8-1995

Autocorrelation in Single-Subject Data: A Meta-Analytic View

Laura L. Methot
Western Michigan University

Follow this and additional works at: http://scholarworks.wmich.edu/dissertations
Part of the Applied Behavior Analysis Commons

Recommended Citation
http://scholarworks.wmich.edu/dissertations/1781

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.
AUTOCORRELATION IN SINGLE-SUBJECT DATA:
A META-ANALYTIC VIEW

by
Laura L. Methot

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Psychology

Western Michigan University
Kalamazoo, Michigan
August 1995
Recent work by Huitema and McKean (1991, 1994a, 1994b, 1994c, in press) has shown that the most frequently used statistical methods for performing conventional time-series analyses lead to gross distortions of results when these approaches are applied in the context of the typical behavioral research study. Most of these problems could be avoided if researchers were aware that the time-series methods recommended in many areas are not generally needed. The appropriate evidence regarding the need for complex time-series methods requires a meta-analysis of the autocorrelation present in behavioral studies. The project involved: (a) sampling several hundred research articles published in the Journal of Applied Behavior Analysis (JABA) during 1990-1995, (b) extracting the quantitative data from the time-series data displays in 200 sample articles, (c) analyzing each data set using recently developed statistical methods, and (d) performing a meta-analysis of the individual analyses using a new methodology developed for time-series data.

Conventional lag-1 autocorrelation estimates $r_1$ (known to be negatively biased) were computed on (a) the raw data and (b) the residuals obtained from regressing the raw data on time in order to remove the linear trend. New unbiased autocorrelation estimators ($r_{\text{F1}}$ and $r_{\text{HM}}$) were applied to both the raw and detrended data. New tests of the homogeneity of the coefficients in the distributions of $r_{\text{HM}}$ and $r_{\text{F1}}$ were computed, and confidence intervals on the means of the distributions were computed.
Finally, new tests of the significance of the obtained $r_{F1}$ and $r_{HM}$ values were computed.

Results indicate that recent conclusions regarding the presence of the autocorrelation "problem" in typical behavioral experiments should be questioned. When the unbiased estimator ($r_{F1}$) is used, there appears to be a significant proportion of behavioral data sets exhibiting autocorrelation, generally in the positive direction. Most of the apparent positive autocorrelation can be explained by linear trend in the data. When the data are appropriately detrended the autocorrelation distribution is similar to that expected by chance when sampling from a population of independent (nonautocorrelated) errors. The conclusion that behavioral data are not highly autocorrelated is drawn. In most cases the explanation for "apparent" autocorrelation is not that the errors are autocorrelated. Rather, the statistical model used to explain the data has been misspecified. Artificial autocorrelation is to be expected if trend (linear or otherwise) and other deterministic components in the data have not been specified in the statistical model. Conventional general linear model solutions are satisfactory for many studies in applied behavior analysis.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
ACKNOWLEDGMENTS

Many people are deserving of recognition for their contributions to the completion of this dissertation. First and foremost, to my advisor Dr. Bradley Huitema, I extend my sincere appreciation for his constant support, guidance, and timely feedback. Without the exemplary mentoring of Dr. Huitema this project would yet to be off the ground.

To my committee members, Drs. Alyce Dickinson, Alan Poling, and Joseph McKean, I extend appreciation for their versatility in timing and their acceptance of the topic of this study. Thanks go also to Drs. Jack Michael and Richard Spates for their participation in my oral examination and for their helpful comments on the final manuscript.

Finally, I wish to thank my family members who modeled persistence every step of the way. I dedicate this work to Pete Maruh, who provided an always safe haven from the daily demands of academe.

Laura L. Methot
TABLE OF CONTENTS

ACKNOWLEDGMENTS.................................................................................................... ii
LIST OF TABLES ............................................................................................................... vii
LIST OF FIGURES ......................................................................................................... viii
INTRODUCTION ............................................................................................................. 1
  Analysis of Single-subject Data............................................................................... 1
    Visual Analysis...................................................................................................... 1
  Statistical Analysis............................................................................................. 3
    Rationale for Use of Formal Statistical Analyses With Applied Behavioral Data..................................................................................................................... 3
    Assumptions Underlying Conventional Linear Models .................................. 4
The Autocorrelation Debate....................................................................................... 9
  The Independence Assumption and Autocorrelation....................................... 11
  The Meaning of Autocorrelation........................................................................ 11
  Computational Basis of Autocorrelation........................................................... 13
Alternative Intervention Models.............................................................................. 15
  ANOVA Time-Series Model............................................................................... 15
  Interrupted Time-Series Regression Models................................................... 15
  Generalized Least Squares Procedures............................................................ 16
  ARIMA Intervention Models.............................................................................. 16
Problems in Applying ARIMA Intervention Models With Behavioral Data........ 17
  Small Samples..................................................................................................... 18

iii

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table of Contents--Continued

Reduced Biased Autocorrelation Estimator Applied to Linear Detrended Data .........................................................38
t-values on Slope as a Standardized Descriptor of Trend in Sample Data Sets ..........................................................38

Inferential Procedures..................................................................................................................................................39
Test Statistic for the $r_{F1}$ and $r_{HM}$ Autocorrelation Coefficients .................................................................39
Chi-square Tests of the Homogeneity of $r_{F1}$ and $r_{HM}$ Autocorrelation Coefficients ..............................................40
95% Confidence Interval on $\rho$ Based on the Pooled Weighted Mean Autocorrelation Coefficient $\bar{r}_{HM}$ ..............42
Population Variance Estimates of $\rho_j$ and $\varepsilon_j$ ............................................................................................42

RESULTS......................................................................................................................................................................45
Descriptive Statistics ....................................................................................................................................................45
Length of Time-Series Data Sets .................................................................................................................................45
Conventional Autocorrelation Estimator $r_1$ Applied to Original Data .................................................................48
$r_{F1}$ Autocorrelation Estimator Applied to Original Data ........................................................................................48
t-values on Slopes as Standardized Descriptors of Trend in Original Data ...............................................................50
Conventional Autocorrelation Estimates Applied to Linear Detrended Data ............................................................53
$r_{HM}$ Autocorrelation Estimator Applied to Linear Detrended Data ....................................................................53

Inferential Procedures................................................................................................................................................56
$z_{F1}$ and $z_{HM}$ Distributions ..................................................................................................................................57
Chi-square Tests of the Homogeneity of $r_{F1}$ and $r_{HM}$ Autocorrelation Coefficients ..............................................59
Confidence Intervals on $\rho_1$ ......................................................................................................................................61
Estimates of Error Variance, Variance in the $r_{HM}$ Distributions, and Variance Due to Other Sources .................63

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table of Contents--Continued

DISCUSSION .....................................................................................................................66
  Conclusions ...............................................................................................................67
    Efficacy of Reduced Bias Estimators and Linear Detrending .....................67
    Potential Moderating Variables ........................................................................68
    Suggestions for Future Research ......................................................................69
    Summary Recommendations for Analysts .......................................................70

APPENDICES
  A. List of Figures From Which the Conventional Baseline Subsample
     (N=150) of the 1990-1995 JABA Sample Was Drawn ......................................74
  B. List of Figures From Which the Functional Analysis (N=50)
     Subsample of the 1990-1995 JABA Sample Was Drawn ..............................87

BIBLIOGRAPHY ...........................................................................................................91
LIST OF TABLES

1. Example of Residuals Based on the ANOVA Model .............................................. 7

2. Example of Basic and Lagged Residuals ................................................................ 14

3. Number of Data Sets, Mean and Median Number of Observations, and Range of Values in Each Sample Contained in the Study ............................................. 47

4. Unweighted Means on $r_I$, $r_{1R}$, $r_{F1}$, and $r_{HM}$ Computed for Four Samples ........ 48

5. Mean Test Statistics Computed on the $r_I$, $r_{1R}$, $r_{F1}$, and $r_{HM}$ Distributions for Four Samples ............................................................ 57

6. Number of Test Statistics Falling Beyond the ±2.0 Limits for $z_{FI}$ and $z_{HM}$ Distributions for the 1990-1995 JABA Sample, the Conventional Baseline and Functional Analysis Subsamples, and the Busk and Marascuilo Sample ......................................................... 60

7. Chi-Square Obtained Values Associated With $r_{FI}$ and $r_{HM}$ for Four Samples ............................................................ 61

8. Weighted Means on $r_{F1}$ and $r_{HM}$ Computed for Four Samples ......................... 62

9. 95% Confidence Intervals Around $\hat{r}_{HM}$ on the Process Autocorrelations Computed for Four Samples ......................................................... 62

10. Variance Estimates $\hat{\sigma}_r^2$ and $\hat{\sigma}_\varepsilon^2$ Computed for Four Sample Distributions ....... 65

11. Estimated Proportion of Total Variance in $r_{FI}$ and $r_{HM}$ Explained by Process Autocorrelation Variance ($\hat{\sigma}_r^2$) and Error Variance ($\hat{\sigma}_\varepsilon^2$) ................. 65

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF FIGURES

1. Example of Raw Scores Varying Around Phase Means for a Two-Phase Study................................................................. 7

2. Example of Residual Values Varying Around Means of Zero for a Two-Phase Study................................................................. 8

3. Example of Upward Trending Data With a Fitted Regression Line............. 8

4. Frequency Distributions for the Number of Observations in Four Samples of Applied Behavioral Data Sets (Censored for n < 6 Observations). ................................................................. 46

5. Frequency Distributions for the $r_1$ Coefficients Computed on Four Samples................................................................. 49

6. Frequency Distributions for the $r_F$ Coefficients Computed on Four Samples................................................................. 50

7. Frequency Distributions for the $r_{H1}$ Coefficients Computed on Four Samples................................................................. 51

8. $r_F$ Values Plotted Against the Absolute Values of the Standardized Slopes for the 1990-1995 JABA Sample and the Busk and Marascuilo Sample............................................................................. 52

9. Frequency Distributions for the $r_{IR}$ Coefficients Computed on Four Samples................................................................. 54

10. Frequency Distributions for the $r_{HM}$ Coefficients Computed on Four Samples....................................................................... 55

11. Mean Unweighted Autocorrelation Estimators $r_1$, $r_{IR}$, $r_F$, and $r_{HM}$ for Four Samples................................................................. 56

12. Frequency Distributions for the $z_F$ Coefficients Computed on Four Samples....................................................................... 58

13. Frequency Distributions for the $z_{HM}$ Coefficients Computed on Four Samples....................................................................... 59

14. Example of Nonconstant Variance in a Conventional Baseline Phase.................. 72

15. Example of Extreme Instability in a Conventional Baseline Phase............... 72
List of Figures--Continued

16. Example of Ceiling Effect in a Conventional Baseline Phase ..................73
INTRODUCTION

The appropriate analysis of single-subject data has been debated in numerous articles throughout the history of behavior analysis (e.g., Gentile, Roden & Klein, 1962; Jones, Vaught & Weinrott, 1977; Matyas & Greenwood, 1991; Michael, 1974). On one side of the debate are the proponents of visual analysis of graphically presented data (Baer, 1977; Michael, 1974; Baer & Parsonson, 1978, 1992; Parsonson & Baer, 1986; Sidman, 1960), and on the other side are those who assert that more formal statistical analyses are necessary or useful aids in the analysis of behavioral data (Busk & Marascuilo, 1988; Gottman & Glass, 1978; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). Those in favor of more formal statistical analyses are further divided on the issue of the most appropriate formal methods. A brief overview of the history of these issues is provided in subsequent sections to create the relevant context for the problem investigated in this dissertation.

Analysis of Single-subject Data

Visual Analysis

The results of single-subject research are typically presented in graphic form for visual inspection. Data are plotted over time, baseline and intervention phases are clearly identified, and the goal is to determine whether changes in experimental conditions are associated with changes in the behavior from one phase to another. The merits of visual analysis are extolled by Kazdin in the following passage:

Visual inspection is regarded as a relatively unrefined and insensitive criterion for deciding whether the intervention has produced a reliable change. The unsophisticated features of the method are regarded as a
The criterion is somewhat crude, only those interventions that produce very marked effects will lead the scientific community to agree that the intervention produced a change. Weak results will not be regarded as meeting the stringent criteria of visual inspection. Hence, visual inspection will serve as a filter or screening device to allow only clear and potent interventions to be interpreted as producing reliable effects (1982, p. 232).

This passage suggests that visual inspection is viewed as superior to a more formal statistical analysis because the latter may result in the identification of small or weak effects. In fact, behavioral researchers assert that one undesirable characteristic of statistical analysis is its sensitivity in detecting weak intervention effects (e.g., Baer, 1977; Johnston & Pennypacker, 1991). Human judgement of intervention effects via visual analysis of graphed data is not, however, flawless. The lack of concrete decision rules for determining intervention effects has been cited as a potential weakness of the visual inspection method (Kazdin, 1982, p. 239). Further, the reliability of conclusions across judges appears moderate at best (DeProspero & Cohen, 1979; Jones, Weinrott, & Vaught, 1978). Finally, some methodologists suggest that trends and variability in the data may mask effects and render them undetectable by means of visual analysis (e.g., Kazdin, 1982, p. 240). It is not uncommon to encounter phases in which performance gradually deteriorates or improves over time. Some methodologists argue that determining whether trend within one phase is simply a continuation of trend from a previous phase, or if an actual treatment effect is present, can be difficult to judge visually. Unless there is an obvious change in level or slope from one phase to the next, treatment effects might go undetected (Barlow & Hersen, 1984). While most behavior analysts argue that the answers to these problems lie in generating better experimental control to minimize intrasubject response variability and to maximize intraphase response stability (e.g., Johnson & Pennypacker, 1991; Michael, 1974; Sidman, 1960), others suggest that this is not always possible,
particularly when operating within the constraints of applied research (e.g., Barlow & Hersen, 1984).

**Statistical Analysis**

Statistical analyses are used to describe and summarize data and to help determine whether different experimental conditions are associated with differences in the nature of the behavior under two or more levels of an independent variable. Statistical procedures have generally been employed with sets of data collected from groups of subjects, but may also be used to compare data gathered from a single subject measured repeatedly across time. Data collected in this manner from a single subject are frequently described as a time-series and the statistical analysis of such data is often described as a time-series analysis.

**Rationale for Use of Formal Statistical Analyses With Applied Behavioral Data**

The rationale for supplementing visual with statistical analysis in behavioral studies generally follows one of two lines of reasoning. First, detecting intervention effects through visual inspection of data requires a stable baseline. When baseline data show trends, the argument goes, intervention effects may be masked. Similarly, when nontrending baseline data show large variability it may be difficult to detect relevant functional relations (Kazdin, 1982, p. 242-244). The resultant argument is that statistical decision aids may be valuable in such cases to conclude that the intervention did or did not affect the behavior of interest.

A second rationale for supplementing a visual analysis with a statistical analysis is that small effects may be important (Kazdin, 1982, p. 244); it is an advantage rather than a disadvantage if statistical analyses are more sensitive than visual analysis for detecting treatment effects. It should be pointed out, however, that the data supporting
the claim of higher sensitivity for statistical analysis in the case of behavioral data is not particularly convincing (Huitema, 1986b).

Statistical analyses for single-subject data have received much attention since Gentile et al. (1962) proposed that the analysis of variance (ANOVA) be employed in analyzing data from single-subject designs. This method of analysis entails comparison of the means of two or more phases based on the data collected from a single subject. Many other statistical methods have been proposed to compare behavior under different conditions. A general inferential question posed by all methods is, "Do the data provide enough evidence to conclude that the intervention had an effect?" The adequacy of the inferential statements associated with all statistical methods depends upon the extent to which the data conform to the assumptions of the statistical model underlying the analysis.

Assumptions Underlying Conventional Linear Models

Inferential statistical procedures involve fitting a model to the data. A statistical model is an equation that explains behavior on the dependent variable using one or more parameters. The models sparking the debate over the applicability of inferential procedures in the analysis of behavioral data are analysis of variance models.

There are three assumptions underlying the analysis of variance. It is assumed that the errors of the model are independent, that they have the same variance under each treatment condition, and that they are normally distributed. Central to the current debate regarding the analysis of behavioral data is the assumption of independence. In order to understand the assumption of independence it is helpful to look at the components of the ANOVA model. The ANOVA model in the case of a one-factor between subjects design is as follows:
\[ Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \]

where \( Y_{ij} \) is the observed score for the \( i \)th subject in the \( j \)th group, \( \mu \) is the population grand mean for all groups in the study, \( \alpha_j \) is the effect of falling in the \( j \)th group, and \( \varepsilon_{ij} \) (called the error component) is the discrepancy between the \( i \)th subject's observed score and the mean (\( \mu_j \)) for the \( j \)th group.

For any collection of sample raw scores, there will be a corresponding collection of residuals, defined as

\[ Y_{ij} - \bar{Y}_j = e_{ij} \]

where \( Y_{ij} \) is the observed score for the \( i \)th subject of the \( j \)th group, \( \bar{Y}_j \) is the mean of the group from which the score was drawn, and \( e_{ij} \) is the resulting residual, which may be plotted as a frequency distribution in the same manner as the raw scores. These residuals are estimates of the errors in the ANOVA model; they provide information on the extent to which the observed sample data conform to the three assumptions. The errors of the ANOVA model represent random variation that is unexplained by the other terms in the model.

Similarly, the simple linear regression model is written as follows:

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

where \( Y_i \) is the observed score for the \( i \)th subject. \( \beta_0 \), the population intercept, is the elevation on \( Y \) when \( X \) is equal to zero. \( \beta_1 \), the population slope, is the number of units increase or decrease on \( Y \) for one unit increase on \( X \). \( X_i \) is the level on the predictor variable associated with \( Y_i \), and \( \varepsilon_i \) (the error component) is the discrepancy between the \( i \)th subject's observed score and the expected value of \( Y_i \) given \( X_i \) (i.e., \( \beta_0 + \beta_1 X_i \)). As with the ANOVA model, for any collection of sample raw scores there will be a corresponding collection of residuals associated with the regression model.
(defined as \((Y_i - \hat{Y}_i = e_i;\) where \(\hat{Y}_i\) is the value of \(Y_i\) predicted from the fitted regression line). The residuals of the regression analysis are used to evaluate the adequacy of the regression model as a description of the data in the same way that the residuals of an ANOVA are used to evaluate the adequacy of the ANOVA model in describing data.

While the analysis of variance and regression models specified above were described as appropriate for explaining the behavior of various subjects in "group" designs, these models can be written in a modified form to explain the behavior of a single subject in a time-series design. In the case of the ANOVA model the behavior to be explained is \(Y_{ij}\); this is the dependent variable score for the one subject in the study measured at time \(t\) under condition \(j\). The explanatory terms are \(\mu, \alpha_j,\) and \(e_{ij}\) where \(\mu\) is the overall level of the time-series, \(\alpha_j\) is the effect of the condition applied during phase \(j\) of the experiment, and \(e_{ij}\) is the error defined as \(Y_{ij} - [\mu + \alpha_j]\). It is assumed once again that the errors are independent with constant variance across all phases, and normally distributed. Table 1 presents an example of a set of residuals computed by subtracting the phase mean from each of a set of raw scores. The original raw scores and their corresponding residual values are plotted in Figures 1 and 2. As can be seen in Figure 1, the raw scores vary around the phase means of 9 and 59 while Figure 2 shows that the residual values vary around a mean of zero for each phase.

Correspondingly, the simple linear regression time-series model explains the behavior \((Y_t)\) of the one subject in the study at time \(t\) rather than the behavior of various subjects at one point in time. The explanatory variable is time rather than \(X\). Hence the model is written: \(Y_t = B_0 + B_1t + \epsilon_t\). It is assumed that the errors are independent with the same variance at each time point and normally distributed. Figure 3 shows a sample data set of 10 observations. An upward trend is clearly evident in this graph. The
Table 1
Example of Residuals Based on the ANOVA Model.

<table>
<thead>
<tr>
<th>Observation number</th>
<th>Raw score $Y_{ij}$</th>
<th>Phase Mean $\bar{Y}_j$</th>
<th>Residual $e_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>9</td>
<td>+2</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>9</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>9</td>
<td>-2</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>9</td>
<td>+1</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>59</td>
<td>+1</td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>59</td>
<td>-2</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>59</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>61</td>
<td>59</td>
<td>+2</td>
</tr>
<tr>
<td>10</td>
<td>59</td>
<td>59</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1. Example of Raw Scores Varying Around Phase Means for a Two-Phase Study.
Figure 2. Example of Residual Values Varying Around Means of Zero for a Two-Phase Study.

Figure 3. Example of Upward Trending Data With a Fitted Regression Line.
A fitted regression line has a slope \((b_1)\) of 2.72; that is, \(\hat{Y}\) increases 2.72 units for every one unit increase on \(X\). The sample intercept \((b_0)\) is equal to 24.3, and the residual value is computed by subtracting \(\hat{Y}\) (the level on \(Y\) predicted by the regression line) from the raw score \(Y\). The terms \(b_0, b_1,\) and \(e_{ij}\) are sample estimates of the unknown population values \(\beta_0, \beta_1,\) and \(\epsilon_i\).

Violation of the independence assumption results in invalidation of the ANOVA and linear regression \(F\) tests because the probability value associated with the test will be conservative or liberal, depending on the degree and direction of dependency among the errors (Huitema, 1985). When applying inferential procedures such as ANOVA with between-group studies involving independently treated subjects, there is little question regarding the independence of the errors. The question arising with respect to behavioral time-series data, however, is "to what extent can the errors of data gathered over time from a single subject be considered independent?" This question can be investigated empirically by computing coefficients that estimate dependency among the errors of the relevant model. Such coefficients are called autocorrelations. If the errors are autocorrelated an alternative model should be considered. But it is important to thoroughly understand the nature of the "relevant" autocorrelation before considering the issues in selecting an alternative model.

The Autocorrelation Debate

A series of papers appearing in the 1970s and 80s argued for alternatives to conventional parametric statistical tests in the analysis of behavioral data (Busk & Marascuilo, 1988; Hartmann, Gottman, Jones, Gardner, Kazdin, & Vaught, 1980; Jones, Vaught, & Reid, 1975; Jones et al., 1977; Jones et al., 1978; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). The justification for this argument rested on the notion that behavioral data necessarily violate the independence...
assumption underlying the analysis of variance and other variants of the general linear model. The lack of independence was ostensibly illustrated empirically through the computation of non-zero autocorrelation coefficients on single-subject data sets (Busk & Marascuilo, 1988; Jones et al., 1977; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). Additionally, the logical stance that human performance at one point in time is influenced by performance at an earlier point was taken as evidence that behavioral data violate the independence assumption (Busk & Marascuilo, 1988; Jones et al., 1977; Sharpley, 1988).

Interrupted time-series models that contain parameters to explain the assumed nonindependence have been widely recommended as necessary alternatives for the statistical analysis of behavioral data (e.g., Jones et al., 1975, 1977; Hartmann et al., 1980). Huitema (1985, 1986a, 1986b, 1988) asserts that the recommendation to routinely use these techniques is unfounded. He notes that "because applied behavior data are collected across time it has been presumed that time-series analysis is required" (1986b, p. 225). A fundamental misunderstanding of the assumption of independence is the foundation for the continued erroneous arguments against the use of conventional general linear model procedures (including time-series regression intervention models and the analysis of variance) and in favor of formal (e.g., ARIMA) time-series analysis of behavioral data. Clearly earlier attempts to explicate this critical assumption (Huitema 1985, 1986a, 1988) have not been entirely successful. The purposes of the three following sections are to summarize the "autocorrelation problem" and its relationship to the independence assumption, review the body of literature that has been generated by the debate over this alleged problem, and describe the intent of the research carried out in this dissertation.
The Independence Assumption and Autocorrelation

It has been pointed out in all papers that argue for the use of complex time-series methods that the independence assumption will be violated in the case of behavioral time-series data. This presumption has been frequently supported with empirical evidence regarding the autocorrelation coefficients associated with behavioral data. Unfortunately, conceptual and computational errors appear to surround all of these empirical studies (Huitema, 1986a, 1988, 1990). The conceptual and computational issues associated with autocorrelation coefficients and independence are presented next.

The Meaning of Autocorrelation

Inaccurate verbal definitions of autocorrelation as well as computational inaccuracies characterize discussions of autocorrelation in the behavioral literature. Autocorrelation, or serial dependence, in a time series has been defined as follows: the level obtained at one point in the series is to some extent predictable based on knowledge of the levels at previous points in the series (Glass, Willson, & Gottman, 1975). But this definition is misleading; it has been pointed out (Huitema, 1985, 1986a) that the level of behavior at one point in the series is somewhat predictable from the previous data in the series even when there is absolutely no autocorrelation. It may be appropriate to avoid verbal definitions of autocorrelation and to instead focus on the nature of the computations involved in computing autocorrelation coefficients.

An autocorrelation coefficient is a correlation computed between the observations (or residuals) of a time series and the lagged values of that series (Huitema, 1994). It is possible to compute autocorrelations on raw data or the
residuals of any number of statistical models, the relevant computation being dictated by the nature of the data and the question posed by the investigator.

Several papers (Huitema, 1985, 1986a, 1988) have pointed out that earlier investigators (e.g., Busk & Marascuilo, 1988; Jones et al., 1975; Jones et al., 1977; Jones et al., 1978) based their recommendations for using complex time-series analysis techniques on the faulty assumption that it is autocorrelation of the observed data that invalidates the use of procedures based on the general linear model, such as the analysis of variance. To state, though, that behavioral data are not necessarily plagued with an autocorrelation problem, as Huitema has done (1985, 1986a, 1988), has generated the erroneous logical position that (a) if there is no autocorrelation in behavioral data, then (b) people behave randomly over time, thus (c) the prediction and control of behavior is nonexistent (see Huitema, 1986b). He points out that this argument is faulty because zero autocorrelation of the errors $e_t$ of a statistical model does not imply that behavior is unpredictable. Behavior at any point in time in a series may be predicted using the mean of the data in a non-trending series, or by using a regression equation fitted to the data when a trend is present.

The presence of autocorrelation in the residuals of either ANOVA or simple linear regression generally means that the correct statistical model has not been identified. A regression model, for example, will provide a better fit to data trending over time than a model based on the mean. Similarly, some environmental determinant of the behavior under study can cause autocorrelation of the residuals if the environmental event is not included in the model. Thus, the logic that no autocorrelation means that behavior is random and unpredictable does not hold because the statistical models used to explain behavior contain both deterministic and random components. The lack of autocorrelation of the errors simply means that the error at
one point in time is not predictable from previous errors; this says nothing about the predictive usefulness of the other terms in the model (Huitema, 1986a).

**Computational Basis of Autocorrelation**

Serial dependence of the errors of a model is assessed by computing a correlation between the residuals of the time series and the lagged values of the series. A lag-1 residual autocorrelation, for example, is derived by pairing residuals at time $t$ with those at time $t + 1$ and computing a correlation coefficient similar to (but not identical with) the Pearson coefficient. The result, which is labeled $r_1$, is a summary measure of the degree of linear dependence between contiguous values. Table 2 (adapted from Huitema, 1994) contains columns of residuals that are relevant in computing $r_1$ through $r_{50}$. That is, the values in column $e_t$ are correlated with those in column $e_{t+1}$ to compute $r_1$; those in column $e_t$ are correlated with those in $e_{t+2}$ to compute $r_2$, and so on. Huitema states that "it is possible to compute as many autocorrelations as lags, but generally only about one-sixth to one-fourth of all possible autocorrelations are computed" (1994, p. 430). This limit on the number of coefficients is recommended because the autocorrelation estimates become less stable as the lag increases. Note that the lag-1 autocorrelation in Table 2 is based on 199 pairs of numbers, but the lag-50 autocorrelation is based on 150 pairs of numbers (i.e., 200 observations in the original series minus 50 lags). One pair of numbers in the computation is lost for every lag, a fact which severely impinges upon the stability of sample autocorrelations for small data sets. The whole set of sample autocorrelation coefficients (e.g., $r_1$ through $r_{N/4}$, where $N/4$ is the number of lags considered in the computations) is sometimes called the sample autocorrelation function. The conventional formula for the lag-$k$ autocorrelation coefficient is:
Tests of significance can be computed on the sample autocorrelation coefficients to determine whether the underlying population series (called the process) of errors of some specified model is autocorrelated. Many tests are available for this purpose (e.g., Bartlett, 1946; Box & Pierce, 1970; Huitema & McKean, 1991, 1994c, submitted; Ljung & Box, 1978). If the tests suggest that autocorrelation is present it is concluded that the assumption of independent errors is violated. The identification of autocorrelated errors means that the model of behavior (Y_t) is misspecified and that an alternative model is required to explain the data.

Table 2

Example of Basic and Lagged Residuals

<table>
<thead>
<tr>
<th>Time</th>
<th>e_t residual at Time t</th>
<th>e_{t+1} residual at Time t + 1</th>
<th>e_{t+2} residual at Time t + 2</th>
<th>...</th>
<th>e_{t+50} residual at Time t + 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>e_1 = -2.33</td>
<td>e_2 = 4.67</td>
<td>e_3 = 0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>e_2 = 4.67</td>
<td>e_3 = 0.67</td>
<td>e_4 = -5.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>e_3 = 0.67</td>
<td>e_4 = -5.33</td>
<td>e_5 = 6.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>e_4 = -5.33</td>
<td>e_5 = 6.67</td>
<td></td>
<td></td>
<td>e_{200} = -4.3</td>
</tr>
<tr>
<td>5</td>
<td>e_5 = 6.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td></td>
<td>e_{200} = -4.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>e_{200} = -4.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alternative Intervention Models

If a researcher is interested in statistically analyzing the data from a behavioral intervention study, there are several approaches that can be taken. Among them are the ANOVA model, an interrupted time-series regression model, an ARIMA intervention model, and a new "double bootstrap" method. Each of these methods is most appropriate under different conditions. A brief description follows.

ANOVA Time-Series Model

If there is no apparent trend in the data, that is, the data appear stationary around some mean value within each phase of the study, then the ANOVA model may provide an appropriate fit. This model has been found to be adequate for many behavioral time-series (Huitema, 1985, 1986a). It is not highly unusual, however, to encounter autocorrelation with these designs. To rule out the presence of serial dependency, the residuals of the ANOVA model should be assessed. While software for conventional tests for autocorrelated errors of this model exist (they are known as the runs test and the Durbin-Watson test), recent research suggests that these tests are inappropriate in the case of samples of the size generally found in applied behavior analysis research. An alternative test proposed by Huitema and McKean (1994d) is both simple and effective for ANOVA and other variants of the general linear model applied to time-series data. A statistically significant outcome of this test implies that a more complex model should be considered.

Interrupted Time-Series Regression Models

If the residuals of the ANOVA model are significantly autocorrelated it is possible that within-phase trend in the data is responsible for the autocorrelation,
which case an appropriate interrupted time-series regression model should be identified. If this model is applied to the data and the residuals do not have significant autocorrelation, the conventional tests associated with the model are valid. It has been found that a general linear model which allows for a different slope in each phase of the experiment often fits behavioral data quite well (Huitema, 1985, 1988). Details on the regression approach for the analysis of time-series designs containing two, three, or four phases can be found in Huitema, McKean, and McKnight (1994). Once again, if the residuals of this approach are autocorrelated, some more complex alternative model should be considered. The form of the increased complexity may be in terms of added independent variables or parameters to model directly the errors of the model.

**Generalized Least Squares Procedures**

When the residuals of a time-series regression model are autocorrelated, a method of modeling the autocorrelation should be found. A well known approach for solving this problem is known as generalized least-squares. Two versions of this approach have been studied recently (McKnight, 1994); both were found to be clearly unacceptable in the case of small samples.

**ARIMA Intervention Models**

ARIMA intervention models (Box & Tiao, 1965, 1975; Glass et al., 1975) provide a general approach to describing time-series data. It is one of the approaches that is at the basis of the ongoing debate surrounding the statistical analysis of behavioral data. Many investigators recommend the use of ARIMA modeling over ANOVA or other conventional models for behavioral data because of the alleged autocorrelation problem (e.g., Hartmann et al., 1980). Some have attempted to explain ARIMA modeling in terms that can be readily understood by behavioral researchers to
enable them to use ARIMA procedures effectively (e.g., Jones et al., 1977; Hartmann, et al., 1980). Unfortunately, ARIMA techniques are complex and much has been lost in the attempted translation to simplicity.

General ARIMA models are specified using three components \((p,d,q)\); where \(p\) is the autoregressive (AR) component, \(d\) is a term that describes a maneuver called differencing that removes trend from the data set, and \(q\) describes the moving average (MA) component\(^1\). Three numbers specify the order of the three components. ARIMA \((0,0,0)\) specifies the white noise model (Glass et al., 1975), where no autoregressive component is identified, no differencing is required to remove trends, and no moving-average component is inherent. Data that conform to the white noise model are not autocorrelated and any variation in the data is considered to be a random process. A first-order autoregressive model contains one AR parameter, but requires no differencing or moving average parameters. It is identified as an ARIMA \((1,0,0)\). Complex models can contain autoregressive, differencing, and moving average components of high order (e.g., \(3,2,2\)), but this is rare. The major tools for specifying the order \((p,d,q)\) are the sample autocorrelation and sample partial autocorrelation functions (Box & Jenkins, 1976).

Problems in Applying ARIMA Intervention Models With Behavioral Data

There are at least three problems with attempting to apply ARIMA modeling to behavioral data: (1) there generally is an inadequate number of data points for accurate model identification, (2) differencing to remove trends from the data introduces both

\(^1\) Moving averages is a complex phenomenon, better labeled as a "moving sum." The MA model describes the error at time \(t\) as a function of previous and current "random shocks" to the system. Most behavioral time series do not involve MA components, thus it is neither necessary nor desirable to attempt an explanation of moving averages here.
statistical problems and interpretation issues, and (3) a subjective component is involved in model identification. Each of these will be considered subsequently.

**Small Samples**

The foremost problem with applying ARIMA modeling to behavioral data is related to the number of data points typically available in applied behavioral studies. Box and Jenkins (1976) state that a minimum of 50 observations is necessary to obtain a stable estimate of the true autocorrelation function. The modal number of observations in baseline phases of studies published in *JABA* from 1968 to 1977 was about 4 (Huitema, 1985). Clearly this number comes nowhere near the recommended minimum. Large samples of time-series data points are needed "for knowledge of $p$, $d$, and $q$, so that the dependence among observations can be properly accounted for in statistical tests of intervention effects" (Glass et al., 1975, p. 112). Huitema (1988) clarifies:

Since the identification of an ARIMA model depends upon the estimation of a substantial number (say a dozen) of reasonably stable autocorrelations, the $n = 50$ recommendation is sensible. It must be understood that a major purpose of the autocorrelation function in ARIMA modeling is the identification of an appropriate model. This is accomplished by comparing the estimated autocorrelation and partial autocorrelation functions with the autocorrelation and partial autocorrelation functions known to characterize various ARIMA models. (p. 271)

Jones et al. (1975) and subsequent investigators (Busk & Marascuilo, 1988; Sharpney & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987) have failed to recognize the importance of model identification. Further, Sharpney (1987) has suggested that model identification may not even be necessary, stating that "the argument that accurate model-identification is difficult with short series is based on the assumptions that exact model-identification is both possible and necessary" (p. 41). He goes on to say that "it would appear that the need to be specific in model-identification is not so obvious as..."
was once thought" (p. 41). This position is indeed peculiar when he later reports on a personal communication from Glass (November 10, 1984):

> The often-quoted figure of 50 to 100 observations as necessary for the reliable use of ITSA [interrupted time-series analysis] is a vast overestimate, and...as few as eight baseline and eight intervention observations may be sufficient to test for treatment effects, providing the model of the series is known [italics added] (p. 42).

Sharpley seems to be saying here that accurate model identification is difficult and unnecessary, but that in order to use interrupted time-series analysis accurately, the model of the series must be known.

**Differencing to Eliminate Trend**

Differencing is a procedure for removing trend from a data set. Glass et al. (1975) describe the procedure as follows:

If the series is not stationary, successive differences are taken until the resulting series is stationary; for example, if the values of a series are 1, 3, 5, 7, 9, etc., there is a linear trend. First differencing reduces the series to 2 (=3-1) [sic], 2 (=5-3), 2 (=7-9) [sic], 2 (=9-7) [sic] and the resulting series of all 2's is clearly stationary. If the values of a series are 1, 4, 9, 16, 25, 36, there is a quadratic trend. First differencing gives 3, 5, 7, 9, 11, and differencing again gives 2, 2, 2, 2. In general, first differencing eliminates linear trend, second differencing eliminates quadratic trend, and so on (p. 79).

While trend in a baseline data set may mask intervention effects under some circumstances (e.g., where the trend is in a therapeutically desirable direction), behavior analysts do not typically recommend statistically removing the trend. Rather, the preferred option is to wait for the data to stabilize before further experimental manipulations are initiated. Sidman (1960) asserts that when an experiment produces unstable behavior, the investigator must rule out extraneous variables that may influence the dependent measure. Once the experimenter is assured that the instability is a product of the experimental manipulation, the next task is to "examine the instability and describe it as fully as his [or her] available methods will allow" (p. 271). The
rationale for this recommendation is that accurate description of the form of the
instability and its relevant controlling variables is necessary if subsequent experimenters
are to successfully deal with the phenomenon as it arises.

Kazdin (1982) states that "a problem may emerge, at least from the standpoint
of the design..." (p.263) when trend in a therapeutically desirable direction is apparent
in baseline data. In this case it may be difficult to detect intervention effects. He
presents several design alternatives for dealing with trending data, including variants on
the reversal, multiple-baseline, and changing-criterion designs. The differential effects
of treatment can often be detected under these design options, even when there is
overall trend in the data. In cases where trending overshadows or makes ambiguous
the intervention effects, a statistical aid may be employed.

Subjective Nature of ARIMA Model Identification

Even with large sample sizes, ARIMA model identification is complex, time
consuming, and inefficient, especially for "mathematically unsophisticated researchers"
(Velicer & McDonald, 1984, p. 39). Human judgement is required at several stages in
the procedure, and there is "no reason to assume that given a standard set of data, all
persons would arrive at exactly the same model" (Horne, Yang, & Ware, 1982, p.
187). Michael's (1974) advice regarding the use of complex inferential techniques is
skeptical:

The various judgemental aids do not achieve their simplifying effects
without some cost...they are easier to react to in part because they are
abbreviations...Further, the scientist must spend some time learning
about them, time he [or she] might be spending in other activities
relevant to his [or her] subject matter (p. 647).

The amount of effort required to gain the necessary repertoire for accurate model
identification, were it possible, is significant. The cost of such complex procedures, to
both the behavioral researcher and the research consumer, in terms of time and effort

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
might greatly outweigh any benefit accruing from their use. Recent developments in automated model identification algorithms (Liu, 1993) may reduce this objection to ARIMA modeling. It is obvious from the current work on such algorithms, however, that they are not designed to be used with sample sizes that are typical in behavioral psychology. The shortest series in a large collection of recent examples demonstrating automated identification (Liu, 1993) had 70 observations.

New Double Bootstrap Procedure

The poor small-sample performance of the conventional versions of the generalized least squares procedure and several problems that accompany ARIMA modeling motivated the development of a new approach. This approach employs a computer intensive algorithm known in statistics as the "bootstrap" during two stages of the analysis. The original evaluation of the application of this method to the analysis of the interrupted time-series design reported very satisfactory results (McKnight, 1994). Current work confirms this evaluation (McKnight, McKeen, & Huitema, in preparation).

Benefits of GLM and New Double Bootstrap Models Over ARIMA Modeling

There are at least two benefits associated with using the conventional general linear model and the new double bootstrap methods when analyzing behavioral data. First, they are not bound by large-sample requirements in the same manner as are the ARIMA techniques. Second, they produce results that are easier to understand than those produced through ARIMA modeling. Each of these points will be considered below.
Solution to Small Sample Problems

It has been repeatedly found that a relatively large N (at least 50) is required for adequate estimation of autocorrelation parameters and model identification when employing either ARIMA procedures (e.g., Box & Jenkins, 1976; Horne et al., 1982; Huitema, 1985, 1986a, 1988; McKnight, McKeen, & Huitema, in preparation) or generalized least squares (McKnight, 1994). Further, it is evident that the number of observations typically generated in studies published in *JABA* comes nowhere near the recommended minimum of 50. Rather, they typically contain fewer than ten observations per phase (Huitema, 1985).

While small sample sizes are often an issue when employing conventional general linear models such as ANOVA, the concern is one of power rather than of the stability of model identification and parameter estimation; that is, there is a potentially low probability of concluding there is an experimental effect when in fact one has occurred. However, the power of an inferential test is also related inversely to the amount of within-phase variation. Thus, where experimental control is effective and data show little variability within phases, the power of tests based on the general linear model can be relatively high even in small-N cases.

In the case that the errors of the general linear model are autocorrelated, the new bootstrap procedure mentioned above has been shown to be effective with small-N data sets. This procedure enables the researcher to accurately estimate autocorrelation parameters and intervention effects without relying on large-sample theory that is of questionable relevance in the case of small samples.
Type III Errors

Behavior analysts generally favor the most direct approach to data analysis: visual inspection of the graphed data to determine if intervention effects are evident. This approach leaves little room for speculating on which aspects of the stimulus array generated by graphed behavioral time-series data control the investigator's conclusion that an effect did or did not occur. Said another way, when we visually inspect data, we can point to the parts of the graph that have led us to our conclusions regarding the effectiveness of the treatment. Similarly, when conventional general linear models such as ANOVA are used, the single-subject researcher can easily point to the relevant aspects of the stimulus array to explain what the significance test is comparing. Unfortunately, behavioral researchers are now encouraged to use software that will fit complex ARIMA intervention models. A problem associated with processing one's data through such software is the high likelihood of misunderstanding the nature of the comparison generated. Huitema (1986b) calls this a Type III error. He points out that the output generated by complex time-series intervention analysis (ARIMA) programs contain terms and probability values that are likely to answer questions that are very different from the question the researcher thinks is being answered.

Summary of the Major Methodological Problems Associated With Previous Surveys of Autocorrelation in the Journal of Applied Behavior Analysis

Many misunderstandings regarding autocorrelation and time-series analysis have been published in the behavioral literature. A detailed account of most of these issues can be found in several critical articles (Huitema, 1985, 1988) and chapters (Huitema, 1986a, b). The argument underlying all published recommendations to use ARIMA (and other) complex time-series methods in the analysis of applied behavior analytic data rests on the presumption that the errors of such studies are autocorrelated.
This implies knowledge of the distribution of appropriate autocorrelation coefficients associated with a population of applied behavior analytic data. All previous attempts to characterize this distribution have been flawed. The most important flaws associated with each of the earlier attempts have been presented by Huitema (1990). He labeled these errors as (a) model misspecification type I (the problem of the 1970s), (b) model misspecification type II (the problem of the 1980s), and (c) the universal estimation problem; these are summarized in the remainder of this section.

The first problem is associated with the work published in the 1970s; an early study by Jones et al. (1977) involved the computation of autocorrelation coefficients on 24 sets of raw data from studies published in the *Journal of Applied Behavior Analysis* (*JABA*). The authors concluded that 83% of the studies had significant high levels of autocorrelation and that ARIMA time-series models were required for the analysis of these studies. The principal recommendation of the Jones et al. (1977) paper that complex time-series analyses be used as a supplement to visual analysis of behavioral data is questionable because it is based in irrelevant autocorrelations. The error committed in this work can be viewed as a type of model misspecification. The model the researchers used in conceptualizing their analyses was not formally stated in their work, but verbal descriptions of their methods, the obtained autocorrelation coefficients, and subsequent reanalyses of their data leave no ambiguity whatsoever that the following statistical model captures the strategy they used:

\[
(1970s \text{ Model}) \quad Y_t = \mu + e_t. \tag{1}
\]

This model implies that behavior is a function of \( \mu \), the overall level of the whole observed series (i.e., data collected across all conditions or phases), and \( e_t \), random error. It is known that this model misspecifies the actual form of the behavior in any time-series in which there is a difference in level between phases (Huitema, 1985,
1986a). Because the model does not contain a term that acknowledges differences in the levels of the different phases, the model is said to misspecify the sources of variation in the experiment. This means that if there are different levels of behavior in the different phases, artificial autocorrelation will necessarily be introduced to the errors.

While the authors of the work carried out in the 1970s claimed that their autocorrelation results were proof that data analysis methods based on the general linear model (such as independent sample t-tests and ANOVA F-tests) were inappropriate for behavioral time-series data, their results were shown to be completely irrelevant to such models. The autocorrelation coefficients computed under the 1970s model provide no information on the autocorrelation associated with the errors of any realistic intervention model (Huitema, 1986a). For example, if the ANOVA model describes the data, the residuals for a study with two or more phases are based on deviations from each phase mean rather than from one common mean for the entire series. If the raw data are inappropriately combined into a single series, a high positive autocorrelation coefficient will generally be obtained, even if the residuals of the ANOVA model have zero or negative autocorrelation (Huitema, 1985, 1986a). Unfortunately several subsequent studies appeared, also claiming that formal time-series analysis of behavioral data is not only desirable but necessary (e.g., Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987).

The problem just described was acknowledged in the subsequent surveys of autocorrelation in behavioral data carried out in the late 1980s and early 1990s by Busk and Marascuilo (1988) and Matyas and Greenwood (1985, 1991). These researchers also failed to explicitly describe their model, but descriptions of their work imply the following:
This model is much more realistic because it is a representation of the behavior that occurs within a phase rather than across all phases. Autocorrelations computed on the errors of this model are meaningful if the data are not trending upward or downward across time. If the data from two or more phases are conceptualized with one model rather than as a collection of individual phase models, it can be written as

\[ Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}. \]  

(3)

This is the ANOVA model for a time-series design containing multiple phases. The term \( \mu \) is the overall level of the whole series (across all phases), \( \alpha_j \) is the effect of being exposed to condition \( j \) in phase \( j \), and \( \varepsilon_{ij} \) is the error. If the errors of model 2 for each individual phase in a multiple phase design are independent, the errors for the ANOVA model (3) associated with this design will also be independent. The problem with models (2) and (3) is that they are unrealistic in the case of data that trend within phases. This is another case of misspecification error. Trends within phases will introduce artificial autocorrelation to the errors of models (2) and (3) just as differences between levels of phases introduces artificial autocorrelation in model (1). Both level changes between phases and trending within phases are deterministic aspects of the data; such characteristics should be built into the model of behavior. If the model does not contain parameters to explain these characteristics when they exist, the model is misspecified and the errors will be artificially autocorrelated. The models recently studied by Huitema, McKeen, and Zhao (in press) contain parameters for both differences in level and trend for these very reasons.

The third error described by Huitema is called the universal estimation problem. All of the previously published surveys of autocorrelation in behavioral data employed
the conventional autocorrelation estimator. This estimator is known to yield biased estimates. The degree of bias has been shown to be serious in the case of small samples (Huitema & McKean, 1991). This implies that all previous surveys are flawed in the sense that the measure of autocorrelation universally used in these studies is not adequate for the purpose of characterizing the true level of autocorrelation in the field of applied behavioral research. The major purpose of the present research is to provide an unbiased estimate of the level of autocorrelation present in the errors of the population of recently published behavioral studies.
METHODS

Introduction to Meta-analysis

In 1985 Huitema concluded that the residuals of general linear models fit to applied behavioral data have little or no autocorrelation. Given that the typical baseline phase in studies published in *JABA* is around four to six observations, one might question the means by which Huitema came to this conclusion. The approach was to cumulate the results from several hundred applied behavioral data sets in order to summarize the overall level of autocorrelation in the population of behavioral studies. The general approach taken by Huitema (1985) and in the current study is commonly referred to as "meta-analysis."

Logic of Meta-analysis

Meta-analyses are statistical methods for combining the results of a collection of individual studies. The logic behind these procedures is as follows:

[If each] computed effect size from a series of studies estimates a single population effect size, then it is sensible to compute a single pooled estimate that will be far more precise than any of the individual estimates, and the corresponding tests of significance will be substantially more powerful than any of the separate tests (Raudenbush, 1991, p. 242).

Two important features of meta-analytic procedures relevant to the current study, and those of Huitema (1985, 1986a, 1988), are that they enable the researcher to (1) summarize an overall relationship between two variables by cumulating the results of independent studies, and (2) determine which factors, called moderating variables, are
associated with variations in results across individual studies (Rosenthal, 1991). A brief overview of the steps taken in conducting a meta-analysis follows.

**General Steps in Meta-analyses**

The starting point in a meta-analysis is to bring together a collection of studies from a common area, each with their individually computed results. Johnson, Mullen, and Salas (1995) observed that the results of empirical research are typically reported in terms of effect size (i.e., the magnitude or strength of the effect) or significance level (i.e., the likelihood of having obtained the observed results given that the null hypothesis of no effect is true). Several metrics are available for estimating and reporting effect size; three commonly employed methods are Pearson's $r$, the transformation of Pearson's $r$ to Fisher's $Z$, and the standardized mean difference, sometimes known as Cohen's $d$. Similarly, various methods are available for reporting significance levels; commonly used metrics include standard normal deviates ($z$ scores) and probability ($p$) values (Johnson et al., 1995). It is conceivable that a collection of studies from a common area may employ multiple means of measuring the behavior of interest and reporting on effect sizes and significance levels. It is not necessary for the dependent measures of these studies nor their corresponding summary statistics to be in the same metric in order for the results to be meaningfully combined (see Rosenthal, 1991, chapter 2).

Because meta-analyses include statistical tests of many individual analyses, evidence for the appropriateness of pooling the individual results is needed. The consistency of study results can be assessed with a chi-square test of homogeneity of the individual effect size coefficients (Huitema, 1988; Raudenbush, 1991). A test of the hypothesis of homogeneity with a nonsignificant outcome is evidence that the results of the independent studies are indeed based on samples taken from a common
population. Thus, pooling the results is justified. The pooled coefficient is an optimal estimate of a common population effect that is weighted by sample size. Those individual coefficients associated with larger sample sizes are given more weight than are those associated with smaller samples.

Once the pooled weighted estimate is computed, it is then reasonable to construct confidence intervals around the pooled estimate and perform significance tests against the null hypothesis of no effect in the population. Tests of significance on the pooled estimate will be considerably more powerful than any of the tests on the results of the individual studies.

Identifying Moderator Variables

If the hypothesis of homogeneity is rejected, the next step in the meta-analysis is the search for moderator variables: factors independent of the treatment that might affect the outcome of a study. Rosenthal (1991, p. 9) provides an example of how moderator variables can be useful in explaining variation in effect size across several studies. In his example, the effect size of interest is the correlation between teacher expectations and pupil performance in six separate studies. Assume that a test of the homogeneity of the six correlation coefficients was conducted and the p-value produced was small (i.e., \( p \leq .05 \)); the hypothesis of homogeneity would be rejected and we would have no evidence to support pooling those coefficients. Rosenthal provides a second set of numbers, mean ratings of teacher excellence, that might provide an explanation for the inconsistent correlation coefficients found in the six studies. Assume also that a significant relationship between the original effect size estimates (i.e. the correlation between teacher expectations and pupil performance), and the moderator variable (mean ratings of teacher excellence) exists. We could empirically demonstrate that the level of teacher excellence moderates the magnitude of the
relationship between teacher expectations and pupil performance, concluding that the results of the six studies are consistent when the moderator variable is considered in the statistical model assumed to explain the data.

Current Development of Meta-analysis Methods for Autocorrelation Coefficients

A meta-analysis of the degree of autocorrelation in typical applied behavior analytic data is possible using the procedures outlined above with slight modifications. For instance, instead of a collection of effect size coefficients, we have a number of autocorrelation coefficients computed on observations from several behavioral time-series data sets. Because autocorrelation coefficients are not generally found in studies published in applied behavioral journals, the appropriate steps in such a meta-analysis would begin with: (a) sampling a number of time-series data plots from published applied behavioral studies, (b) extracting the quantitative data from those plots, (c) computing individual autocorrelation coefficients for each sampled data set, and (d) computing test statistics on the amount of variation away from zero observed in the distribution of the autocorrelation estimates. A test of the homogeneity of the autocorrelation coefficients—a modified version of a chi-square test designed specifically to assess autocorrelation coefficients—would follow. Depending on the results of the test of homogeneity, one would then either (a) compute a pooled weighted mean autocorrelation estimate (given nonsignificant results from the homogeneity test) and construct confidence intervals around that estimate or (b) search for moderator variables that might account for the apparently inconsistent results of the individual studies and include information on those moderator variables in the appropriate statistical model fit to the data. This was the general routine followed in the current study. Details of sampling procedures and meta-analytic procedures modified for use with autocorrelation coefficients can be found in the sections to follow.
Data Acquisition Procedures

Two primary samples of data were used in this study; one newly sampled from *JABA* and one obtained from Busk and Marascuilo (1988). To remain consistent with previous attempts at identifying autocorrelation in behavioral data (e.g., Busk & Marascuilo, 1988; Huitema, 1985; Sharpley & Alavosius, 1988; Suen & Ary, 1987), the analysis was restricted to data sets with a minimum of six observations within the first phase of the design. By limiting the data to those meeting this phase length criterion, there were no concerns regarding the properties of the autocorrelation estimators or the associated standard errors. Also, issues regarding the distortion of autocorrelation estimates caused by deterministic intervention effects were avoided by focusing on data obtained before interventions were applied.

Sampling Methods for Selecting Data Sets From *JABA*

The period covered in *JABA* from volume 23, issue 1 (1990), through volume 28, issue 1 (1995) was the source of 200 time-series data sets employed in this study. First baselines, defined as the first phase of data collected under baseline or functional analysis conditions prior to any intervention or subsequent conditions, were sampled. Thus, the first A of an ABA design would qualify for inclusion, but the second A (i.e., the reversal or withdrawal phase) would not. Similarly, the first phase of a pairwise test-control design involving multiple assessment phases (Iwata, Dorsey, Slifer, Bauman, & Richman, 1994) would qualify for inclusion but the second and subsequent phases would not. Data sets plotted with some measure of time on the abscissa were included in this finite population. Excluded were cumulative graphs, bar graphs, histograms, and other graphs not plotted as a time series.
Two subcategories of "first baseline" were defined prior to sampling: (1) functional analysis baselines, and (2) baselines from conventional reversal, multiple-baseline and other designs. Functional analyses consisted of conditions in which target behaviors were observed across two or more well defined environments. Data obtained under different functional analysis conditions are plotted on separate data paths and two or more data paths may be presented for a single assessment phase. Under the functional analysis paradigm, physical and social aspects of the environment that might affect the occurrence of a target behavior are manipulated and a base rate of behavior under these conditions is observed (Iwata et al., 1994). Functional analysis baselines are a homogenous group of conditions in which the behavior of interest is typically some form of self-injurious behavior (SIB) or stereotypy, although the latter is studied less often using this assay. Conditions typically employed to assess behavior include social-positive reinforcement, social-negative reinforcement, automatic reinforcement and control conditions (see Iwata, Pace, et al., 1994 for examples). Of the 495 functional analysis first baselines appearing in the JABA issues sampled, 260 contained six or more observations; 50 of these were randomly selected for inclusion in the meta-analysis.

Baselines from reversal, multiple-baseline and other designs consist of a heterogeneous collection of first phases in which some behavior is measured over time under existing environmental conditions. These conditions are free to vary throughout the baseline prior to any intervention. Thus, these baselines contain data from a wide variety of subjects and behaviors and may be based on individual subjects, larger molar units including well defined (stable) groups, or dynamic groups such as the general public in community-based interventions. Of the 1,955 data sets found in this category, 1,173 contained six or more observations; 150 of these were randomly selected for inclusion in the meta-analysis.
Data sets that were selected as the original sample of 50 functional analysis series or as the original sample of 150 conventional baselines were excluded from these samples if all observations fell at the minimum or maximum of the relevant metric (e.g., all at 0% or all at 100%). If such a data set was selected, it was discarded and a new set was drawn randomly\(^2\).

**Data Extraction Procedures**

Quantitative data were extracted from the behavioral time-series graphic data displays using the computer program DataThief (Huyser & van der Laan, 1992). DataThief, developed to aid nuclear physicists in the reverse engineering of data plots, allows the user to generate raw data from existing graphic displays. Graphic displays of sampled data were scanned using a Hewlett Packard ScanJet IIcx and WordScan™ software. Each graphic image was then saved in a PICT format and subsequently opened into the DataThief program. When opened into DataThief, the data plot appears on the screen and the program prompts the user to; (a) point the cursor and click the mouse once on each end of the horizontal axis, (b) enter the values corresponding to each end of the axis, (c) point the cursor and click the mouse once on each end of the vertical axis, (d) enter the values corresponding to each end of the axis, and (e) move the cursor to the first data point to be read and click the mouse once, and point and click on each subsequent data point. When the user has clicked on all the data points, the values are saved as an ASCII file that can be read by any spreadsheet or graphics program. Generated data were imported to Minitab statistical software for analysis.

\(^2\) From here on, the term "data sets" will refer to the individual time series contained in each sample, and the term "sample" will refer to the groupings of time-series data sets subsumed under the 1990-1995 *JABA* sample, the conventional baseline and functional analysis subsamples and the Busk and Marascuilo sample.
Reliability of Data Extraction Procedures

Twenty data sets were randomly selected from the 1990-1995 JABA sample (N=200) to be used for estimating the reliability of data extraction procedures. The PICT formats for these 20 data sets were opened into the DataThief program and the data were extracted a second time following the procedures outlined in the previous section. The conventional autocorrelation estimator $r_1$ was computed for the data extracted from the 20 reliability data sets and for their corresponding 20 original sample data sets using the formula presented on page 13 (Computational Basis of Autocorrelation), where $k=1$. A Pearson correlation coefficient was computed between the $r_1$ values associated with the 20 reliability and 20 original extracted data sets; the correlation between these two sets of autocorrelation coefficients was .998.

Busk and Marascuilo (1988) Data

The second sample of data employed, that used by Busk and Marascuilo (1988), was obtained from the authors for reanalysis. Busk and Marascuilo sampled 248 time-series data plots containing six or more observations that were published in JABA from 1975 to 1985. Of these 248 data sets, 101 first baselines (i.e., the first baseline in an experiment) were included in the current reanalysis (Busk & Marascuilo, 1988).

Computation of Summary Statistics

The computations described below refer to "raw data" and "linear detrended data." Because only single phase baseline data sets were used, applying the formulas to the "raw data" are equivalent to applying them to the residuals of the following linear model: $Y_t = \mu + \epsilon_t$. "Linear detrended data" are the residuals obtained by regressing
the raw data on time in order to remove trend. That is, they are the residuals that estimate the errors in the following model: \( Y_t = \beta_0 + \beta_1 t + \varepsilon_t \). All subsequently described descriptive and test statistics were computed on four samples of data; (1) the 200 newly sampled JABA (1990-1995) first baseline data sets, the subsamples of (2) 150 baselines and (3) 50 functional analysis baselines (separated out of the combined sample of 200), and (4) the 101 first baselines obtained from the Busk and Marascuilo (1988) sample.

**Conventional Autocorrelation Estimator \( r_1 \) Applied to the Raw Data**

Conventional lag-1 autocorrelation estimates \( r_1 \) were computed on the raw data \( (Y) \) of each series. This estimator is known to be negatively biased (Huitema & McKean, 1991, 1994b). If, for example, the process autocorrelation is zero, the expected mean of a frequency distribution based on a large number of sample autocorrelation coefficients would be expected to equal \( r \). It is important to recognize the negative bias inherent in this estimator; the autocorrelation coefficients from the individual studies provide the foundation for subsequent meta-analyses and decisions based on those analyses. The \( r_1 \) formula is written

\[
 r_1 = \frac{\sum_{i=1}^{n-1}(Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}
\]

This is the formula most frequently presented in time-series analysis textbooks and most often used in widely available statistical software packages for computing autocorrelation coefficients on time-series data.
Modified Autocorrelation Estimator $r_{F1}$ Applied to the Raw Data

A newly developed autocorrelation estimator $r_{F1}$ (Huitema & McKean, 1994b) was the second coefficient computed on the raw data of each series. The coefficient $r_{F1}$ was developed to avoid the problem of bias found in the conventional estimator. The $r_{F1}$ formula is written

$$r_{F1} = \left[ r_1 + \frac{1}{n} \right] \left[ 1 + \frac{5}{n} \right]$$

where $r_1$ is the conventional lag-1 autocorrelation coefficient applied to the original series (prior to detrending) and $n$ is the number of observations upon which $r_1$ is based.

Huitema and McKean (1991) demonstrated that the amount of bias in the $r_1$ estimator is a function of the true value of the autocorrelation in the process ($\rho$), and the sample size. The estimator $r_1$ is highly biased in the case that the process autocorrelation $\rho$ is a high positive value, especially when based on sample sizes that are typical of applied behavioral studies, whereas $r_{F1}$ is almost unbiased with most levels of process autocorrelation, even with small $n$ (Huitema and McKean, 1994).

Conventional Autocorrelation Estimates Applied to Linear Detrended Data

A second source of bias that artificially inflates autocorrelation estimates is present in many applied behavioral studies; trend in the data series introduces a positive bias into the already biased $r_1$. It is expected that when the data are appropriately detrended the autocorrelation distribution will be similar to that expected by chance when sampling from a population of independent (nonautocorrelated) errors. To detrend the data, the raw data were regressed on time. The conventional autocorrelation estimate computed on the detrended data is denoted $r_{1R}$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Reduced Biased Autocorrelation Estimator Applied to Linear Detrended Data

The reduced bias \( r_{F1} \) estimator was developed for use with raw scores. A corresponding unbiased estimator for use on the residuals of linear detrended data is not currently available, and little is known about the properties of the \( r_{F1} \) estimator when applied to the residuals of a simple linear regression. Further, a corresponding test of significance of \( r_{F1} \) applied to detrended data is not available. Thus, an estimator developed by Huitema and McKean (1995) for use in testing for autocorrelated errors in time-series regression models was employed. This estimator is defined as follows:

\[
 r_{HM} = \left[ r_{1R} + \frac{P}{n} \right]
\]

where \( r_{1R} \) is the \( r_1 \) estimator computed on the detrended data, \( P \) is the number of parameters in the model, and \( n \) is the number of observations upon which the coefficient is based. The value of \( P \) in the case of the simple linear regression model used for detrending is two. This estimator is essentially unbiased when sampling from error processes having autocorrelation values near zero.

\( t \)-values on Slope as a Standardized Descriptor of Trend in Sample Data Sets

When the raw data are regressed on time for each individual data set in order to remove the influence of trend from the autocorrelation analysis, a \( t \) value can be obtained to test the null hypothesis that the population slope is zero; it may also serve a second, descriptive, function. The \( t \) statistic is computed by dividing the sample slope coefficient by an estimate of its standard error as follows:

\[
 t = \frac{b_1}{\sqrt{\frac{MS_{Residual}}{\sum x^2}}} 
\]
where \( b_1 \) is the sample slope, \( MS_{\text{Residual}} \) is the mean square residual obtained from the analysis of variance of the regression of \( Y \) on time, and \( \Sigma x^2 \) is the corrected sum of squares on the time variable. The \( t \) values resulting from performing the regression procedure on each data set contained in the samples were plotted as frequency distributions. These distributions provide descriptions of the standardized slope in the data sets contained in the samples. The problem with the raw slope as a measure of trend is that they are not all based on the same metric (on either the time dimension or the dependent variable). Dividing the slope by the estimated standard error of the slope standardizes the \( b_1 \) values and hence converts them to a common metric.

Inferential Procedures

Test Statistic for the \( r_{F1} \) and \( r_{HM} \) Autocorrelation Coefficients

The test statistic associated with the reduced bias autocorrelation estimator \( (r_{F1}) \) for the raw data series was computed using the formula presented in Huitema and McKean (1994b); this statistic is interpreted as a standardized normal deviate and is defined as:

\[
Z_{F1} = \frac{r_{F1}}{\sqrt{\left(1 + \frac{5}{n}\right) \left(\frac{(n-2)^2}{n^2(n-1)}\right)}}
\]

where

\[
r_{F1} = \left[r_1 + \frac{1}{n}\right] \left[1 + \frac{5}{n}\right]
\]

Similarly, the test statistic associated with the estimator \( r_{HM} \) for the detrended data was computed using the formula which yields a standardized normal deviate:
2.5. Chi-square Tests of the Homogeneity of $r_{F1}$ and $r_{HM}$ Autocorrelation Coefficients

A chi-square statistic developed by Huitema, McKean, and Zhao (in preparation, c) was used to test the homogeneity of the autocorrelations. This statistic tests $H_0: \rho_1 = \rho_2 = \ldots = \rho_J$; that is, it is a test of the hypothesis that the process autocorrelation is the same in all $J$ studies. If the null is retained, this is evidence that the variation of the sample autocorrelation coefficients around the central value may be due to sampling error alone (i.e., the sample $r$ coefficients were drawn from a population of studies having the same degree of process autocorrelation). If the coefficients are considered homogeneous it is appropriate to pool them. If the null is rejected, then factors influencing the variation away from the central value (i.e., moderator variables) need to be found and included in the statistical model describing the data.

The chi-square test for the homogeneity of the $r_{F1}$ coefficients was computed as follows:

$$z_{HM} = \frac{r_{HM}}{\sqrt{\frac{(n-2)^2}{n^2(n-1)}}}$$

where

$$r_{HM} = \left[ r_{1R} + \frac{2}{n} \right]$$

Chi-square Tests of the Homogeneity of $r_{F1}$ and $r_{HM}$ Autocorrelation Coefficients
where \( v^2 \chi^2 \) is the test statistic, \( r_{Fij} \) is the autocorrelation estimate based on sample \( j \), \( J \) is the number of sample data sets contributing to the chi-square statistic, and \( n_j \) is the number of observations in the \( j \)th data series. This test statistic is evaluated against the critical value of the chi-square distribution based on \( J-1 \) degrees of freedom.

The mean of the \( r_{F1} \) coefficients weighted by sample size, \( \bar{r}_{F1w} \), was computed as

\[
\frac{\sum_{j=1}^{J} n_j \left( r_{F1j} \right)}{\sum_{j=1}^{J} n_j}
\]

Similarly, the chi-square test of the distribution of \( r_{HM} \) coefficients was computed as follows:

\[
\chi^2_{HM} = \sum_{j=1}^{J} \left( \frac{r_{HMj} - \bar{r}_{HMw}}{\left( \frac{n_j - 2}{n_j^2(n_j - 1)} \right)} \right)^2
\]

where \( \chi^2 \) is the test statistic, \( r_{HMj} \) is the autocorrelation estimate associated with the \( j \)th data set, \( J \) is the number of data sets in the sample contributing to the chi-square statistic, and \( n_j \) is the number of observations in the \( j \)th data series. This test statistic is evaluated against the critical value of the chi-square distribution based on \( J-1 \) degrees of freedom.

The mean of the \( r_{HM} \) coefficients weighted by sample size, \( \bar{r}_{HMw} \), was computed as

\[
\frac{\sum_{j=1}^{J} n_j \left( r_{1Rj} + \frac{2}{n_j} \right)}{\sum_{j=1}^{J} n_j}
\]
95% Confidence Interval on \( \rho \) Based on the Pooled Weighted Mean Autocorrelation Coefficient \( \tilde{r}_{HM} \)

Confidence intervals around \( \rho_1 \) based on the pooled weighted mean \( \tilde{r}_{HMw} \) were computed as follows:

\[
95\% \text{ Confidence Interval on } \rho_1: \quad \tilde{r}_{HMw} \pm \frac{1}{\sqrt{\Sigma n_j}} (1.96)
\]

where \( \tilde{r}_{HMw} \) is the pooled weighted autocorrelation for the \( j \)th group and \( \Sigma n_j \) is the sum of the number of observations contained in the individual data sets contributing to the weighted mean.

Population Variance Estimates of \( \rho_j \) and \( \epsilon_j \)

It is recognized that sampling error, expected in any coefficient based on sample data, is exacerbated in small-\( n \) cases. The model used in this dissertation to explain autocorrelation in sample data using, for example, the \( r_{F1} \) estimator is

\[
r_{F1j} = \rho_j + \epsilon_j
\]

where \( r_{F1j} \) is the sample lag-1 autocorrelation coefficient based on the \( j \)th process, \( \rho_j \) is the lag-1 autocorrelation associated with the \( j \)th process, and \( \epsilon_j \) is sampling error \( (r_{F1j} - \rho_j) \). It is assumed that the errors are normally and independently distributed around zero. Thus, the sample autocorrelation can be conceptualized as the sum of two components: the true but unknown process autocorrelation plus sampling error (Huitema, 1988). This statistical model explains why variation in sample autocorrelation coefficients can be anticipated when the autocorrelation \( (\rho_j) \) is constant across different processes.
The model does not explain the total expected variation in autocorrelation coefficients when they are sampled from processes with differing $\rho_j$ values. In such a case, more terms are needed in the statistical model because sample autocorrelation coefficients may vary due to both sampling error and differences among the process autocorrelations. This model is written:

$$r_{F1j} = \bar{\rho} + (\rho_j - \bar{\rho}) + \epsilon_j$$

where $r_{F1j}$ is the sample lag-1 autocorrelation coefficient based on process j, $\bar{\rho}$ is the mean of the process autocorrelations, and $\epsilon_j$ is sampling error $(r_{F1j} - \rho_j)$ associated with $r_{F1j}$. Thus $r_{F1j}$ is the sum of three components: the overall mean of the process autocorrelation coefficients, the difference between $\rho_j$ and $\bar{\rho}$, and sampling error (Huitema, 1988).

Huitema (1988) proposed three essential questions based on the above three-component model: (1) What is the level of the overall process autocorrelation? (2) What is the variance of the lag-1 process autocorrelations? and (3) What is the variance of the sampling errors $(r_{F1j} - \rho_j)$? He contended that the answers to these questions provide a basis for the appropriate interpretation of obtained autocorrelation distributions. The $z$ tests described in the previous section address the question of the overall level of the process autocorrelations by testing the obtained $r_{F1}$ values against the postulated $\rho = 0$. Of particular relevance to this section is the difference between the two theoretical variances that explain the variance in the observed distribution of autocorrelation coefficients. The partitioning of the variance can be modeled as

$$\sigma_{F1}^2 = \sigma_{\rho_j}^2 + \sigma_{\epsilon_j}^2$$

where $\sigma_{F1}^2$ is the variance of the population distribution of $r_{F1}$ values, $\sigma_{\rho_j}^2$ is the variance of the process autocorrelations $\rho_j$, and $\sigma_{\epsilon_j}^2$ is the variance of the errors. If all
If \( r_{F1} \) coefficients are sampled from a single process, then \( \sigma^2_{r_{F1}} \) will be equal to \( \sigma^2_\varepsilon \). Thus if we compute estimates of \( \sigma^2_{r_{F1}} \) and \( \sigma^2_\varepsilon \), the difference between these two estimates will give a reasonable estimate of the remaining variance due to differences among process autocorrelations.
RESULTS

Results regarding the characteristics of the four samples of behavioral time-series data sets are described in this section. A description of the length of the many time-series data sets is presented first. Next, results on the distributions of various measures of autocorrelation are described in detail. Last, the characteristics of the observed distributions are compared with relevant theoretical distributions.

Descriptive Statistics

Length of Time-Series Data Sets

The length of the time-series data sets contained in the four samples is illustrated in Figure 4. Each distribution in the figure illustrates the number of observations included in each time-series data set. Summary statistics based on the distributions are presented in Table 3. The 1990-1995 JABA sample (N=200) has a slightly lower mean number of observations per data set than does the sample of 101 first baselines employed by Busk and Marascuilo (1988), but the subsample mean of 150 conventional baselines is almost identical to the Busk and Marascuilo sample. The functional analysis subsample (N=50), however, is characterized by a much lower mean number of observations per data set and a narrower range. The maximum series length of the subsample of conventional baseline data sets (N=150) and the 101 Busk and Marascuilo baselines are 106 and 69 observations, respectively; the maximum length of the functional analysis subsample (N=50) is only 17 observations. It should be kept in mind that one criterion for selection of these time-series data sets was that they contain a minimum of six observations. There were 495 functional
Figure 4. Frequency Distributions for the Number of Observations in Four Samples of Applied Behavioral Data Sets (Censored for n < 6 Observations).
Table 3

Number of Data Sets, Mean and Median Number of Observations, and Range of Values in Each Sample Contained in the Study

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of data sets contained in the sample (N)</th>
<th>Mean number of observations (n) per data set</th>
<th>Median number of observations per data set</th>
<th>Range of values in each sample (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990-1995 JABA sample</td>
<td>200</td>
<td>13.69</td>
<td>8</td>
<td>6-106</td>
</tr>
<tr>
<td>Conventional Baseline Subsample</td>
<td>150</td>
<td>15.21</td>
<td>10</td>
<td>6-106</td>
</tr>
<tr>
<td>Functional Analysis Subsample</td>
<td>50</td>
<td>9.12</td>
<td>10</td>
<td>6-17</td>
</tr>
<tr>
<td>Busk &amp; Marascuilo Sample</td>
<td>101</td>
<td>15.75</td>
<td>13</td>
<td>6-69</td>
</tr>
</tbody>
</table>

analysis and 1,955 conventional baseline time-series in the volumes of *JABA* spanning 1990 (volume 23, issue 1) to 1995 (volume 28, issue 1). Forty-seven percent of the functional analysis data sets (N=235) and 40% of the conventional baselines (N=782) contained less than five observations and were excluded from the present analysis.

A summary of the means of the autocorrelation coefficients computed on the original and detrended data for the 1990-1995 *JABA* sample, its two subsamples, and the Busk and Marascuilo sample appears in Table 4. The frequency distributions associated with these samples appear in Figure 5. Further analysis of the findings follows.
Table 4
Unweighted Means on $r_1$, $r_{1R}$, $r_{F1}$, and $r_{HM}$ Computed for Four Samples

<table>
<thead>
<tr>
<th>Data Series</th>
<th>Autocorrelation Estimator</th>
<th>1990-1995 JABA Sample</th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$r_1$</td>
<td>-.020</td>
<td>-.020</td>
<td>-.020</td>
<td>.076</td>
</tr>
<tr>
<td>Original</td>
<td>$r_{F1}$</td>
<td>.119</td>
<td>.108</td>
<td>.151</td>
<td>.219</td>
</tr>
<tr>
<td>Detrended</td>
<td>$r_{1R}$</td>
<td>-.175</td>
<td>-.168</td>
<td>-.194</td>
<td>-.038</td>
</tr>
<tr>
<td>Detrended</td>
<td>$r_{HM}$</td>
<td>.029</td>
<td>.023</td>
<td>.046</td>
<td>.126</td>
</tr>
</tbody>
</table>

Conventional Autocorrelation Estimator $r_1$ Applied to Original Data

Figure 5 (panel A) shows the frequency distribution of the $r_1$ autocorrelation estimates computed on the original observations ($r_1$) from the 1990-1995 *JABA* sample. The unweighted mean (i.e., the arithmetic average) of this distribution is -.020. Panels B and C show the frequency distribution of the $r_1$ autocorrelation estimates for the conventional baseline and functional analysis subsamples of the 1990-1995 *JABA* sample. The general shapes of these distributions are consistent with the overall pattern of the larger sample. The distribution for the Busk and Marascuilo sample is shown in panel D. This distribution has a higher mean ($\bar{r}_1 = .076$), with a larger proportion of positive than negative values.

$r_{F1}$ Autocorrelation Estimator Applied to Original Data

The autocorrelation estimator $r_1$ is known to be negatively biased, thus the bias may affect any conclusions based on a meta-analysis employing this estimator. Hence, the reduced bias autocorrelation estimator $r_{F1}$ was applied to all data sets. It can be
seen in Figure 6 that the results based on applying $r_{F1}$ to the original data series differ from those based on the $r_1$ estimator. The $r_{F1}$ distribution in panel A appears to be stretched toward the upper and lower limits and is centered around a higher positive mean (.119) value than the distribution of $r_1$. When separated into the conventional baseline (N=150) and functional analysis (N=50) subsamples, the same overall pattern is evident (Figure 6, Panels B and C). These distributions are centered around .108 and .151 respectively, and both are pulled out to the upper and lower limits. Again, the

![Figure 5. Frequency Distributions for the $r_1$ Coefficients Computed on Four Samples.](image-url)
same pattern is clearly evident in the distribution of the $r_{F1}$ values from the Busk and Marascuilo sample (Figure 6, panel D).

$t$-values on Slopes as Standardized Descriptors of Trend in Original Data

Figure 7 (panel A) shows the frequency distribution of $t$ values associated with the slopes ($b_1$) from individual data sets in the 1990-1995 $JABA$ sample (N=200).
Although the distribution is centered close to zero ($\bar{t}_{b_1} = -0.108$), it can be seen that there is a large number of substantial values in the distribution. Values greater than approximately two identify statistically significant trend. Similar patterns are seen when the distributions are plotted for the conventional baseline and functional analysis subsamples and for the Busk and Marascuilo sample (panels B, C, and D of Figure 7). Slope in the data contributes an artificial positive inflation in both the $r_1$ and $r_{p1}$ values.

Figure 7. Frequency Distributions for the $t_{b1}$ Coefficients Computed on Four Samples.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 8 contains bivariate distributions of the $r_{F1}$ values plotted on the absolute values of the standardized slopes for the 1990-1995 JABA sample and the Busk and Marascuilo sample. These graphs demonstrate the effects of slope on the $r_{F1}$ coefficients. In both cases, higher absolute values for the slopes are strongly associated with higher positive values on $r_{F1}$.

Figure 8. $r_{F1}$ Values Plotted Against the Absolute Values of the Standardized Slopes for the 1990-1995 JABA Sample and the Busk and Marascuilo Sample.
The lines fit through the two distributions in Figure 8 are based on quadratic regressions of the $r_{\text{F1}}$ values on the absolute standardized slopes and the squared values of those standardized slopes. These results lead to the conclusion that autocorrelation coefficients are strongly affected by trends in the data. The results of linear detrending can be seen in the next section by comparing the autocorrelation based on the original data with those based on detrended data. The influence of nonlinear trend will be discussed subsequently.

Conventional Autocorrelation Estimates Applied to Linear Detrended Data

It can be seen in Figure 9 that when the positive bias resulting from linear trend is removed from the data, the frequency distribution of the $r_1$ values is centered lower on the scale. Panel A shows the frequency distribution of the $r_1$ autocorrelation estimates computed on the detrended observations ($r_{1R}$) contained in the individual data sets of the 1990-1995 JABA sample; the mean of this distribution is -.175. Again, this pattern is consistent for the conventional baseline and functional analysis subsamples of the 1995-1990 JABA sample (panels B and C of Figure 9, respectively). Panel D shows that the Busk and Marascuilo sample contains a larger proportion of positive than negative $r_{1R}$ coefficients in comparison to panels A through C.

The $r_{1R}$ coefficients, although devoid of the positive bias introduced by linear trend, contain the negative bias inherent in the $r_{1R}$ estimator. It is for this reason that the reduced bias estimator $r_{\text{HM}}$ was also applied to the detrended data.

$r_{\text{HM}}$ Autocorrelation Estimator Applied to Linear Detrended Data

The $r_{\text{HM}}$ autocorrelation coefficient provides a less biased estimate of the degree of autocorrelation in applied behavioral studies than does the conventional estimator $r_{1R}$. Figure 10 indicates that the $r_{\text{HM}}$ coefficients plotted as frequency distributions for

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
1990-1995 *JABA* sample, the conventional baseline and functional analysis subsamples and the Busk and Marascuilo sample appear more normal than the $r_{1R}$ distributions. Further, these distributions are centered close to zero, and have fewer coefficients in the extreme positive and negative tails. That is, when the effects of both the negative bias inherent in the $r_{1R}$ autocorrelation estimator is reduced and the positive bias introduced by trend in the data is removed, the empirical distributions more closely approximate

---

**Figure 9.** Frequency Distributions for the $r_{1R}$ Coefficients Computed on Four Samples.
normality and are centered around zero.

A summary plot of the unweighted means of the $r_1$, $r_{1R}$, $r_{F1}$, and $r_{HM}$ distributions for the 1990-1995 JABA sample, the two subsamples of that sample, and for the Busk and Marascuilo sample can be seen in Figure 11. This figure illustrates the effects of the different estimators and detrending for each sample. In order to assess whether the variation of the $r_{F1}$, and $r_{HM}$ distributions around zero is more than
that which would be expected on the basis of sampling alone, test statistics associated with the $r_{F1}$ and $r_{HM}$ distributions were computed. The results of these and other inferential tests are described in the subsequent two sections.

Inferential Procedures

Hypothesis tests of $H_0: \rho_1 = 0$ associated with the reduced biased autocorrelation estimators for the original data series and for the linear detrended data were computed. The distribution of the $z_{F1}$ test statistic associated with the $r_{F1}$ distributions indicates that there were more coefficients falling in the tails of the distributions than would be expected if the coefficients differed from zero only as a result of sampling error. However, the distributions of the $z_{HM}$ test statistic associated with the $r_{HM}$ distributions show that the number of coefficients falling in the extreme tails are consistent with that which would be predicted on the basis of sampling error.

![Figure 11. Mean Unweighted Autocorrelation Estimators $r_1$, $r_{IR}$, $r_{F1}$, and $r_{HM}$ for Four Samples.](image-url)
alone. That is, when the negative bias of the $r_1$ estimator is controlled for, and the positive bias introduced by linear trend in the data is eliminated, fewer than 5% of the coefficients are more extreme than the critical values of ±2.0. Further description of the inferential analysis based on test statistics follows.

**$z_{FI}$ and $z_{HM}$ Distributions**

Table 5 contains the means of the test statistics $z_{FI}$ and $z_{HM}$ computed for the $r_{FI}$ and $r_{HM}$ distributions for the 1990-1995 JABA sample, its two subsamples, and the Busk and Marascuilo sample. The mean of the $z_{FI}$ distribution for all samples falls between .347 and .701, while the $z_{HM}$ distributions are centered closer to zero. Inspection of the distribution of $z_{FI}$ test statistics (Figure 12) shows that all samples contain more coefficients in the tails than would be predicted on the basis of sampling error alone. The proportion of coefficients falling beyond ±2.0 ranges from 9.33% to 13.86% for the four distributions; 5% or less would be predicted if the coefficients differed from zero only on the basis of sampling error.

Inspection of the $z_{HM}$ distributions (Figure 13) confirms that the large proportion of extreme $z_{FI}$ scores (Figure 11) can be explained by the presence of linear

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>1990-1995 JABA Sample</th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{FI}$</td>
<td>.352</td>
<td>.347</td>
<td>.368</td>
<td>.701</td>
</tr>
<tr>
<td>$z_{HM}$</td>
<td>.149</td>
<td>.139</td>
<td>.176</td>
<td>.567</td>
</tr>
</tbody>
</table>
trend in the individual data sets. That is, the distributions of test statistics associated with the reduced bias coefficients computed on the detrended data ($r_{HM}$) contain fewer than 5% more extreme than ±2.0. The actual numbers of coefficients found beyond ±2.0 and their corresponding percentages for the $z_{FI}$ and $z_{HM}$ distributions can be found in Table 6.

![Histbograms](image.png)

Figure 12. Frequency Distributions for the $z_{FI}$ Coefficients Computed on Four Samples.
Chi-square Tests of the Homogeneity of $r_{F1}$ and $r_{HM}$ Autocorrelation Coefficients

The chi-square statistics used to test the homogeneity of the obtained $r_{F1}$ and $r_{HM}$ autocorrelation distributions around their central (weighted mean) values can be found in Table 7. The obtained $\chi^2$ values associated with the $r_{F1}$ distributions for the 1990-1995 JABA sample, for the conventional baseline subsample, and for the Busk and Marascuilo sample were greater than their respective critical values ($0.5 \chi^{2}_{1,4}$). These

![Graphs A, B, C, D showing frequency distributions for different samples.](image)

Figure 13. Frequency Distributions for the $z_{HM}$ Coefficients Computed on Four Samples.
Table 6

Number of Test Statistics Falling Beyond the ±2.0 Limits for $z_{F1}$ and $z_{HM}$ Distributions for the 1990-1995 *JABA* Sample, the Conventional Baseline and Functional Analysis Subsamples, and the Busk and Marascuilo Sample

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>1990-1995 <em>JABA</em> Sample</th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{F1}$</td>
<td>19 (9.5%)</td>
<td>14 (9.33%)</td>
<td>5 (10.0%)</td>
<td>14 (13.86%)</td>
</tr>
<tr>
<td>$z_{HM}$</td>
<td>6 (3.0%)</td>
<td>6 (4.0%)</td>
<td>0 (0.0%)</td>
<td>5 (4.95%)</td>
</tr>
</tbody>
</table>

results indicate that the $\rho_j$ parameters associated with the process in the behavioral studies are not homogenous. The obtained $\chi^2$ value associated with $r_{F1}$ for the functional analysis subsample was less than the critical value of $0.05 \chi^2_{.05}$, thus the data are consistent with the hypothesis of a common $\rho$ and it is appropriate to pool these coefficients for significance testing against $H_0: \rho = 0$.

The results of the tests associated with the $r_{HM}$ distributions show that when the influence of trend is removed from the autocorrelation estimates, the 1990-1995 *JABA* sample and the Busk and Marascuilo sample, as well as the functional analysis subsample are associated with nonsignificant chi-square values ($p$ values range from .06 to .58). These distributions of $r_{HM}$ values are considered homogenous; there is justification for pooling the coefficients to obtain the pooled mean $r_{HM}$ coefficient upon which to construct 95% confidence intervals on $\rho$. The $r_{HM}$ distribution for the conventional baseline subsample yields a significant chi-square statistic ($p=.03$); the distribution appears to contain heterogeneous coefficients. Hence there is limited

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
interest in pooling the coefficients because there is not a common process autocorrelation across the various studies in the sample.

Table 7

<table>
<thead>
<tr>
<th></th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2_{1,1}$ Obtained$F_1$</td>
<td>282.53* (p=.00)</td>
<td>232.77* (p=.00)</td>
<td>142.75* (p=.00)</td>
</tr>
<tr>
<td>$\chi^2_{1,1}$ Obtained$HM$</td>
<td>230.40 (p=.06)</td>
<td>184.14* (p=.03)</td>
<td>114.94 (p=.15)</td>
</tr>
</tbody>
</table>

Confidence Intervals on $\rho_1$

Table 8 contains the pooled weighted $r_{F_1}$ and $r_{HM}$ autocorrelations computed for the four samples. The coefficients associated with the original data series for each sample are considerably larger than those associated with the detrended data. Though it appears that there is little difference between the conventional baseline and functional analysis subsamples of the 1990-1995 JABA sample and that the weighted coefficient computed for the Busk and Marascuilo sample is much larger, examination of the 95% confidence intervals on $\rho_1$ provide more adequate descriptions of the process autocorrelations.

Table 9 contains the confidence intervals on the process autocorrelations for the four samples based on the pooled weighted mean $r_{HM}$. As expected, the only interval containing zero is that of the functional analysis subsample. The statistical model that explains the obtained functional analysis autocorrelation coefficients best is the simple
model \( r_{HMj} = \varepsilon_j \), where the observed \( r_{HM} \) is a function of sampling errors \( \varepsilon_j \) that have a mean of zero.

The remaining three confidence intervals (i.e., those of the 1990-1995 JABA sample, the conventional baseline subsample, and the Busk and Marascuilo sample) do not contain zero, although they are relatively narrow. The statistical model that explains the obtained autocorrelation coefficients for these samples is the three-component model

\[
\begin{align*}
    r_{HMj} &= \bar{\rho} + (\rho_j - \bar{\rho}) + \varepsilon_j,
\end{align*}
\]

where the observed \( r_{HM} \) is a function of the overall mean of the process autocorrelation coefficients \( \bar{\rho} \), the difference between \( \rho_j \) and \( \bar{\rho} \), and sampling error.

Table 8

<table>
<thead>
<tr>
<th>Autocorrelation Estimator</th>
<th>1990-1995 JABA Sample</th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_{F1} )</td>
<td>.144</td>
<td>.123</td>
<td>.156</td>
<td>.218</td>
</tr>
<tr>
<td>( r_{HM} )</td>
<td>.056</td>
<td>.057</td>
<td>.053</td>
<td>.139</td>
</tr>
</tbody>
</table>

Table 9

95% Confidence Intervals Around \( \bar{r}_{HM} \) on the Process Autocorrelations Computed for Four Samples

<table>
<thead>
<tr>
<th>1990-1995 JABA Sample</th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>(.02, .09)</td>
<td>(.02, .10)</td>
<td>(-.04, .14)</td>
<td>(.09, .19)</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The results of the confidence intervals around the mean weighted coefficients suggest that the mean process autocorrelation is close to zero after removing the effects of linear trend. A second issue, beyond the mean of the distributions, is whether the variance of the distributions is explainable as sampling error. To appropriately determine the amount of observed variance in $r_{HM}$ values that can be attributed to sampling error relative to the amount that can be attributed to variation in process autocorrelation, estimates of both contributions were computed. These results are presented next.

Estimates of Error Variance, Variance in the $r_{HM}$ Distributions, and Variance Due to Other Sources

The homogeneity test on the $r_{HM}$ coefficients reported above indicate that a reasonable statistical model for the variance in the $r_{HM}$ distribution for the functional analysis subsample is:

$$\sigma^2_{r_{HM}} = \sigma^2_e.$$

This model suggests that the variation in the obtained distribution is solely due to sampling error. A reasonable model for the $r_{HM}$ variance in the 1990-1995 *JABA* sample, the conventional baseline subsample and the Busk and Marascuilo sample includes the error term, plus an additional term to explain differences in process autocorrelations, and can be written as:

$$\sigma^2_{r_{HM}} = \sigma^2_{\rho_{ij}} + \sigma^2_e.$$

If the variance of the $r_{HM}$ values is a function of (a) error variance and (b) variance in process autocorrelations, then the amount of variance observed above that predicted on the basis of sampling error can be interpreted as the variance of the unknown $\rho_{ij}$ values. The latter variance can be estimated from the sample data and knowledge of the

---

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
theoretical variance of the sample $r_{HM}$ values. Table 10 shows the variance estimates associated with the $r_{F1}$ and the $r_{HM}$ distributions computed for the four samples. Rows 1 and 2 contain the estimates for the processes containing linear trend. Row 3 contains the differences between rows 1 and 2 ($\hat{\sigma}_{r_{F1}}^2 - \hat{\sigma}_{m}^2$); these differences estimate the amount of variation $\hat{\sigma}_{p_j}^2$ in the distributions beyond that explained by sampling error. Because the processes upon which the sample $r_{F1}$ autocorrelations are based are known to contain linear trend, it is not surprising to find that all computed values of $\hat{\sigma}_{p_j}^2$ are larger than zero, although the estimate for the functional analysis is substantially smaller than for the other samples.

Rows 4 and 5 of Table 10 contain the variance estimates for the linear detrended distributions. The difference between rows 4 and 5 ($\hat{\sigma}_{r_{HM}}^2 - \hat{\sigma}_{m}^2$) is the amount of variation remaining in the distributions above that predicted by sampling error alone ($\hat{\sigma}_{p_j}^2$). The values of $\hat{\sigma}_{p_j}^2$ computed on the three samples with corresponding significant chi-square test statistics (i.e., the 1990-1995 JABA sample, the conventional baseline subsample of that sample, and the Busk and Marascuilo sample) range from .0090 to .0137. The $\hat{\sigma}_{p_j}^2$ value for the functional analysis subsample is a small negative (-.007) which indicates that sampling error accounts for all of the variation in the $r_{HM}$ distribution for this sample. The proportion of the estimated variance of the $r_{F1}$ and $r_{HM}$ distributions that appears to be explained by variance among the process $p_j$ values can be seen in Table 11.
### Table 10

Variance Estimates $\hat{\sigma}_f^2$ and $\hat{\sigma}_e^2$ Computed for Four Sample Distributions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>$r_{FI}$</td>
<td>$\hat{\sigma}<em>{r</em>{FI}}^2$</td>
<td>.158</td>
<td>.153</td>
<td>.190</td>
<td>.127</td>
</tr>
<tr>
<td>2)</td>
<td></td>
<td>$\hat{\sigma}_e^2$</td>
<td>.115</td>
<td>.102</td>
<td>.161</td>
<td>.094</td>
</tr>
<tr>
<td>3)</td>
<td></td>
<td>$\hat{\sigma}<em>{r</em>{FI}}^2 - \hat{\sigma}_e^2$</td>
<td>.043</td>
<td>.051</td>
<td>.009</td>
<td>.033</td>
</tr>
<tr>
<td>4)</td>
<td>$r_{HM}$</td>
<td>$\hat{\sigma}<em>{r</em>{HM}}^2$</td>
<td>.062</td>
<td>.061</td>
<td>.065</td>
<td>.058</td>
</tr>
<tr>
<td>5)</td>
<td></td>
<td>$\hat{\sigma}_e^2$</td>
<td>.052</td>
<td>.048</td>
<td>.072</td>
<td>.049</td>
</tr>
<tr>
<td>6)</td>
<td></td>
<td>$\hat{\sigma}<em>{r</em>{HM}}^2 - \hat{\sigma}_e^2$</td>
<td>.010</td>
<td>.014</td>
<td>-.007</td>
<td>.009</td>
</tr>
</tbody>
</table>

### Table 11

Estimated Proportion of Total Variance in $r_{FI}$ and $r_{HM}$ Explained by Process Autocorrelation Variance ($\hat{\sigma}_{r_{FI}}^2$) and Error Variance ($\hat{\sigma}_e^2$)

<table>
<thead>
<tr>
<th>Series</th>
<th>AC Estimator</th>
<th>Proportion explained by</th>
<th>1990-1995 JABA Sample</th>
<th>Conventional Baseline Subsample</th>
<th>Functional Analysis Subsample</th>
<th>Busk and Marascuilo Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original $r_{FI}$</td>
<td>$\hat{\sigma}<em>{r</em>{FI}}^2$</td>
<td>.272</td>
<td>.333</td>
<td>.047</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\sigma}_e^2$</td>
<td>.728</td>
<td>.667</td>
<td>.953</td>
<td>.74</td>
<td></td>
</tr>
<tr>
<td>Detrended $r_{HM}$</td>
<td>$\hat{\sigma}<em>{r</em>{HM}}^2$</td>
<td>.161</td>
<td>.23</td>
<td>.00</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\sigma}_e^2$</td>
<td>.839</td>
<td>.77</td>
<td>1.0</td>
<td>.84</td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
DISCUSSION

Many empirical and expository articles on the "autocorrelation problem" with single-subject time-series data have appeared in the behavioral literature (Busk & Marascuilo, 1988; Gottman & Glass, 1978; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). These articles generally conclude that ARIMA (and other) complex time-series methods are essential for the analysis of applied behavior analytic data. They argue that the errors of such data sets are autocorrelated. However, previous attempts to characterize the distribution of autocorrelation coefficients associated with a population of applied behavior analytic data have been flawed.

Huitema (1990) presented the most important errors in earlier attempts to characterize autocorrelation in applied behavioral time-series data sets. Early studies ignored intervention effects in computing autocorrelation coefficients, thus using an inappropriate model upon which to base claims of high autocorrelation. This problem was recognized by later researchers, but those researchers failed to go beyond an ANOVA model in explaining applied behavioral data, thus ignoring the effects of trend on the magnitude of autocorrelation estimators (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen & Ary, 1987). Finally, the universal estimation problem due to the use of the negatively biased \( r_1 \) estimator was recognized in some studies of the "autocorrelation problem," however, no reduced biased estimator was available (Huitema, 1988). The major purpose of the present research was to provide an unbiased estimate of the level of autocorrelation present in the errors of the population of recently published applied behavioral studies through specification of more
reasonable statistical models and use of reduced bias estimators of lag-1 autocorrelation.

Conclusions

Efficacy of Reduced Bias Estimators and Linear Detrending

The first analysis conducted in this study involved computing conventional $r_1$ autocorrelation coefficients on the individual time-series data sets contained in each sample. These computations produced low mean values (-.02 to .08). The negative bias inherent in the conventional $r_1$ estimator is the source of what Huitema (1990) referred to as the universal estimation problem. Application of the reduced bias $r_{F1}$ estimator produced a collection of higher positive mean coefficients (.11 to .22). These results appear to be consistent with those of Busk and Marascuilo (1988). That is, when the unbiased estimator ($r_{F1}$) is used, there appears to be a significant proportion of behavioral data sets exhibiting autocorrelation, generally in the positive direction. Most of the apparent positive autocorrelation can, however, be explained by linear trend in the data. This characterizes what Huitema (1990) labeled model misspecification type II.

Application of the conventional $r_1$ estimator on linear detrended data produced negative mean autocorrelation coefficients ranging from -.04 to -.12. Computation of these conventional autocorrelation estimators on detrended data corrected for the model misspecification flaw, but the results retained the negative bias of $r_1$. For this reason, a reduced bias estimator $r_{HM}$ was employed. These results showed that when the data are appropriately detrended and a reduced bias estimator is employed, the autocorrelation distribution is similar to that expected by chance when sampling from a population of independent (nonautocorrelated) errors.
The conclusion that behavioral data are not highly autocorrelated is drawn. In most cases the explanation for "apparent" autocorrelation is not that the errors are autocorrelated. Rather, the statistical model used to explain the data has been misspecified. Artificial autocorrelation is to be expected if trend (linear or otherwise) and other deterministic components in the data have not been specified in the statistical model.

Potential Moderating Variables

A second issue, beyond the mean of the distributions was addressed in this study; that is, "What proportion of the variance of the $r_{HM}$ distributions is explainable as sampling error?" It was concluded that sampling error alone accounts for the total variation in obtained $r_{HM}$ values for the functional analysis subsample. A portion of variation beyond sampling error is left to be explained for the remaining three samples. Several sources of artificial autocorrelation were postulated that included model misspecification problems. These problems were apparent when inspecting the original data plots from a collection of sampled studies. Other potential moderator variables that may account for the variation beyond sampling error are suggested by the differing results found for the functional analysis subsample when compared to the three other samples. The samples employed in this study suggest that "baseline type" can function as a moderator variable influencing the magnitude of autocorrelation in the 1990-1995 JABA sample.

Functional analysis data sets are a more homogeneous grouping than are the conventional baseline time series. In functional analyses the analyst identifies environmental variables that may influence the level of target behavior. These variables are precisely controlled by the investigator and baseline rates of behavior are obtained under the different conditions. The behavior of interest is typically self injury, but
other topographies such as stereotypy may also be assessed under this assay. Thus, there are several dimensions upon which functional analysis and conventional baselines differ. The functional analysis assay is used to study a narrow range of problematic behavioral topographies, the metrics typically employed are measures of rates or proportions of behavior, sessions are generally short (10-15 minutes), the characteristic subject is a person with a developmental or related disability, and data are gathered under well controlled experimental conditions. The collection of conventional baselines that make up the remaining portion of the sampled JABA studies vary on every dimension. A wide range of behavioral topographies are studied using an equally wide range of metrics. Sessions vary in length from short samples (e.g., 5-10 minutes) to large chunks of time (e.g., the work day, school day, or week), there is no one characteristic subject, and data are gathered under conditions ranging from tight experimental control to community settings where control is nearly impossible. Any one of these dimensions may prove to be a moderator variable that contributes to autocorrelation of errors in applied behavioral data.

Suggestions for Future Research

It is suggested that future research would benefit from two methodological changes. First, the reduced bias $r_{HM}$ autocorrelation estimator for use with detrended data is not completely free of bias. At low values of $\rho$, $r_{HM}$ will produce essentially unbiased estimates of the process autocorrelation. When $\rho$ is a large value, $r_{HM}$ is biased toward zero. Although the bias is less than that of $r_1$, the results of employing $r_{HM}$ with a process autocorrelation of large magnitude is to pull coefficients with extreme values in toward the center of the distribution. A reduced biased estimator for use on detrended data corresponding to the $r_{F1}$ coefficient is currently under development (McKean, McKnight & Huitema, in preparation). Future research on
autocorrelation in applied behavioral data could benefit by employing an autocorrelation estimator that is less biased at extreme values of $\rho_1$.

A second direction for future research involves modeling the complete design. Only baseline phases with a minimum of six observations were sampled in this study. Applied behavioral studies typically contain several baseline, intervention, and follow-up phases, often with fewer than six data points in a single phase. Thus, the sampled phases do not wholly reflect the characteristics of applied behavioral experimental designs. Future research employing reduced bias estimators should include attempts to model pre- and post-intervention phases across a variety of single-subject design types.

**Summary Recommendations for Analysts**

This study has shown that autocorrelation of the errors of a reasonable model is not a problem for applied behavioral time-series data. Complex time-series analysis procedures such as ARIMA models are not necessary, and are generally inappropriate, for analyzing applied behavioral data. Alternative intervention models such as ANOVA and interrupted time-series regression models generally provide a reasonable fit to behavioral data. When autocorrelated errors are present a new procedure based on generalized least squares is now available (McKnight, McKeon, & Huitema, in preparation).

The sound conclusions drawn by Huitema (1988) stand as appropriate recommendations for the analysis of behavior analytic time-series data. His conclusions are repeated as follows:

1. The residuals of appropriate conventional linear models fit to a population of typical applied behavioral studies have little or no autocorrelation.
2. The major implication of little or no autocorrelation is that conventional statistical methods based on the general linear model should be considered for the statistical analysis of behavioral data.

3. The most appropriate method for a given study will depend on the data and the researcher's question. In many cases Shewhart charts, one-factor analysis of variance, or piecewise-discontinuity regression models will provide a satisfactory solution.

4. The use of time-series models should be considered if sufficient data are available to justify the use of such models.

However, it should be understood that though these models generally fit, there are assumptions other than independence that should be satisfied before adopting any general linear model. Examples of data that cannot be fit with such models, or with any conventional time-series model, are routinely encountered in the behavioral literature. For example, Figures 14, 15 and 16 present data indicating obvious departures from (a) the constant variance assumption, and (b) the normality assumption. The use of conventional models in the situations described in these figures are very likely to lead to inappropriate conclusions. The use of statistical models that are inconsistent with the data should be discouraged.
Figure 14. Example of Nonconstant Variance in a Conventional Baseline Phase (Adapted from Austin, Hatfield, Grindle, and Bailey, 1993).

Figure 15. Example of Extreme Instability in a Conventional Baseline Phase (Adapted from Charlop and Trasowech, 1991; ).
Figure 16. Example of Ceiling Effect in a Conventional Baseline Phase (Adapted from Krantz, McClannahan, and Poulson, 1994; ).
Appendix A

List of Figures From Which the Conventional Baseline Subsample (N=150) of the 1990-1995 JABA Sample Was Drawn
For ease of retrieval, the following list of baseline figures (N=150) and the articles that they were drawn from is organized according to the order in which they appear in the *Journal of Applied Behavior Analysis* beginning with Issue 1, *Volume 23* (1990). Each article listed generally comprises more than one data set, and two or more data sets may have been drawn from the same article or figure, therefore each sample is identified with sufficient information to clearly distinguish it from other data sets within the figure or article.

- Figure 2, Noreen, SIB, page 20
- Figure 2, Noreen, compliance, page 20

- Figure 1, child 7, page 95
- Figure 2, child 2, page 96

- Figure 1, Jenny, page 116

- Figure 1, student 3, page 122
- Figure 2, student 3, page 124

- Figure 2, Jen, free-choice table, page 134

- Figure 1, top panel, page 156
- Figure 1, fifth panel, page 156

- Figure 7, bottom panel, page 177


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Figure 2, middle panel, training, page 467


Figure 1, third panel, page 491


Figure 2, top panel, page 544


Figure 1, bottom panel, doll, page 568


Figure 1, third panel, page 594


Figure 1, Doris, lunch, page 601


Figure 4, pool B, children, page 684


Figure 2, page 690


Figure 1, bottom panel, page 709


Figure 1, parent, sixth panel, page 741

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 1, bottom panel, imitated, page 754
Figure 2, bottom panel, incorrect, page 755
Figure 2, bottom panel, spontaneous, page 755

Figure 1, middle panel, page 133

Figure 2, teacher C, page 160
Figure 3, classroom C, page 162

Figure 1, staff D, across-program generalization 2, page 171
Figure 4, staff P, across-program generalization 1, page 176

Figure 1, bottom panel, idiomatic, page 186

Figure 1, Sara, page 254

Figure 2, bottom panel, theme related, page 273

Figure 2, top panel, page 285

Figure 1, Quenton, peer initiations, page 312
Figure 1, Lee, peer initiations, page 312
Figure 1, Lee, teacher prompts, page 312
Figure 2, Quenton, page 313

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
  Figure 1, Adam, school/home, page 348
  Figure 3, Ian, page 351

  Figure 1, Dave, desirable, page 360

  Figure 1, water faucet on, video assessment, page 379
  Figure 1, chair in the aisle, video assessment, page 379

  Figure 3, Vicky, training sessions, page 394
  Figure 4, class 2, generalization sessions, page 396

  Figure 1, Mary, oven, page 481
  Figure 2, Sara, refrigerator, page 482
  Figure 3, Ann, refrigerator, page 484
  Figure 6, Fran, refrigerator, page 487

  Figure 2, Ava, high-high, page 496

  Figure 1, Karen, middle panel, page 606
  Figure 2, Susie, middle panel, page 607
  Figure 3, Pauline, middle panel, page 608
  Figure 3, Pauline, bottom panel, page 608
  Figure 4, Ernie, middle panel, page 609


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


- Figure 1, top panel, page 364
- Figure 1, bottom panel, page 364


- Figure 1, middle panel, page 374


- Figure 1, bottom panel, page 383


- Dave, percentage of intervals eye contact, low stimulation community, page 402
- Nancy, percentage intervals head orienting, high-stimulation community, page 402


- Figure 1, positive attention, Bob, page 428
- Figure 3, bottom panel, page 430


- Figure 2, Sam, pre-treatment, page 445


- Figure 2, middle panel, page 27


- Figure 1, middle panel, page 55
- Figure 2, top panel, page 57
- Figure 2, bottom panel, page 57

Figure 2, bottom panel, page 80


Figure 2, bottom panel, page 109


Figure 4, 7th period, student 3, page 126
Figure 4, 7th period, student 6, page 126


Figure 1, top panel, page 150


Figure 2, bottom panel, social affect, page 166


Bottom panel, page 174


Figure 2, middle panel, SIB, page 312


Figure 1, bottom panel, page 329


Figure 2, top panel, page 337


Figure 1, Tom, page 454

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 3, bottom panel, inappropriate behavior, page 477


Figure 1, bottom panel, page 538
Figure 2, bottom panel, page 539


Figure 1, middle panel, low-P request, page 626


Figure 1, bottom panel, page 644


Figure 2, middle panel, page 654
Figure 3, middle panel, page 656


Figure 1, Neal-Training Trials, pantomime response, matches, page 691
Figure 3, Heidi-Training Trials, pantomime response, non-matches, page 693
Figure 3, Heidi-Training Trials, pantomime response, matches, page 693
Figure 3, Heidi-Probe Trials, pantomime response, non-matches, page 693
Figure 4, Seth-Training Trials, vocal response, non-matches, page 695
Figure 4, Seth-Training Trials, pantomime response, non-matches, page 695


Figure 2, bottom panel, receptive test, page 10


Figure 1, middle panel, inappropriate, page 51
Figure 2, bottom panel, appropriate, page 56

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 1, bottom panel, page 66

Figure, page 84
Appendix B

List of Figures From Which the Functional Analysis (N=50) Subsample of the 1990-1995 JABA Sample Was Drawn

87
For ease of retrieval, the following list of functional analysis baseline figures (N=50) and the articles that they were drawn from is organized according to the order in which they appear in the Journal of Applied Behavior Analysis beginning with Issue 1, Volume 23 (1990). Each article listed generally comprises more than one data set, and two or more data sets may have been drawn from the same article or figure, therefore each sample is identified with sufficient information to clearly distinguish it from other data sets within the figure or article.

Figure 1, Noreen, demand, page 16  
Figure 1, Gina, demand, page 16  
Figure 1, Gina, attention, page 16  
Figure 1, Nellie, play, page 16  
Figure 1, Donald, play, page 16  
Figure 1, Donald, medical, page 16  
Figure 1, Lana, play, page 16

Figure 1, Kathy, demand, page 366  
Figure 1, Kathy, attention, page 366

Figure 4, Ann, social attention, page 463

Figure 1, top panel, page 500

Figure 2, top panel, social disapproval, page 511

Figure 1, high % bites graph, bites, page 147  
Figure 1, low activity graph, activity prompts, page 147  
Figure 1, low % bites graph, food prompts, page 147

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 1, middle panel, task disengagement, page 556


Figure 1, Lonnie, SIB alone, page 438
Figure 2, self restraint, alone, page 440
Figure 3, restraint unavailable, SIB play, page 441
Figure 4, SIB, play, page 443
Figure 4, alone, self-restraint, page 443
Figure 4, play, self-restraint, page 443


Figure 1, Brenda, alone, page 15
Figure 1, Bonnie, demand, page 15


Figure 1, Marc, attention, page 187
Figure 4, bottom panel, page 193


Top panel, parent attention, page 256


Figure 1, Holly, no antecedent interaction (therapist present), page 302
Figure 1, Larry, antecedent instruction, page 302


Figure 1, top panel, alone, page 357
Figure 1, top panel, play, page 357
Figure 2, bottom panel, SIB (extinction), page 358


Figure 2, bottom panel, page 137
Figure 3, demand, page 138

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 2, child 2, alone, page 205
Figure 2, child 5, social disapproval, page 205


Figure 1, top panel, diverted attention/ignore, page 273
Figure 1, middle panel, social attention, page 273
Figure 1, bottom panel, diverted attention/ignore, page 273


Figure 1, open triangles, page 303


Figure 1, Wanda, demand, page 321
Figure 1, Laurie, demand, page 321
Figure 1, Lynn, demand, page 321


Figure 1, top panel, alone, page 361
Figure 1, top panel, play, page 361


Figure 1, bottom panel, alone, page 653


Figure 1, top panel, tangible, page 710
Figure 1, bottom panel escape, page 710


Figure 1, top panel, escape, page 21
BIBLIOGRAPHY


* Asterisk marks references for articles contained the 1990-1995 JABA sample and subsamples used in the meta-analysis

91

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
