MARS session, participants read selected passages from a military novel and answered questions to induce retinal slippage. After completion of the MARS session, participants again completed the PVT, MSQ, and PBA.

The data were analyzed using a one-way repeated measures ANOVA over the three lighting conditions. There was no significant difference between the three lighting conditions for each of the measures. To obtain statistically significant results from the ANOVA analysis for a medium effect size, with statistical power .80, the sample size would have needed to be 28.

**Single Case Multiple Baseline on Subjects Design**

Substituting a single case multiple baseline on subjects design in this study requires only 3 participants instead of the 9 in the original one-group within subjects repeated measures design. All 3 participants would begin by experiencing 20-minute MARS sessions with no strobe light. These sessions would continue until the data across all three instruments were stable for each participant, with a minimum of five sessions required for each participant.

When stability is achieved, Participant 1 experiences the MAR sessions with the 4 Hertz (Hz) strobe lighting. The other participants continue the MARS sessions with no strobe lighting. When Participant 1 reestablishes stability of results while experiencing the 4 Hz strobe lighting, Participant 2 begins MARS sessions with the 4 Hz strobe
lighting. This procedure continues until Participant 3 achieves stability of results with the 4 Hz strobe lighting. Participant 1 would then have the 8 Hz strobe lighting introduced for the MARS sessions. Again, this procedure would be continued until all participants had reestablished stability of results with the 8 Hz strobe lighting. Subjecting the participants to multiple MARS sessions to induce sickness may be a Human Subjects Institutional Review Board ethics issue. However, subjecting the participants to only one MARS session for each lighting condition may not induce sickness and would therefore be insufficient to establish the validity of the effect of the strobe lighting.

The data would then be analyzed using a suitable statistical method for multiple baseline designs, such as Huitema (2011) mean and trend model.

**Discussion**

This study suggests that (a) very few of the experimental studies identified in the three aviation journals discussed statistical power, much less calculated it; and (b) even fewer studies mention effect sizes. This would suggest that sample sizes were not calculated a priori, but were probably a matter of convenience or necessity. Fortunately, for most of these studies, the effect size was sufficiently large that the lack of statistical power did not prevent the results from being significant. However, the lack of a priori calculations of statistical power and effect sizes would suggest that the significance of the results was obtained by chance, rather than by research design based on acquired knowledge.
Based on the post hoc sample size calculations for the three substitution examples, the sample sizes required are considerably larger than generally used in these studies. There would be an obvious reduction in the number of participants, the time required, and probably cost for these studies, if a MBD had been used in place of the original design. However, without replicating each of the studies and replacing the original design with a MBD, it would be impossible to state that the research conclusions would be the same. It is the intention of the author of this study to replicate a single-intervention two-group design experiment (Whitehurst & Rantz, 2012) using a multiple baseline design on subjects in order to compare research conclusions and to try and estimate if this methodology would also be more cost-effective.

If the same research conclusion can be reached sooner and using fewer participants, without giving up the statistical stability, then resources are saved. In his study of statistical power of articles in these three journals, Ison (2011) stated,

It is readily apparent that aviation research studies are often underpowered and neglect to provide critical components necessary to confirm the soundness of studies. If one considers a small effect size, there is only a slightly better than a 1 in 4 chance of detecting a difference. Considering a medium effect size, the average power was .685 which is still short of the generally acceptable .80 value. Only if considering large effect size, which it is important to note is “roughly twice as large as medium” (Cohen, 1962, p. 150), would researchers exceed the .80 threshold. (p. 79)

Given the reduction in the number of participants and time needed for the single-subject design, researchers may wish to consider the option of using the MBD over the traditional between-group design as an effective, and possibly a more cost-effective, alternative.
References


CHAPTER III

A COMPARISON OF TWO DESIGNS: THE RESULTS FROM MULTIPLE BASELINE ACROSS SUBJECTS SINGLE CASE RESEARCH AND A BETWEEN-GROUP EXPERIMENT∗

Literature Review

Although it has been only 110 years since Wilbur Wright made his historic 12-second flight over the sand dunes of Kitty Hawk, NC, aircraft now have the capability of flying more than 500 passengers for over 15 hours. The advances in aircraft technology that have made this possible have also made today’s aircraft safer than ever before. But with all the advances in technology, it is still the pilots and their knowledge and skills gained during flight training that is the biggest safety factor aboard any aircraft.

It was only 10 years ago that pilots gained their knowledge and skills in aircraft equipped with limited mechanical (analog) flight instrumentation. However, aircraft, and specifically the cockpit in which today’s students are taught how to fly, have changed considerably. Nowadays, many training aircraft are Technically Advanced Aircraft (TAA) equipped with advanced avionic displays, autopilots, global positioning systems (GPS), and, in many cases, moving map displays and flight management systems (FMS). These airplanes have the computing power, functionality, and automation capabilities

∗This paper is under review by the journal Aviation Psychology and Applied Human Factors.
previously found only on larger commercial aircraft (Hamblin, Gilmore, & Chaparro, 2006).

Since 2002, general aviation aircraft manufacturers, producing airplanes powered by small engines and propellers, have offered digital instrumentation in new aircraft (McDermott, 2005). Only recently have flight training providers begun to incorporate TAA in their training fleets in the belief that this would better equip pilots with the knowledge and skills required for today’s hi-tech commercial aircraft. Proponents of TAA believe that providing more information and automation will make flying safer. The logic is that the safety of commercial aviation has increased in part to similar technologies; therefore, general aviation should also benefit (Hamblin et al., 2006). The growing use of these aircraft presents unique challenges to the aviation infrastructure, since, according to AOPA Air Safety Foundation (2005), “the bulk of the existing 180,000-plus light GA (General Aviation) airplanes still use steam gauges.” The disproportionately large number of analog equipped aircraft remaining in general aviation suggests that transitions between digital and analog will become more numerous, especially as more and more pilots are trained in TAA aircraft without ever flying on “steam gauges,” the old mechanical analog flight instruments (Whitehurst & Rantz, 2011).

The results of research by Hamblin et al. (2006) suggest that the knowledge and skill needed to fly TAA are distinct from those needed to fly airplanes equipped with traditional mechanical gauges, and that flying traditional avionics is mostly unrelated to operating TAAs. This would suggest that the reverse would seem equally likely.
Currently, there is no requirement to obtain transition training between digital and analog instrumented aircraft specified in the Federal Aviation Administration regulations Title 14 part 61.31, (Federal Aviation Administration, 2012). Therefore, as the TAA training fleet continues to expand the potential for transitional incidents and accidents is likely to increase.

**Purpose**

The purpose of this paper is to replicate the between-group digital-to-analog flight instrumentation transition study by Whitehurst and Rantz (2012) using a multiple baseline design (MBD), and to compare the results of the two studies. A previous study (Whitehurst, in press) suggests that a MBD was an appropriate and efficient alternative to the group methodology commonly used. An example of how a MBD could have been used as an alternative design was included. However, the study fell short in terms of any empirical comparisons between the results from the two study designs. This study will correct that deficiency by replicating a between-group design study using a MBD, and comparing the two sets of results.

**Research Methodology**

**Research Design**

In the original (Whitehurst & Rantz, 2012) experimental design, the flight performance of a group of student pilots transitioning from flying digital flight instrumentation (DFI) equipped aircraft to flying analog flight instrumentation (AFI)
equipped aircraft was compared to a control group of student pilots who flew only DFI aircraft. The study was conducted as a feasibility study because the treatment and the control groups each had only 3 participants and therefore had very low statistical power (.06). The data were collected over a 4-week period, with each participant completing two 1-hour sessions flying a Personal Computer Aviation Training Device (PCATD), which was capable of emulating a Cessna 182 Glass, a DFI aircraft (see Figure 3.1), and a Cessna 182 Skylane RG, an AFI aircraft (see Figure 3.2). 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 (see Appendix A). For the first session, both groups flew a DFI configured aircraft. For the second session the control group again flew a DFI configured aircraft, whereas the treatment group flew an AFI configured aircraft. Each session was recorded electronically by the computer flight software, and visually by a video camera, to enable appropriate analysis.

Figure 3.1. Digital flight instrumentation.  Figure 3.2. Analog flight instrumentation.
In the current study, only one flight profile was flown each session and the order participants flew the sessions was randomized at the beginning of the study. The sessions were not flown concurrently, due to scheduling and equipment constraints. However, all participants completed each session, within 48 hours, before the next session could be flown by the first participant. The MBD requires data to be collected on all participants prior to any intervention to provide baseline flight performance data for each participant. Ideally, the baseline is expected to have little or no trend and to have little or no variability, thus giving “stable” data. However, in this study, the participants flew unfamiliar equipment and continued to “learn” during the study; therefore, both the baseline phase data and intervention phase data were expected to have a “downward” trend (fewer errors) and some variability. The expected “downward” trend and the reduction in the variability of the data in the baseline phase would show an improvement in performance, as a result of the participant becoming more familiar with the equipment and the environment. The expected “downward” trend and the reduction in the variability of the data in the intervention phase would show an improvement in performance, as a result of the participant improving in the assimilation of the information being presented in a different format.

When all participants achieved a “downward” trend and a reduction in variability in their flight performance flying the DFI aircraft, participant 1 began flying the AFI aircraft (intervention phase). There were no specified limits to the variability of the data for it to be considered “stable,” but limits were determined by the researcher based on experience and prior knowledge. Data from the original study and discussions with
certified flying instructors–instrument (CFIs) suggested that, at this stage of their training, participants achieving plus or minus 2 PTS errors of a downward trend line showed good consistency or “stability” of performance. Therefore, for this study, a participant’s performance was considered “stable” when they achieved plus or minus 2 PTS errors of a downward trend line for two or more continuous sessions.

The other participants continued flying the DFI aircraft until participant 1 again achieved stability of flight performance, but in the intervention phase. Participant 2 then began flying the AFI aircraft, participant 1 continued flying the AFI aircraft, and the other participants continued flying the DFI aircraft. This procedure was repeated until all participants were flying the AFI aircraft. Each phase required a minimum of five data points, even if the downward trend and reduced variability were achieved earlier, and a minimum of six phases (three participants) were required in order to meet evidence standards for SCDs suggested by Kratochwill et al. (2010) in their SCD technical documentation.

Participants

Participants were recruited from flight students in a 4-year university flight science degree program who met the following criteria: (a) must have completed their private certification, (b) must be within 15 flights of completing their instrument certification, and (c) must have never flown an aircraft equipped with analog flight instrumentation. These criteria were confirmed during an initial interview that also collected data on flight hours, actual and simulated, pilot-in-command hours and types of
aircraft flown. The 15 hours to instrument certification criterion was selected to ensure proficiency in instrument flying, but also to provide sufficient time to complete the research project before participants had 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 negating one of the requirements for participation.

Four student pilots who met study inclusionary criteria participated in the study; due to other commitments, one participant withdrew during the first intervention phase. The three remaining participants completed the study, providing the minimum six phases required to meet evidence standards. Only the data from the three participants who completed the study were used in the data analysis.

**Method**

To ensure consistency of instructions, before each flight a pre-flight instructional script (see Appendix B) was used to brief participants. These instructions highlighted the engine power and throttle settings and airspeeds required for each sector of flight (see Appendix C). Each participant flew the PCATD, emulating the DFI for the baseline phase and the AFI for the intervention phase. During each simulated flight, participants were asked to fly a radar-vectored flight pattern and to complete an instrument approach. Each flight was visually recorded for later analysis of the participant’s flight
performance. The dependent variable for measuring participant flight performance consisted of the number of times the aircraft deviated from the criteria listed in the Federal Aviation Administration’s Practical Test Standards (PTS) for instrument flight check rides. The criteria are: for heading, ±10°; for altitude, ±100 feet; and for speed, ±10 knots. A deviation beyond any one of the three limits was recorded as one PTS error.

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 170° 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 an 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.

**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, if required, as a back-up for analysis, data collection and interrater reliability checks.

**Flight Patterns**

In an effort to minimize any practice effects, four different radar-vectored flight patterns were used on a random basis (see Appendix D). By using vectored instrument flight patterns and not having system faults, the flight environment allowed for consistent flight performance. 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.

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.

**Data Collection**

Data were collected from the participants over a period of 8 weeks. The participants each flew one flight pattern per session, two or three times per week, based on their academic and flying schedules. The participants were randomly assigned to their order of participation and this order was the 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 that (a) each recording could be assessed by more than one rater, and (b) recordings could be stop and/or rewound to confirm accuracy of assessment. Each time a participant exceeded the limits set out in the Practical Test Standard for the Instrument Rating Certification, in altitude, speed or heading, a check mark was recorded against the participant on the assessment form. The dependent variable was the total number of check marks (errors) recorded during a flight pattern.

**Data Analysis**

Single case research designs are found predominantly in the social sciences and the type of data analysis most commonly used is visual analysis. In 2002, an initiative of the U.S. Department of Education’s Institute of Education Sciences (IES) created the What Works Clearinghouse (WWC) as a central and trusted source of scientific evidence for what works in education. In 2010, WWC assembled a panel of national experts in single case design (SCD) and analysis to draft SCD standards (Kratochwill et al., 2010) for use by WWC reviewers when reviewing SCD research papers. Kratochwill et al. set out rules for visual analysis and specified six features to assess the effects of the intervention within SCDs. However, methods of statistical analysis and standards that could be applied to them were not included in the documentation. Although not included in the WWC documentation, there are statistical analyses applicable for SCD. In this
study, both a visual and a statistical analysis were used to analyze the data. Comparisons were then made to the original study.

**Visual Analysis**

The oldest and still the most popular method for the analysis of MBD data is visual inspection (Bulté & Onghena, 2009). 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 (Zhan & Ottenbacher, 2001). 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 Figure 3.3). On the basis of these graphs, a judgment is reached about the reliability or consistency of intervention effects (Long & Hollin, 1995). This method of data analysis undoubtedly has some advantages, such as the speed of making the graphs, yielding conclusions, and deriving hypotheses (Parsonson & Baer, 1992); in many cases, however, these advantages do not outweigh the difficulties.

The main problems with visual analysis are the lack of concrete decision rules, the requirement of a particular pattern of the data (e.g., stable baselines without a trend in the direction of the expected change), and the overlooking of small but systematic effects (Kazdin, 1982). The accuracy and reliability of this method have been questioned, because there has often been a lack of agreement among judges (DeProspero & Cohen, 1979). Especially when there is variability within phases, both Type I and Type II error rates are elevated to unacceptable levels (Matyas & Greenwood, 1990).
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 are used 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 (Fisher, Kelley, & Lomas, 2003; Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; Morgan & Morgan, 2009; Parsons & Baer, 1978) (see Figure 3.4). The 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

Figure 3.3. Graphed multiple baseline design data for three participants.
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 (Furlong & Wampold, 1981), that is, to 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 (Kratochwill et al., 2010).
Inter-rater Reliability

There are a number of statistics which can be used to determine interrater reliability. The intraclass correlation coefficient (ICC) is an index of the reliability of the ratings 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 (Wuensch, 2010).

**Figure 3.4.** Visual analysis features.
The ICC is the ratio of the variance between subjects over the total variance

\[
\text{ICC} = \frac{MS_{BS} - MS_{WS}}{MS_{BS} + MS_{WS}}
\]

(1)

where

- \(MS_{BS}\) is the mean square error between subjects
- \(MS_{WS}\) is the mean square error within subjects

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 (Rosner, 2005).

**Statistical Analysis**

Morley and Adams (1991) recommended complementing visual analysis with a statistical analysis of the data, whenever possible. Several statistical tests have been developed for use with multiple baseline data, but unlike in group research designs, their use is the exception rather than the rule.

Huijema (2011) stated that several statistics are useful in analyzing data from a MBD and recommended three: (1) a measure describing overall level change (across the different series), \(LC_{\text{Overall}}\); (2) a test for overall level change, \(LC_{\text{OverallTest}}\); and (3) measures of overall effect size, \(ES_{\text{Overall}}\).

To compute the three statistics recommended by Huijema, the data first need to be fitted to one of four time-series regression models. The models differ in the number of parameters used to describe the intervention effects and the assumed nature of the errors
(independent or autocorrelated). The four parameter models focus on both level change and slope change, whereas the two parameter model focus on level change only. Autocorrelation among the errors of the model refers to errors measured at one time point being correlated to errors at an earlier time point and therefore not independent.

The four time-series regression models are:

For independent errors;

Four parameter model;

\[ Y_{jt} = \beta_{0j} + \beta_{1j}T_t + \beta_{2j}D_t + \beta_{3j}S_C_t + e_{jt} \]  \hspace{1cm} (2)

Two parameter model;

\[ Y_{jt} = \beta_{0j} + \beta_{2j}D_t + e_{jt} \]  \hspace{1cm} (3)

For autocorrelated errors;

Four parameter model;

\[ Y_{jt} = \beta_{0j} + \beta_{1j}T_t + \beta_{2j}D_t + \beta_{3j}S_C_t + \epsilon_{jt} \]  \hspace{1cm} (4)

Two parameter model;

\[ Y_{jt} = \beta_{0j} + \beta_{2j}D_t + \epsilon_{jt} \]  \hspace{1cm} (5)

where

\( Y_{jt} \) is the dependent variable score at time \( t \) for participant \( j \);

\( \beta_{0j} \) is the expected elevation on \( Y \) at time period zero for participant \( j \);

\( \beta_{1j} \) is the slope in the baseline phase for participant \( j \);

\( \beta_{2j} \) is the level change measured at time \( n_1 + 1 \) (difference between the predicted values of \( Y \) at \( n_1 + 1 \) using baseline data and intervention data) for participant \( j \);
\( \beta_{3j} \) is the change in slope from the baseline phase to the intervention phase for participant \( j \);

\( T_t \) is the value of the time variable \( T \) at time \( t \);

\( D_t \) is the value of the level-change dummy variable \( D \) (0 for the first phase and 1 for the second phase) at time \( t \);

\( SC_t \) is the value of the slope-change variable \( SC \) at time \( t \), defined as \( [T_t - (n_1 + 1)]D_t \);

\( e_{jt} \) is the error \( t \) for participant \( j \) at time \( t \);

\[ e_{jt} = \varphi_1 e_{j(t-1)} + u_{jt} \] where \( \varphi_1 \) is the lag-1 autoregressive coefficient; and

\( u_{jt} \) (the disturbance for participant \( j \) at time \( t \)) = \( Y_{jt} - \beta_{0j} + \beta_{2j} D_t + \varphi_1 e_{t-1} \)

or \( Y_{jt} = \beta_{0j} + \beta_{1j} T_t + \beta_{2j} D_t + \beta_{3j} SC_t + \varphi_1 e_{t-1} \).

It is assumed that \( u_{jt} \) is independent and normally distributed with mean zero and a common variance \( \sigma^2 \) for all time points.

**Model Selection**

The correct model is selected either by using prior knowledge or by fitting the data to the models and conducting model comparison tests. In this replicated study, the four-parameter model was selected based on knowledge from the original study. The original data showed an improvement in participants’ performance as the number of sessions progressed; this suggests a slope (trend) in the data and therefore the four-parameter model is required. The errors are assumed to be independent to simplify the initial analysis. The data are fitted to the selected model and a linear regression is used to
compute the parameters’ estimates and the Durbin-Watson statistic $d$ to check for autocorrelated errors.

To check if there is autocorrelation of errors, the Durbin-Watson (D-W) test can be used. The null hypothesis associated with the Durbin-Watson test statistic $d$ is:

$H_0$: The value of lag-1 autocorrelation among the process errors $= 0$

There is an upper and lower bound of the test statistic $d$ with associated areas for rejecting or accepting $H_0$ (see Figure 3.5).

The linear regression parameter estimates for each of the participants are then used to calculate the three statistics $LC_{Overall}$, $LC_{OverallTestl}$, and $ES_{Overall}$.

![Figure 3.5. Durbin-Watson Test statistic $d$.](image)

**Calculation of Overall Level Change ($LC_{Overall}$)**

Based on Huijema (2011), the overall level-change is a weighted average of the $j$ level-change coefficients, where the weights are the reciprocals of the error variance estimates for the individual level-change coefficients, calculated by:
\[
LC_{\text{Overall}} = \frac{\sum_j \frac{1}{\hat{\sigma}^2_{zj}} \beta_{zj}}{\sum_j \frac{1}{\hat{\sigma}^2_{zj}}}
\]  

(6)

where

\( j \) is the number of units (number of participants);

\( \beta_{zj} \) is the linear regression level-change coefficient estimated for the \( j \)th unit;

\( \hat{\sigma}_j \) is the linear regression estimated standard error for the \( j \)th level-change coefficient;

\( z_j \) is the value of \( z \) associated with the linear regression \( p \)-value for unit \( j \);

\( \hat{\sigma}^2_j \) is the linear regression variance estimate associated with \( j \)th LC coefficient;

\( \frac{1}{\hat{\sigma}^2_{zj}} \) is the reciprocal of \( \hat{\sigma}^2_j \), and is the weight used for the weighted average.

**LC_{\text{Overall}} Test Statistic**

The test statistic is based on the amount of evidence associated with each level-change, indicated by the \( z \)-score associated with the \( p \)-value for the level-change for each phase. In this study there is one phase change for each participant. The test statistic is the overall \( z \)-score calculated dividing the sum of the \( z \)-scores by the square root of the number of units.

\[
Z_{\text{Overall}} = \frac{\sum_j z_j}{\sqrt{j}}
\]  

(7)

where

\( j \) is the number of units (number of participants);

\( z_j \) is the value of \( z \) associated with the linear regression \( p \)-value for unit \( j \).
The $Z_{\text{Overall}}$-statistic is distributed approximately as a standard normal deviate and so will provide the associated $p$-value.

**Overall Standardized Effect Size**

An effect size is exactly equivalent to a “Z-score” of a standard Normal distribution. For example, an effect size of 0.8 means that the score of the average participant in the intervention phase is 0.8 standard deviations above the average participant in the baseline phase. The overall standardized effect size is the overall level change standardized by the within-unit variation. The within-unit variation is calculated by taking the square-root of the sum of the residual sum of squares divided by the pooled within-unit residual degrees of freedom.

$$
ES_{\text{Overall}} = \frac{LC_{\text{Overall}}}{\sqrt{\frac{\sum_j SS_{\text{Residual}_j}}{\sum_j (N_j - P)}}}
$$

(8)

where

- $SS_{\text{Residual}_j}$ is the sum of squares residual from fitting the two phase model to unit $j$;
- $\sum_j SS_{\text{Residual}_j}$ is the pooled within residual sum of squares;
- $N_j$ is the number of observations in the series associated with unit $j$;
- $P$ is the number of parameters in the intervention model applied to unit $j$;
- $(N_j - P)$ is the residual degrees of freedom for unit $j$;
- $\sum_j (N_j - P)$ is the pooled within unit residual degrees of freedom.
Results

Interrater Reliability

All videos were reviewed and rated by the researcher and a random selection of 20% of the videos, from each phase of each participant, were reviewed and rated by a Certified Flight Instructor Instrument Rating - 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. Using SPSS one-way random effect model, the single measure ICC was .924, with 95% CI = (.769, .977). These results show excellent reproducibility (Rosner, 2005) for this study.

Visual Analysis

For visual analysis, the data were plotted for each of the three participants. The three graphs were then positioned above one another, with a vertical dashed line indicating when the intervention, the change from digital to analog flight instrumentation, took place (see Figure 3.6). It can be seen from the dotted trend-lines in the baseline, Phase A, there is a “downward” trend towards zero errors and that the variability about the trend line reduces, as predicted, within the first five sessions, for all three participants. The intervention phase, Phase B, began at session 6 for Participant 1, when all three participants had achieved the predicted trend and reducing variability, within the first five sessions of the baseline phase. Participant 1, after a marked increase in the number of
PTS errors, again achieved a “downward” trend toward zero errors and an acceptable variability about the trend line by session 10, the first five sessions of the intervention phase. Therefore, Participant 2 began the intervention phase at session 11. Participant 2, after a marked increase in the number of PTS errors, achieved a “downward” trend toward zero errors and an acceptable variability about the trend line by session 15. So Participant 3 began the intervention phase at session 16, and also after a marked increase in the number of PTS errors, achieved a “downward” trend toward zero errors and an acceptable variability about the trend line by session 20. At this point, the study was concluded.

*Figure 3.6. Graphed data for all participants with trend lines leading to intervention.*
It can be seen that, for each participant, there was a marked increase in the number of PTS errors at the introduction of the intervention. The intervention was introduced at different times for each of the participants, and would therefore suggest that the degradation in flight performance, indicated by the marked increase in PTS errors, experienced by each participant is directly related to the change from digital flight instruments to analog flight instruments.

In Phase A, the means and standard deviations are: Participant 1, sessions 1–5, mean = 1.4 and $SD = 1.52$; Participant 2, sessions 1–10, mean = 7.10 and $SD = 2.06$; Participant 3, sessions 1–15, mean = 9.00 and $SD = 3.82$.

In Phase B, the performance of Participant 1 degraded from an error rate of close to zero PTS errors per session to 6 PTS errors per session for two flights before gradually improving back towards a zero error rate. Participant 2 experienced a higher performance degradation, moving from an error rate of close to 6 PTS errors per session to 19 PTS errors per session, before also gradually improving back towards baseline error rate. Participant 3 experienced performance degradation similar to Participant 2, moving from an error rate of close to 7 PTS errors per session to 19 PTS errors per session, before gradually improving back toward baseline error rate.

In Phase B, the means and standard deviations are: Participant 1, sessions 6–20, mean = 2.33 and $SD = 1.88$; Participant 2, sessions 11–20, mean = 9.80 and $SD = 6.03$; Participant 3, sessions 16–20, mean = 16.80 and $SD = 2.68$. 
**Statistical Analysis**

Data were fitted to Equation (1) and SPSS statistical software was used to regress Y on T, D, SC, and to obtain the D-W test statistic d for each of the three participants (see Table 3.1).

The D-W d, for 3 predictors and 20 observations has an upper and lower bound of \(d_U = 1.676\) and \(d_L = .998\). The fail to reject \(H_0\) areas (see Figure 3.5) for autocorrelation of data are:

\[
d_U = 1.676 < d < 2 + d_L = 2.998
\]

For all three participants \(1.676 < d < 2.998\) (see Table 3.1) and therefore the conclusion is to fail to reject \(H_0\). The data are confirmed as having no lag-1 autocorrelation, the errors are independent and the four-parameter model with independent errors, Equation (2), is the correct model to use.

The parameter estimates and associated inferential results from the separate regression analyses required for the level-change and effect size computation previously discussed are shown in Table 3.1.

The overall level-change statistic \(LC_{\text{Overall}}\) (see Equation 6) is 9.545, which is the weighted average of the \(j\) level-change coefficients, where the weights are the reciprocals of the error variance estimates for the individual level-change coefficients. The overall level-change test statistic \(Z_{\text{Overall}}\) (see Equation 7) is 5.716, and the associated one-tailed \(p\)-value is <.0001, so the overall level-change is highly significant. The overall standardized effect size, the difference between means in standard deviation units is \(ES_{\text{Overall}}\) (see Equation 8) is 4.935. This represents an extremely large effect size.
Table 3.1

Summary of Values From Regression Required for Computation of Weighted Overall Level Change

<table>
<thead>
<tr>
<th>Unit</th>
<th>$\beta_j$</th>
<th>$\delta_{zj}^2$</th>
<th>$t_j$</th>
<th>$p_i$</th>
<th>$z_j$</th>
<th>$\delta_{zj}^2$</th>
<th>$\frac{1}{\delta_{zj}^2}$</th>
<th>SS Residual</th>
<th>D-W d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.563</td>
<td>1.314</td>
<td>4.327</td>
<td>.001</td>
<td>3.10</td>
<td>1.727</td>
<td>0.579</td>
<td>20.576</td>
<td>1.915</td>
</tr>
<tr>
<td>2</td>
<td>13.297</td>
<td>1.645</td>
<td>8.083</td>
<td>&gt;.001</td>
<td>3.40</td>
<td>2.706</td>
<td>0.370</td>
<td>53.321</td>
<td>2.275</td>
</tr>
<tr>
<td>3</td>
<td>15.543</td>
<td>2.558</td>
<td>6.007</td>
<td>&gt;.001</td>
<td>3.40</td>
<td>6.543</td>
<td>0.153</td>
<td>116.911</td>
<td>2.131</td>
</tr>
</tbody>
</table>

$$\sum_1^j z_j = 9.90 \quad \sum_1^j \left(\frac{1}{\delta_{zj}^2}\right) = 1.102$$

Discussion

The data were both visually and statistically analyzed. Visual analysis is commonly used for analysis for large effect sizes in all fields using single case research designs, but statistical analysis is becoming more common when variability in the data and smaller effect sizes prevents researchers from making conclusive visual analysis.

Visual Analysis

Applied aviation research is a field in which single subject research designs have been very rarely used, and in the studies using SCDs, the criteria for visual analysis suggested by Kratochwill et al. (2010) had not yet been published. The six features suggested by the experts in other fields as standards for visual analysis of data will need to be considered as they pertain to the field of aviation.
“Level” is the mean score for the data within a phase and is a feature that 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. In this study, the “level” is the mean number of PTS errors committed each participant. At the introduction of the intervention to each participant there was a significant increase in PTS errors indicating degradation in flight performance by each participant.

“Trend” is the slope of the best-fitting straight line for the data within a phase. For fields of research in which the intervention is expected to have a distinct effect or even a reversal of the slope this would be a useful feature. However, in aviation the intervention may not produce any noticeable change in the slope and a reversal of the slope would be an exception. In this study, the slope represents the rate of change in PTS errors, achieved by the pilots flying more or less accurately. The intervention changes how the information pilots require to fly accurately is presented. Once the new method of presenting the information has been assimilated, the pilot’s ability to fly accurately and keep errors to a minimum would be expected to return towards the baseline rate. Trend would not be a suitable feature for visual analysis of this study and the usefulness of this feature in other studies would need to be assessed on a study by study basis.

“Variability” is the range or deviation of the data points about the trend line. This will depend on the participants and would be an important feature to analysis in all fields. As with trend, this feature would need to be assess on a study-by-study basis. For this study, it is a useful feature to use as an indication when the intervention phase should be
introduced. Each participant would be expected to “learn” how to assimilate the new information and, therefore, reduce the variability in their flight performance (PTS errors).

“Immediacy of the Effect” shows how quickly the intervention produces a measureable effect in the dependent variable, essential if the effect of the intervention by chance is to be discounted. This is a feature essential for all types of research, including applied aviation research. For this study, there was an immediate increase in the number of PTS errors for each participant when the intervention was introduced.

“Overlap” is the values of the data points in the intervention phase approaching the values of the data points in the baseline phase. This feature 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.

“Consistency of data patterns across similar phases” is self-explanatory and also an essential feature for fields of research if the effect of the intervention by chance is to be discounted.

Of the six visual features suggested for visual analysis, four were suitable for analysis of this study and possibly other applied aviation research studies: (1) level, (2) variability, (3) immediacy of the effect, and (4) consistency of data patterns across. Based on these four visual features, the visual analysis of the data showed that the intervention had a negative effect on pilot performance as measured by the dependent variable “Number of PTS Errors.” In other words, the three participants were unable to fly with the same accuracy after the intervention as they had flown before the intervention.
Statistical Analysis – Comparison of Findings

A comparison of the results of the original study (Whitehurst & Rantz, 2012) with the results from the Huitema method of analysis shows that both results were similar in that they were statistically significant. The effect size for the original study was 3.44 and the effect size for this study was 4.94. In other words, in the original study, the average participant in the intervention phase was 3.44 standard deviations above the average participant in the baseline phase in their number of PTS errors, and in the MBD study, the average participant in the intervention phase was 4.94 standard deviations above the average participant in the baseline phase in their number of PTS errors. Although both studies produced extremely large effect sizes, the MBD produced the larger effect size showing that it is more sensitive to the active design variable (independent variable).

Advantages of Single Case Designs in This Context

There are several advantages to using the MBD in place of the between-group design used in the original study. The main advantage is the reduction in the number of subjects required. For this design, there is a requirement for only three participants to meet evidence standards (Kratochwill et al., 2010). In both the original study and this study, one of the difficulties was finding volunteers that met the criteria for the study. The most important criterion was the requirement that a participant could not have flown any aircraft equipped with the analog flight instrumentation. However, once student pilots have received their private pilot certification, they are able to lease aircraft to build up their flight hours and experience. The newer digitally equipped aircraft are more
expensive to lease than the older analog equipped aircraft, so many student pilots will have flown analog equipped aircraft before reaching the stage of instrument training required as another criteria for the study. The pool of eligible student pilots is therefore severely limited, creating a problem when trying to recruit the number of participants to provide the statistical power required for a between-group study.

Other advantages that stem from the reduction in participants are reduced time spent recruiting; reduced time required to complete the study, and, very important in today’s economy, a greatly reduced cost associated with both the researcher’s time and the cost of the use of any equipment, such as simulators.

**Cost Analysis**

Both this study and the original study used a PCATD to simulate flight conditions, and although PCATD have been approved for use in flight training, they do not simulate the real airplane to the same degree as even an advanced aviation training device (AATD) like the Redbird FMX would. The Redbird FMX is a full motion AATD with wrap-around visuals and a fully enclosed cockpit. If funding had been available, the AATD would have provided a more realistic environment for the research study. Using the costs associated with the Redbird FMX, a simple cost calculation can be made to compare the cost of the original between-group study and the MBD study. If the participants had been available, the original study would have needed 42 participants to provide the accepted level of statistical power (0.8) to detect a large effect size. Each participant would have flown four flight profiles pre-intervention and four flight profiles
post-intervention, a total of 336 flight profiles. Each flight profile takes approximately 30 minutes, 20 minutes to fly and 10 minutes to set up and debrief, a total time of approximately 168 hours. The Redbird FMX simulator costs $75 per hour at the college from which the participants were recruited. The use of the AATD would have cost approximately $12,600.

The MBD study required only 3 participants, with each participant flying 20 flight profiles. At 30 minutes per flight profile, the study required a total time of \(3 \times 20 \times 0.5 = 30\) hours, at a cost of $75 per hour gives a cost for this study of $2,250. Not including any other costs, such as principal investigator (PI), co-PI and/or assistant’s time required for the simulator flights, or reviewing the videos and data analysis, by simply using a MBD the cost saving would have been $10,350 or 460% more than the cost of the MBD.

**Disadvantages of Multiple Baseline Single Case Designs in This Context**

The disadvantages of using MBDs in this context stem from the bias toward between-group designs and the belief that only group designs provide internal and external validity. Kazdin (1982) stated, “Internal validity refers to the extent to which an experiment rules out alternative explanations of the results” (p. 77) and “External validity addresses the broader question and refers to the extent to which the results of an experiment can be generalized or extended beyond the conditions of the experiment” (p. 81).

MBDs can rule out threats to internal validity and improve external validity by demonstrating the effect of the intervention occurs only when the intervention is applied,
at different time points across the different participants and phases, hence the requirement for a minimum of six phases, three interventions at three different time points, for a MBD recommended by Kratochwill et al. (2010).

Conclusion

The findings from this study suggest that the MBD may be a very appropriate and efficient design for replicating this study and possibly for other studies within applied aviation research. The Huitema method for statistical analysis produced similar results to the ANCOVA used in the original study, which would suggest it is a suitable method for statistical analysis of MBDs. The intervention effect size in this study was 144% larger than the effect size produced by the original between-group study, clearly showing that the MBD is more sensitive to the active design variable.

The extremely large effect size also ensured it could be observed easily using visual analysis, an analysis that, if the effect size is large enough, can be very quick and simple to use. However, not all of the six visual features recommended by Kratochwill et al. (2010) for visual analysis were suitable for this study and possibly may not be suitable for other applied aviation studies. The two visual features not suitable for this study were (a) “trend”—this was expected to be similar pre- and post-intervention; and (b) “overlap”—the expected trend post-intervention would suggest possible overlap. These two features, although not suitable for this study, may well be suitable for other applied aviation studies and should not be disregarded unless prior knowledge suggests otherwise.
Finally, it took several years and several aircraft accidents for the FAA to initiate the FITS training program for transitioning from analog to TAA. It is hoped that the results of this study and the original study will help to speed up the process of establishing a training program for the transition from any digitally instrumented aircraft, including TAA, to analog instrumented aircraft.

References


CHAPTER IV
THE COST OF INCREASED VALIDITY: COMBINING A MULTIPLE BASELINE DESIGN WITH AN ABAB DESIGN∗

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 (2010) describes 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 $\alpha$-value) (Ferrin, Bishop, Tansey, Frain, Swett, & Lane, 2007).

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

∗This paper is under review by the Global Science and Technology Forum’s Journal of Aviation Technology.
overwhelming cost and availability of flight simulators or aircraft. One solution to the problem of sample size proposed 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. These 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.

**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 most threats to internal validity in SCD research. Acceptable evidence standards for showing intervention effects suggested by Kratochwill et al. (2010) 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). Figure 4.1 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.

**Horizontal Design**

![Horizontal Design Image]

**Horizontal/Vertical Design**

![Horizontal/Vertical Design Image]

*Figure 4.1. Visual representation of designs.*
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 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. (2010): “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.
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 (Hayes, 1981). The most common form of the design that meets the replication criteria advanced by Horner et al. (2005) 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. Internal validity was critical, whereas the external validity was not a primary concern. A review of the literature shows that in most fields using single case research, this is still the case. However, with other fields 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 the within-series comparisons. The combined design has the advantage of the MBD’s between-series comparisons together with the systematic replication suggested by Hayes (1981) to increase the internal validity of the study. Applied aviation research is a field in which generalization is a necessity as well as strong internal validity, so comparing the
combined design to the MBD in terms of internal and external validity as well as results and cost is very important.

**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 (Whitehurst & Rantz, 2012). The purpose is to compare the combined design to the MBD, used in a previous study (Whitehurst, under review), 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.

**Research Methodology**

**Research Design**

In a previous study (Whitehurst, under review), a 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 Figure 4.2). The intervention data are flight performances of the participants flying the PCATD configured to emulate the Cessna 182 Skylane RG, an AFI aircraft (see Figure 4.3). 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 for 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 were 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. The reduction in the variability of the data would be a result of the participant becoming more familiar with the equipment and environment.

Figure 4.2. Digital flight instrumentation. Figure 4.3. Analog flight instrumentation.

There is no specified limit to the variability of the data for it to be considered “stable.” Kratochwill et al. (2010) 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 (Whitehurst & Rantz, 2012) 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 complete 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.

Participants

The participants in this study were students enrolled in the flight program of a university offering a four-year degree in aviation flight science. To be eligible to
participate, the volunteers needed to meet the following criteria: (a) to have only flown aircraft with digital flight instrumentation, and (b) to be enrolled in and to have completed a minimum of 65 hours (80%) of the flight program’s flight instrument certification course.

The 65 hours criterion was selected after consultation with the flight program’s certified flight instructors (instrument) to ensure the participants would have the necessary level of flight skills needed to fly the instrument flight patterns for the study and to have sufficient time (8 weeks) to complete the study before completing the flight instrument certification course. The latter requirement was due to the possibility of participants flying aircraft with analog flight instrumentation in the next course of their flight training, thus negating criterion a).

Although four participants began the study, due to personal reasons, one participant withdrew during the first intervention phase. The minimum of six phases, required to meet evidence standards suggested by Kratochwill et al. (2010), was met by the three participants who completed the study. Only data from those participants who completed all four phases of the study were used in the data analysis.

**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.
For each A phase of the study, participants flew the PCATD emulating the DFI aircraft, and for each B phase of the study, participants flew the PCATD emulating the AFI aircraft. Each flight was visually recorded for later analysis of the participant’s flight performance, using a Contour Nflightcam video camera 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.

**Dependent Variable**

During the flight instrument certification course, flight students are trained to fly accurate headings, altitudes, and speeds without using visual references. The accuracy of a pilot’s instrument flying is checked during the certification flight test based on the FAA PTS. The limits specified in the FAA’s PTS were used to measure the accuracy of the participant’s instrument flying, which for this study was called flight performance. The dependent variable consisted of the total number of times the participants flew outside the limits specified in the FAA’s PTS for instrument flight check rides. The limits 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. Exceeding any of the three limits was recorded as one error and the total number of errors was recorded for each participant for each session.

**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 intrarater reliability checks, if required.

**Flight Patterns**

Flight patterns for four airports familiar to the participants were used in the study on a random basis to prevent the possibility of “testing,” an internal validity problem, and the familiar airports prevented any stress a new flight environment may have induced. The headings, altitudes, and speeds the participants were required to fly for each leg of the flight pattern were briefed before each session. From an adjacent room, simulating air
traffic control (ATC) giving instructions and radar vectors, the researcher repeated the headings, altitudes, and speeds and other typical ATC instructions using a commercially available intercom system and prepared script for each flight pattern. The flight patterns all included a take-off from the airport’s main runway, a climb to a specified altitude (3,000 or 3,500 feet dependent on the airport’s approach requirement), a 500 feet descent while turning 90º, a 90º level turn, and a heading for localizer interception. At this point the participant was cleared to fly the headings, altitudes, and speeds required to intercept and maintain the localizer and glideslope for an ILS approach to a visual landing.

**Data Collection**

Each PCATD session involved one participant flying one flight pattern and, based on the participant’s schedules and personal commitments, two or three sessions were flown each week. The order the participants flew was randomly assigned at the beginning and then maintained throughout the study. Data were collected during instrument flight conditions, which began on cloud penetration at 300 feet above ground level on climb out and ceased at the decision height of 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.

**Interrater Reliability**

Interrater reliability was determined from the intraclass correlation coefficient (ICC), an index of the reliability of the ratings. Wuensch (2010) suggested using the ICC
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. In this study, an interpretation of the ICC values suggested by Rosner (2005) was used as follows: values less than 0.40 indicates poor reproducibility; values of 0.40 to 0.75 indicate fair to good reproducibility, and an values of greater than 0.75 show excellent reproducibility.

**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.

**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 (Zhan & Ottenbacher, 2001). 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 Figure 4.4). On the basis of these graphs, a judgment is reached about the reliability or consistency of intervention effects (Long & Hollin, 1995).
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 (Fisher, Kelley, & Lomas, 2003; Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; Morgan & Morgan, 2009; Parsonson & Baer, 1978) (see Figure 4.5). 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-

Figure 4.4. MBD with overall mean and trend lines.
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.

Figure 4.5. Visual analysis features.
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 (Furlong & Wampold, 1981), i.e., 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 (Kratochwill et al., 2010).

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. The four features were “Level,” “Trend,” “Variability,” and “Immediacy of the Effect.” 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” and “Consistency of data patterns across similar phases” are essential 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, which shows a change in the “level” and is observable outside the “variability” of the data.

**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,” (Kazdin, 1982, p. 241). Morley and Adams (1991) 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 (2003) also suggested the use of hierarchical linear models (HLM) for single case data. In this study, HLM was used 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 (2003) 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 2 baseline and 2 intervention phases, whereas the MBD overall phase effect was calculated using only 1 baseline and 1 intervention phase. The regression equations for the unconditional model, or the model with no treatment indicator, for both designs are:

For level 1

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

(1)

For level 2

\[ \beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}) \]  

(2)

Where

- \( Y_{ij} \) is the response score of participant \( j \) (\( j = 1, 2, 3 \) for both designs) for occasion \( i \) (\( i = 1 \ldots 20 \) for the 2 phases of the MBD and \( i = 1 \ldots 50 \) for the 4 phases of the combined design);
- \( \beta_{0j} \) is the mean response for participant \( j \);
- \( \gamma_{00} \) is the mean across participants;
- \( u_{0j} \) is the random error associated with participant means, \( \text{var} (u_{0j}) = \tau_{00} \);
- \( e_{ij} \) is the random error associated with occasion \( i \) for participant \( j \), \( \text{var} (e_{ij}) = \sigma^2 \).
The regression equations for the conditional model, or the model with the treatment indicator, are:

Level 1

\[ Y_{ij} = \beta_{0j} + \beta_{1j}(\text{phase})_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2) \]  \hspace{1cm} (3)

Level 2

\[ \beta_{0j} = \gamma_{00} + u_{0j} \quad \text{and} \quad u_{0j} \sim N(0, \tau_{00}^2) \]  \hspace{1cm} (4)

\[ \beta_{1j} = \gamma_{10} \quad u_{0j} \sim N(0, \tau_{00}^2) \]  \hspace{1cm} (5)

where

- \( Y_{ij} \) is the response score of participant \( j \) (\( j = 1, 2, 3 \) for both designs) for occasion \( i \) (\( i = 1 \ldots 20 \) for the 2 phases of the MBD and \( i = 1 \ldots 50 \) for the 4 phases of the combined design);
- \((\text{phase})_{ij}\) is an indicator that equals 1 if occasion \( i \) for participant \( j \) is part of the intervention phase, 0 otherwise;
- \( \beta_{0j} \) is the mean response for participant \( j \) in the baseline phase;
- \( \beta_{1j} \) is the magnitude of the effect of the intervention on participant \( j \);
- \( \gamma_{00} \) is the mean baseline level;
- \( \gamma_{10} \) is the mean intervention effect;
- \( u_{0j} \) is the random error associated with participant means, \( \text{var} (u_{0j}) = \tau_{00} \);
- \( r_{ij} \) is the random error associated with occasion \( i \) for participant \( j \) controlling for (phase) and is a conditional or residual variance, \( \text{var} (r_{ij}) = \sigma^2 \).
In the conditional model, the parameters of interest are the fixed effects $\gamma_{00}$ and $\gamma_{10}$ and the variance parameters $\sigma^2$ and $\tau_{00}$. 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 ($\gamma_{10}$) by the square root of the residual between-person variance ($\tau_{00}$) (Van den Noortgate & Onghena, 2003).

**Results**

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

**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, were 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 is .948, 95% CI = (.894, .975); this shows excellent reproducibility (Rosner, 2005).
Visual Analysis

For visual analysis, the data were plotted for each of the three participants. Figure 4.6 shows the graphed data for the three participants for the combined design. It can be seen from the dotted trend-lines in the first Phase A that “stability,” a “downward” trend, and variability about the trend line, within the plus or minus 2 PTS errors, were achieved for all three participants within the first five sessions. Participant 1 began the B phase at session 6.

![Graphed data for all participants with trend lines leading to intervention.](image)
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 sessions 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, who would appear to have more difficulty assimilating the new information. 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, inferring a causal relationship. The means and standard deviations (SD) of the number of PTS errors for the combined design are presented in Table 4.1. For comparison, means and SDs from the MBD are given in Table 4.2.

Table 4.1

Visual Analysis Results for the Combined Design

<table>
<thead>
<tr>
<th>Phase</th>
<th>Participant</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>1.40</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.30</td>
<td>2.06</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.87</td>
<td>2.23</td>
</tr>
<tr>
<td>B1</td>
<td>1</td>
<td>2.27</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.07</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>12.07</td>
<td>4.23</td>
</tr>
<tr>
<td>A2</td>
<td>1</td>
<td>0.27</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.80</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.00</td>
<td>3.82</td>
</tr>
<tr>
<td>B2</td>
<td>1</td>
<td>0.80</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.90</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11.20</td>
<td>3.83</td>
</tr>
</tbody>
</table>
Table 4.2

Visual Analysis Results for the MBD

<table>
<thead>
<tr>
<th>Phase</th>
<th>Participant</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1.40</td>
<td>1.52</td>
</tr>
<tr>
<td>A1</td>
<td>1</td>
<td>7.30</td>
<td>2.06</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.87</td>
<td>2.23</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.27</td>
<td>1.94</td>
</tr>
<tr>
<td>B1</td>
<td>1</td>
<td>7.07</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12.07</td>
<td>4.23</td>
</tr>
</tbody>
</table>

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 3 and Table 4, respectively. For the combined design, the estimated overall baseline mean ($\gamma_{00}$) is 3.35 and the estimated phase effect ($\gamma_{10}$) 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. In other words, the average participant in the intervention phase was 0.67 standard deviations above the average participant in the baseline phase in their number of PTS errors, a large effect size. For the MBD, the estimated overall intercept ($\gamma_{00}$) is 5.43 and the estimated phase-indicator ($\gamma_{10}$) 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-person variance (3.69) and is 0.95. In other words, the average participant in the intervention phase was 0.95 standard deviations above the
average participant in the baseline phase in their number of PTS errors, 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 decreased by 25%, whereas the square root of the residual between-person variance increased by 4%. The reduction in effect size is mainly due to the decrease in the between-phase effect size, confirming that the “learning” the participants were expected to make, in flying the PCATD and assimilating the new form of information, did occur.

Table 4.3

*Fixed and Random Effects for Combined Design*

| Effects      | Estimate | Standard Error | df | t Value | $\chi^2$ | Pr > |t| | Var. Comp. |
|--------------|----------|----------------|----|---------|---------|-------|---|-----------|
| Fixed        |          |                |    |         |          |       |   |           |
| Intercept ($\gamma_{00}$) | 3.351765 | 2.240330 | 2  | 1.496   | 0.273   | <0.001 |   |           |
| Condition ($\gamma_{10}$) | 2.589803 | 0.637393 | 146 | 4.063   |         |       |   |           |
| Random       |          |                |    |         |          |       |   |           |
| Level ($u_{0j}$) | 1        | 3.80254 | 2  | 99.439  | <0.001  | 14.459 ($\tau_{00}$) |   |           |
| Level ($r_{ij}$) | 2        | 3.85188 |     |         |         | 14.837 ($\sigma^2$) |   |           |
Table 4.4

*Fixed and Random Effects for MBD*

<table>
<thead>
<tr>
<th>Effects</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>df</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Var. Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>5.431867</td>
<td>3.195408</td>
<td>2</td>
<td>1.700</td>
<td>0.231</td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>3.502933</td>
<td>1.040230</td>
<td>56</td>
<td>3.367</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>5.39892</td>
<td>5.68614</td>
<td>2</td>
<td>87.622</td>
<td>&lt;0.001</td>
<td>29.148 (τ₀₀)</td>
</tr>
<tr>
<td>Level 2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.588 (σ²)</td>
</tr>
</tbody>
</table>

**Discussion**

Hayes (1981), Horner et al. (2005), and Kratochwill et al. (2010) 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.
The designs were compared with respect to the internal and external validity and the results of the analyses from the two designs. The costs for the combined design and the MBD were also compared to determine what increase in cost is associated with increasing replications.

**Internal Validity**

Kratochwill et al. (2010) 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 were not 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. (2010).

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, 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.

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 (ABAB) design replicates the intervention effect across the participants at the second B phase. If we look at the ABAB phases as two separate AB phases as similar to the concept of the between-group randomized block design and treat each AB as a block with homogeneous participants, then variability within each block will be less than the variability of the combined ABAB phases and therefore the estimate of the treatment effect within a block is more efficient than estimates across the combined phases. To determine the correlation between the replicated intervention effects, an ICC was calculated. Using SPSS’s one-way random effects model, the single measure ICC is .573 and the 95% CI is (−1.170, −.140). These results show fair to good reproducibility (Rosner, 2005). This suggests that the combined design has an advantage over the MBD with respect to external validity, but by what amount is subjective.

**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, and “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 Figures 4.4 and 4.7).

Figure 4.7. Graphed data for combined design with overall mean and trend lines.

*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 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.

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, the costs associated with the Redbird FMX were used since this is an appropriate AATD to use in this study. A basic cost calculation can be made to compare the cost of the MBD and combined design. The cost calculation is kept simple by basing it on the cost of the 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 3 participants. For the MBD, each participant flew 20 flight profiles. At 30 minutes per flight profile the study required a total time of $3 \times 20 \times 0.5 = 30$ 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 $3 \times 50 \times 0.5 = 75$ hours; at a cost of $75$ per hour for the flight simulator, this cost would be $5,625$, a 250% increase in cost.

**Conclusion**

The MBD has good internal validity due to the replication of the intervention effect across subjects at staggered times; however, the internal validity of the combined design is further improved by the replication of the AB phases. Although the MBD does
have external validity because of this design has between-person as well as within-person comparisons, the combined design has improved external validity compared 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.

References


CHAPTER V

CONCLUSION

The purpose of this final chapter is (a) to summarize the results of each study presented in Chapters II, III, and IV; and (b) to interpret and integrate the results of each study to examine the overall findings.

Review and Summary of Chapters II, III, and IV

Summary of Chapter II

The first study described in Chapter II was a concept paper with two research objectives: (1) to determine if the sample size used in experimental and quasi-experimental studies, published in three prominent U.S. aviation journals, provided the minimum accepted statistical power (.8) to detect meaningful effects; and (2) to provide an alternate methodology to aviation researchers, which requires only small sample sizes. The three journals, *International Journal of Applied Aviation Studies*, *Journal of Aviation/Aerospace Education and Research*, and *Collegiate Aviation Review*, published a total of 470 articles during the established review period. The articles were classified as non-research (151), survey research (255), quasi-experimental (34), and experimental (30). The experimental and quasi-experimental designs were further categorized into single intervention (35) and multiple intervention (29), as the MBD is applicable only for single intervention experiments. A post hoc calculation of the statistical power for each of
the 35 single intervention experimental studies was completed. The calculations were based on a medium effect size (.25), which is generally used in aviation research when no prior research provides a more specific estimate, and \( \alpha = .05 \). Two studies did not include sufficient data for power to be calculated, 7 had power of .8 or greater, the accepted minimum, and 26 had less than the accepted minimum power.

These results showed that information on statistical power and effect sizes was not usually presented, and post hoc calculations of statistical power showed, in most cases, the sample sizes were not large enough to adequately power a study with a medium effect size. Information on statistical power and effect sizes is presently not a requirement of aviation journal editors; therefore, the lack of information may be a function of this deficiency. However, as Ison (2011) stated, “These facts call into question the sample size strategies used in these studies. Further, the validity of the conclusions made upon statistical analyses could therefore be debatable” (p. 82). The use of an experimental design requiring small sample sizes provides a potential solution to this problem. Single case research is a methodology in which sample sizes can be as small as \( N = 1 \).

SCDs, and specifically the MBD, one type of SCD, were proposed as a suitable design for aviation research studies where only small sample sizes were available. The MBD design was described as an alternative design in one of the between-group single intervention studies.
Summary of Chapter III

Paper 1 provided theoretical procedures for replacing between-group designs. In Paper 2, a study by Whitehurst and Rantz (2012) that investigated flight performance of pilots trained on digital flight instrumentation when they transitioned to analog flight instrumentation for the first time was replicated. The original study was a between-groups design. The replicated study was a MBD. Data from the MBD were analyzed both visually and statistically. The effect size of the intervention was sufficiently large that visual analysis of the data clearly suggested an intervention effect. The analysis showed that each of the three participants had a large degradation in their flight performance (dependent variable) at the introduction of the change in flight instrumentation (independent variable), suggesting a causal relationship between the researcher-manipulated independent variable and the change in the dependent variable. The statistical analysis showed an overall level change, a weighted average of the mean difference between the levels, of 9.55 errors, which was significant, \( p < .001 \). The overall standardized effect size was 4.94 standard deviations, an extremely large standardized effect by any standards. The conclusion from these results is that pilots transitioning from digital to analog flight instrumentation for the first time will suffer a significant degradation in their flight performance, potentially a serious flight safety issue for the aviation industry.

Both the statistical analysis from the MBD and from the ANCOVA, used in the original study, suggested a statistically significant intervention effect. Effect sizes from both designs were standardized using the pooled within unit variation. Standardization of
effect sizes allows comparison across studies. The standardized effect size in the MBD study was 199% larger than the standardized effect size produced by the original between-group study, clearly showing that the MBD is more sensitive to the active design variable. A cost analysis of the two designs showed a significant savings in time and resources for the MBD compared to the between-group design, providing support for the MBD as an efficient alternative to the between-group design.

**Summary of Chapter IV**

The third study described in Chapter IV compared two types of SCDs, the MBD and a combined design using the context of the flight performance study by Whitehurst and Rantz (2012). The combined design, a combination of the MBD and the ABAB design, was compared to the MBD with respect to internal validity, external validity, analytical results, and cost. The results suggested there is an improvement in both internal and external validity with the combined design compared to the MBD as a result of the increased number of systematic *replications of the effect* within the course of the experiment (e.g., Hersen & Barlow, 1976; Horner et al., 2005; Kazdin, 1982; Kratochwill, 1978; Kratochwill & Levin, 1992). However, the literature does not specify the number of systematic *replications of the effect* the design needs to be increased by, or by what amount validity will be increased due to the increase in replications. The combined design doubles the number of systematic *replications of the effect* within the course of this experiment, so subjectively an improvement in validity would be expected, but by what amount is not quantifiable from the researcher’s literature review.
The intervention effect was sufficiently large that visual analysis of the data clearly identified the intervention effect regardless of the design. The analysis showed that each of the three participants had a smaller but noticeable degradation in their flight performance at the introduction of the second intervention phase of the combined design, showing that there was a very strong correlation of the intervention effect with the time of introduction of the intervention. The data from the MBD were re-analyzed using the statistical technique (HLM) required for the analysis of the combined design data. The results of the HLM analyses for both designs revealed a significant phase effect. The overall standardized effect size for the combined design was 0.67 standard deviations, a reduction from the 0.95 standard deviations produced by the MBD. These effect sizes are very large compared to effect sizes normally seen in the social sciences. However, in the field of aviation, effect sizes are very rarely calculated or included in articles published in aviation journals, so these effect sizes may not be large for aviation experiments.

The effect size is calculated by dividing the overall between-phase effect by the square root of the residual between-person variance. The overall between-phase effect reduced, while the square root residual between-person variance increased slightly. The reduction in effect size was mainly due to the 25% decrease in the between-phase effect. This decrease in between-phase effect size is an important consideration when selecting the most suitable design for the experiment. In any study involving pilots and flight performance, an important consideration is “learning.” A pilot’s flight performance is a “skill” that is based on knowledge and practice, and therefore the increased exposure, an unavoidable consequence of the increased number of sessions required by the combined
design, to a new environment or intervention will provide the practice necessary for “learning” to occur. This study’s results would suggest that the “learning” the participants were expected to make in flying the PCATD for the first time and assimilating the new form of flight information did occur. Another important challenge of the combined design compared to the MBD is the considerable increase in time required for the additional phases, and a simplified cost analysis showed a considerable increase in cost (250%) of the combined design compared to the MBD.

The results of the combined design were similar to the MBD results, in that both showed a statistically significant intervention effect, albeit with the effect size of the combined design being reduced from .95 to .67, which is still a large effect size. Overall, this study demonstrated that there was subjective, but not quantifiable, evidence that the combined design improved internal validity and external validity, but with an increase in costs. The increase in costs associated with using the combined design instead of the MBD would not appear to be justified in the current economic climate.

**Significance of Findings**

The findings of Papers 1 and 2 show that the MBD is a single case research design that can, both theoretically and practically, replicate a between-group experimental study using a smaller sample size, and thus be more efficient in terms of time and associated costs. The findings of Paper 3 show that the combined design is also a design that is capable of replicating a between-group experimental design and may
provide improved internal and external validity, with similar results, albeit at an increase in time and cost.

The overall significance of these findings is in the area of aviation flight safety. The results of Paper 2 and 3 are extremely important for the aviation industry as they reiterate a serious problem first highlighted by Whitehurst and Rantz (2012)—degradation in flight performance of pilots transitioning from digital to analog flight instrumentation for the first time. The transition of pilots from a traditional analog cockpit to a modern-glass cockpit has been a training challenge for the last two decades (Dahlstrom, Dekker, & Nahlinder, 2006), and many studies have been conducted on how this transition training should be carried out (Casner, 2003a, 2003b; Fanjoy & Young, 2003; Rignér & Dekker, 1999). However, the problem of transitioning from digital to analog has become an issue only over the last decade, as digitally equipped training aircraft are being used by an increasing number of flight schools, and little research has been conducted in this area of flight safety. With the number of pilots trained, being trained, or who will be trained on only digital flight instrumentation increasing daily, the potential for an accident/incident is also increasing rapidly and should be a serious concern for the aviation industry.

**Recommendations**

The findings of this research call for the following suggestions for consideration by the evaluation, measurement, and research community; the aviation research community; and the research community in general.
No matter what their field of study, researchers need to expand their knowledge of available research designs. The literature review for these papers showed that single case research is not a methodology commonly used in research outside of the fields of psychology and special education. This may be indicative of researchers’ lack of knowledge, or forgotten knowledge of this particular methodology. However, in today’s economic climate, where obtaining funding for research is becoming increasingly difficult, selecting the most appropriate research design is crucial, and knowledge of designs requiring only small sample sizes, such as SCDs, is therefore essential if cost is a major consideration. For the aviation industry, the MBD would seem ideal for any study involving small sample sizes, as it provides both internal and external validity without the potential problem of “learning” encountered with the combined design.

For the fields of study that are fairly young in terms of research history, in particular aviation, procedures and best practices used in the more established fields of research, such as psychology and education, need to be studied and possibly incorporated into their research. This also applies to editors of most journals in these research fields. There need to be firmer guidelines, for both researchers submitting papers and the reviewers, to ensure that there are sufficient data, such as sample size calculations and effect size, for the reader to validate the study’s conclusions. When researchers use sample sizes that do not provide the accepted level of statistical power (0.8), then evidence-based reasoning for the sample size and/or effect size used should be provided.
Future Research

In the review of the three prominent U.S. aviation journals in the first study of this dissertation, the most common methodology used was between-group NHST, and many of these studies had low statistical power based on the sample sizes used. It would be of interest to research the methodologies used in international aviation journals to determine if SCDs are used in any aviation research, or if between-group NHST is the common methodology used throughout the aviation industry. It would also be of interest to determine if the internationally published articles showed that low power studies are endemic in aviation research.

One concern for researchers using SCDs is the question of external validity when using such small sample sizes. In the fields where SCDs are most commonly used, psychology and special education, generalizability does not present as big an issue, since the researchers are generally interested in the treatment effect on individual patients or students, rather than the population as a whole. In the third paper of this dissertation, a combined design was used to increase the systematic phase replication (Hayes, 1981; Horner et al., 2005) to improve internal and external validity. Research into quantifying the validity improvement would be of interest, as justifying the increase in costs associated with the improved validity is subjective rather than objective.

In recent years, as SCDs are beginning to be used in other fields of research, other methods of improving generalizability are being investigated. One such method is randomization tests and nonparametric statistical analysis of the data. Bulté and Onghena (2008) stated:
Although randomization tests do not depend on parametric assumptions or on the assumption of random sampling, they do have one requirement: The experimenter has to designate certain times at which the treatment has to be administered and then randomly assign each time to a treatment. (p. 472)

As this method does not require the increase in systematic replication of phases required by the combined design and has been developed for use with the MBD, it may be a more cost-effective way of increasing external validity. It would be of interest to repeat the experiment in the second paper using randomization tests and the nonparametric data analysis and compare the results with the MBD study.

References


Appendix A

PTS Observation Form
Flight by Reference to Instruments

Maintains altitude within ±100 feet during level flight
Check for each deviation

Total deviations ______

Maintains headings within ±10°
Check for each deviation

Total deviations ______

Maintains airspeed within ±10 knots
Check for each deviation

Total deviations ______

Maintains bank angles within ±5° during turns
Check for each deviation

Total deviations ______
Instrument Approach Procedures: Precision Approach

Prior to beginning the final approach segment, maintains the desired altitude ±100 feet, the desired airspeed within ±10 knots, the desired heading within ±10°.

Check for each deviation

Total deviations ______

Maintains a stabilized final approach, from the Final Approach Fix to DA/DH allowing no more than ¾-scale deflection of either the glideslope or localizer indications

Check for each deviation

Total deviations ______

Maintains the desired airspeed within ±10 knots.

Check for each deviation

Total deviations ______

Observer/Inter-rater: _______________________

Additional Comments: ____________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________
Appendix B

Pre-Flight Instructional Script
Pre-flight Instructional Script

We will be conducting one instrument flights per session. One session should last about 30 minutes. The flight today will conclude with an instrument landing system approach to a full stop landing. You will be given assigned headings and altitudes to maintain until you are cleared for the instrument approach, the engine power settings and airspeeds required for the flight are displayed on the console in front of you. As you can see, we have the instrument approach plate for the ILS runway [runway number] at [name of airport]. Please take a moment to familiarize yourself with the ILS approach plate. Here is a copy of the latest Automatic Terminal Information Service (ATIS) information. Please be certain you understand how the PCATD works and that you are comfortable at the PCATD station. So as not to interfere in your flight, I will be leaving the room while you are conducting your flight and will not be able to help you in any way. I will be observing and recording your flight using the video camera, computer monitor, and flight simulation software to permit me analysis the flight at a later time. I will play the role of Air Traffic Control and provide you with appropriate vectors and altitudes. You will need to talk with [name of airport] Tower and [name of airport Approach Control]. We are starting the flight at the hold line of runway [runway number] at [airport]. The before starting engine, engine start, before taxi, and taxiing checklists have been completed.
Do you have any questions before we begin?

If for any reason you feel you need to discontinue the flight, just tell me that by saying it out loud and I will terminate the flight immediately.

Are you ready?

Please wait for my call to announce the beginning of the flight.
Appendix C

PCATD Engine Power and Throttle Settings for Airspeed
# TAKE OFF

<table>
<thead>
<tr>
<th>POWER</th>
<th>Prop (Blue)</th>
<th>2400 RPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Throttle (Black)</td>
<td>26-27 IN</td>
</tr>
</tbody>
</table>

**SPEEDS**

<table>
<thead>
<tr>
<th></th>
<th>Rotate</th>
<th>75 kt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Climb</td>
<td>85-90 kt</td>
</tr>
</tbody>
</table>

# CRUISE

<table>
<thead>
<tr>
<th>POWER</th>
<th>Prop (Blue)</th>
<th>2100 RPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Throttle (Black)</td>
<td>19-20 IN</td>
</tr>
</tbody>
</table>

**SPEED**

<table>
<thead>
<tr>
<th></th>
<th>110 kt</th>
</tr>
</thead>
</table>

# APPROACH

*(Cleared for ILS)*

<table>
<thead>
<tr>
<th>FLAPS 10°</th>
<th>POWER</th>
<th>Prop (Blue)</th>
<th>2000 RPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Throttle (Black)</td>
<td>15-17 IN</td>
<td></td>
</tr>
</tbody>
</table>

**SPEED**

<table>
<thead>
<tr>
<th></th>
<th>95 kt</th>
</tr>
</thead>
</table>

*(Intercept Glideslope)*

<table>
<thead>
<tr>
<th>FLAPS 20°</th>
<th>POWER</th>
<th>Prop (Blue)</th>
<th>1600 RPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Throttle (Black)</td>
<td>12-14 IN</td>
<td></td>
</tr>
</tbody>
</table>

**SPEED**

<table>
<thead>
<tr>
<th></th>
<th>85 kt</th>
</tr>
</thead>
</table>
Appendix D

Flight Pattern Narration
Flight Pattern Narration

Flight Pattern 1 KBTL

(EXPERIMENTER): *Session start, please begin. Contact tower when ready for takeoff.*

(PARTICIPANT): *Battle Creek Tower Western 45 ready for departure runway 23.*

(EXPERIMENTER): *Western 45 you are cleared for departure. Fly runway heading climb and maintain 3,000’.*

(PARTICIPANT): *Fly runway heading climb and maintain 3,000’ Western 45*

**After reaching 1000’ AGL**

(EXPERIMENTER): *Western 45 contact Kalamazoo Approach on 119.2.*

(PARTICIPANT): *Contacting Kalamazoo Approach on 119.2 Western 45.*

(PARTICIPANT): *Kalamazoo Approach Western 45 is with you heading 230 climbing to 2,500’.*

(EXPERIMENTER): *Western 45 roger.*

(EXPERIMENTER): *Western 45 turn left heading of 120.*

(PARTICIPANT): *Turning left to a heading of 120 Western 45.*

(EXPERIMENTER): *Western 45 turn left heading of 050 descend and maintain 2,500’.*

(PARTICIPANT): *Turning left to a heading of 050 descending to 2,500’ Western 45.*

(EXPERIMENTER): *Western 45 turn left to a heading of 320.*

(PARTICIPANT): *Turning left to a heading of 320 Western 45.*

(EXPERIMENTER): *Western 45 turn left to a heading of 270 cleared for the ILS 23 contact Battle Creek Tower 118.1.*

(PARTICIPANT): *Contacting Battle Creek Tower on 118.1 Western 45.*
(PARTICIPANT): Battle Creek Tower this is Western 45 on the ILS 23.

(EXPERIMENTER): Western 45 you are cleared to land runway 23.

(PARTICIPANT): Cleared to land runway 23 Western 45.

When aircraft stopped on the runway after landing

(EXPERIMENTER): This session is over. Please relax and I will join you in a few minutes.
Flight Pattern Narration

Flight Pattern 2 KAZO

(EXPERIMENTER): Session start, please begin. Contact tower when ready for takeoff.

(PARTICIPANT): Kalamazoo Tower Western 45 ready for departure runway 35.

(EXPERIMENTER): Western 45 you are cleared for departure. Fly runway heading climb and maintain 3,500’.

(PARTICIPANT): Fly runway heading climb and maintain 3,500’ Western 45

After reaching 1000’ AGL

(EXPERIMENTER): Western 45 contact Kalamazoo Approach on 121.2.

(PARTICIPANT): Contacting Kalamazoo Approach on 121.2 Western 45.

(PARTICIPANT): Kalamazoo Approach Western 45 is with you heading 350 climbing to 3,500’.

(EXPERIMENTER): Western 45 roger.

(EXPERIMENTER): Western 45 turn left heading of 260.

(PARTICIPANT): Turning left to a heading of 260 Western 45.

(EXPERIMENTER): Western 45 turn left heading of 170 and descend to 3,000’.

(PARTICIPANT): Turning left to a heading of 170 and descending to 3,000’ Western 45.

(EXPERIMENTER): Western 45 turn left to a heading of 080.

(PARTICIPANT): Turning left to a heading of 080 Western 45.

(EXPERIMENTER): Western 45 turn left to a heading of 030 cleared for the ILS 35 contact Kalamazoo Tower 118.3.

(PARTICIPANT): Contacting Kalamazoo Tower on 118.3 Western 45.
(PARTICIPANT): *Kalamazoo Tower this is Western 45 on the ILS 35.*

(OBSERVER: *Western 45 you are cleared to land runway 35.*

(PARTICIPANT): *Cleared to land runway 35 Western 45.*

**When aircraft stopped on the runway after landing**

(EXPERIMENTER): *This session is over. Please relax and I will join you in a few minutes.*
Flight Pattern Narration

Flight Pattern 3 KLAN

(EXPERIMENTER): Session start, please begin. Contact tower when ready for takeoff.

(PARTICIPANT): Lansing Tower Western 45 ready for departure runway 10R.

(EXPERIMENTER): Western 45 you are cleared for departure. Fly runway heading climb and maintain 3,000’.

(PARTICIPANT): Fly runway heading climb and maintain 3,000’ Western 45

After reaching 1000’ AGL

(EXPERIMENTER): Western 45 contact Lansing Approach on 133.475.

(PARTICIPANT): Contacting Lansing Approach on 133.475 Western 45.

(PARTICIPANT): Lansing Approach Western 45 is with you heading 100 climbing to 3,000.

(EXPERIMENTER): Western 45 roger.

(EXPERIMENTER): Western 45 turn right heading of 190.

(PARTICIPANT): Turning right to a heading of 190 Western 45.

(EXPERIMENTER): Western 45 turn right heading of 280 and descend to 2,500’.

(PARTICIPANT): Turning right to a heading of 280 and descending to 2,500’ Western 45.

(EXPERIMENTER): Western 45 turn right to a heading of 010.

(PARTICIPANT): Turning right to a heading of 010 Western 45.

(EXPERIMENTER): Western 45 turn right to a heading of 060 cleared for the ILS 10R contact Lansing Tower 119.9.
(PARTICIPANT): Contacting Lansing Tower on 119.9 Western 45.

(PARTICIPANT): Lansing Tower this is Western 45 on the ILS 10R.

(OBSERVER: Western 45 you are cleared to land runway 10R.

(PARTICIPANT): Cleared to land runway 10R Western 45.

When aircraft stopped on the runway after landing

(EXPERIMENTER): This session is over. Please relax and I will join you in a few minutes.
Flight Pattern Narration

Flight Pattern 4 KJXN

(EXPERIMENTER): Session start, please begin. Contact tower when ready for takeoff.

(PARTICIPANT): Jackson Tower Western 45 ready for departure runway 24.

(EXPERIMENTER): Western 45 you are cleared for departure. Fly runway heading climb and maintain 3,000’.

(PARTICIPANT): Fly runway heading climb and maintain 3,500’ Western 45

After reaching 1000’ AGL

(EXPERIMENTER): Western 45 contact Lansing Approach on 127.3.

(PARTICIPANT): Contacting Lansing Approach on 127.3 Western 45.

(PARTICIPANT): Lansing Approach Western 45 is with you heading 240 climbing to 3,500.

(EXPERIMENTER): Western 45 roger.

(EXPERIMENTER): Western 45 turn left heading of 150.

(PARTICIPANT): Turning left to a heading of 150 Western 45.

(EXPERIMENTER): Western 45 turn left heading of 060 and descend to 3,000’.

(PARTICIPANT): Turning left to a heading of 060 and descending to 3,000’ Western 45.

(EXPERIMENTER): Western 45 turn left to a heading of 330.

(PARTICIPANT): Turning left to a heading of 330 Western 45.

(EXPERIMENTER): Western 45 turn left to a heading of 280 cleared for the ILS 24 contact Jackson Tower 120.7.

(PARTICIPANT): Contacting Jackson Tower on 120.7 Western 45.
(PARTICIPANT): Jackson Tower this is Western 45 on the ILS 24.

(OBSERVER): Western 45 you are cleared to land runway 24.

(PARTICIPANT): Cleared to land runway 24 Western 45.

When aircraft stopped on the runway after landing

(EXPERIMENTER): This session is over. Please relax and I will join you in a few minutes.
Appendix E

Human Subjects Institutional Review Board
Letter of Approval
Date: September 19, 2012

To: Jessaca Spybrock, Principal Investigator
Geoffrey Whitehurst, Co-Principal Investigator

From: Amy Naugle, Ph.D., Chair

Re: HSIIRB Project Number 12-09-09

This letter will serve as confirmation that your research project titled “Multiple Baseline across Subjects Single Case Research Design as an Alternate Methodology for Aviation” has been approved under the expedited category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may only be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., you must request a post approval change to enroll subjects beyond the number stated in your application under “Number of subjects you want to complete the study”). Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: September 19, 2013