DOCUMENT RESUME

ED 228 071 SE 041 268

AUTHOR Mayer, Victor J.; Monk, John S.
TITLE Handbook for Using the Intensive Time-Series Design.
INSTITUTION Ohio State Univ., Columbus.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE [83]
GRANT SED-8016589
NOTE 65p.
PUB TYPE Guides - General (050)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; Computer Oriented Programs; *Data Analysis; *Data Collection; Earth Science; Educational Research; *Elementary Secondary Education; Higher Education; *Item Analysis; Item Banks; Research Design; *Research Methodology; Science Education; Student Attitudes; Test Items
IDENTIFIERS National Science Foundation; *Science Education Research; Time Series Analysis; *Time Series Design

ABSTRACT

Work on the development of the intensive time-series design was initiated because of the dissatisfaction with existing research designs. This dissatisfaction resulted from the paucity of data obtained from designs such as the pre-post and randomized posttest-only designs. All have the common characteristic of yielding data from only one or two points in time. In addition, true experimental designs require random selection of subjects from a population and random assignment of subjects between control and experimental groups. Since such conditions seldom materialize in school situations, a design in which the same group could perform functions of both experimental and control groups would be advantageous for school-based research. The time-series design, with its potential adaptation of baseline, intervention, and followup states, has such potential. Areas addressed in this handbook include background/rationale for the time-series design, instrument development (developing item pools, generating daily instruments, constructing multiple item instruments), administration and collection of data, and analysis procedures. Example items, sample coding sheet, and Rasch item calibration heuristic are provided in appendices. (JN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED228071

HANDBOOK FOR USING THE INTENSIVE TIME-SERIES DESIGN

by
Victor J. Mayer
and
John S. Monk
The Ohio State University, Columbus, Ohio

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Victor J. Mayer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The preparation of the information contained herein, and the study from which examples were drawn, were funded in part by a grant from the National Science Foundation (Grant # SED 8016589)

Printed in USA

SE041268

CONTENTS

BACKGROUND AND RATIONALE

Work on the development of the intensive time-series design was initiated because of dissatisfaction with existing research designs. This dissatisfaction resulted from the paucity of data obtained from designs such as the prepost and randomized posttest only designs. All have the common characteristic of yielding data from only one or two points in time. Even when delayed posttests are given only one or two additional points are obtained. Typically, intervals between these data points are several weeks to months apart. What happens to learning in between these widely spaced points in time? If it were possible to obtain data on learning every day, might not the density of such data reveal new insights into the learning process and into the relationship of various instructional and environmental variables to learning? So the reasoning went. Such a design would best be used, initially at least, in descriptive studies of classroom instruction. New insights obtained would help to identify different assemblages of variables relating to classroom instruction that had not previously been identified as important in learning.

Other difficulties are inherent in the use of traditional designs, most of which have been adapted from the agricultural and physical sciences. True experimental designs require random selection of subjects from a population and random assignment of subjects between control and experimental groups. Such conditions can seldom, in reality, probably never, be met in

typical school situations. Therefore, a design in which the same group could perform the functions of both experimental and control groups would be advantageous for use in school-based research. The time-series design, with its potential adaptation of baseline, intervention, and followup stages, has such potential. Also, with additions of one or more groups, it can be adapted to an experimental design, providing a richness of data not possible with the traditional experimental designs.

Description of Intensive Time-Series Designs

Time-series designs for use in behavioral research are repeated measure designs having a large number of data points equally spaced in time. Thus far, the following designs as described in Glass, et. al., 1975 have been used:

Operant

0 I0 0 I0 0 I0 0 I0 0 I0 0 I0 ...

Single Intervention

0 0 0 0 ... I0 I0 I0 I0 ... 0 0 0 0 ...

Single or Multiple Intervention Multiple Group

0 0 0 0 ... I01 I01 I01 I01 ... 0 0 0 0 ...

0 0 0 0 ... I02 I02 I02 I02 ... 0 0 0 0 ...

Where:

0=Observation

I=Intervention.

In the latter two designs the first stage results in the collection of data on a daily basis. This is termed the BASELINE and provides information on student performance on whatever constitutes the variable of interest prior to the time that it is expected to change. In a sense, the baseline is a sophisticated pretest. In the second stage of the design, called INTERVENTION, data continue to be collected while the independent variable or treatment is introduced. In our studies this has been an instructional unit. The third stage or FOLLOWUP continues the collection of data after the termination of treatment and allows the determination of a forgetting rate.

In the single group designs that have been used so far, the group has consisted of a class or all classes of a certain subject taught by a given teacher. In the multiple group designs, the groups are either individual classes of a teacher or all of a teacher's students, regardless of class, subdivided into groups on some basis for the purpose of the study. For example, in studies on the effect of frequency of testing, a teacher's students are randomly divided into three subgroups each receiving tests on a different schedule.

Thus far, the data collected have consisted of knowledge of the topic of the unit and attitudes toward several different concepts. In each study, data have been obtained through the use of short objective items for knowledge and semantic differential items for attitude. This has permitted data collection to occupy no more than five minutes of class time. The data are collected daily, normally at the end of each class period.

Potential Unique Contributions of the Design

Because of the density of data collection unique to this design, it has the potential for yielding information on a number of questions which should be of interest to teachers. Is there a point during the teaching of a concept or unit at which the class reaches a plateau in its learning? If so, when does this occur? Is it toward the end of a unit, or somewhere prior to the end? Is there a "momentum effect" in learning? That is, is there an increment to understanding of concepts after the unit has been completed? There is a strong suggestion of such an effect in data from the Mayer and Kozlow (1980) and Farnsworth (1981) studies. How rapidly does knowledge deteriorate following instruction? Does this rate vary with level of knowledge, i.e. recall vs. understanding? Does level of cognitive development have any effect on the slope of the learning curves and on the rate of forgetting?

The design can yield valid information on the effects of environmental variables such as the day of week; for example, is there a TGIF effect? What are the effects of teacher attitudes upon student attitudes? Student performance? What effects might temperature, humidity, and barometric pressure have upon student performance and student attitudes? The design is uniquely capable of providing information on these questions and many others.

Resolution of Concerns Regarding Design Validity

A number of threats to the validity of data from a time-series design are discussed in Campbell and Stanley (1966).

These and other concerns with its use must be addressed or at least recognized in the use of the design. One of the most serious threats to internal validity is that of history or maturation. In the traditional time-series or repeated measure design data are collected at relatively large intervals of time. Also, there are relatively few data points. In the intensive design, however, data are collected daily and anywhere from 25 to 70 or more data points are obtained. Under such circumstances the threat posed by history is minimized to the point that it is of no concern. If some event occurs outside the study that influences the data in the study, that influence can be immediately identified and accounted for. The probability of more than one such external influence occurring in one day, (the time interval between data points) is so small that it can be virtually ignored. Also, because of the density of data collection, maturation can be taken into consideration when reporting results.

Experimental mortality, absenteeism in intensive time-series designs, can be a threat to internal validity. A variety of measures have been taken in obtaining and processing data to reduce this threat. Surprisingly, absenteeism has not been a serious problem in the pilot studies except where data were being collected in some of the high sickness months of mid-winter. Analyses reported by Farnsworth (1981) indicate a lack of correlation between absenteeism and achievement suggesting that measures used by her to minimize its effects were effective. Where a student was absent on a given day, his/her scores for the

preceeding and following days were averaged and used as that student's score on the missing day.

Another, and potentially more serious threat to internal validity is that of testing. This could be exhibited in two ways. One we have been referring to as "resentful demoralization". With such frequent testing, will students become resentful and, therefore, not respond accurately to items? The second possible effect of testing would be increased familiarization with the items and with the content simply due to testing. These were the most serious concerns with the design and, therefore, those which were dealt with early in its development in the study by Mayer and Rojas (1982). The students were randomly subdivided into three groups and each group tested on a different schedule: every day, every fourth day, and every eighth day. Analysis of variance of data obtained with a multiple item test administered at the end of intervention revealed no significant differences between these three groups. Also, there were no differences in trends of the data when subjected to linear regression analysis. It appeared, therefore, that frequency of testing did not affect performance on the individual items nor on overall achievement in the unit. Student attitudes were also monitored and showed no effects of "resentful demoralization". There was no detectable negative trend in attitudes about the science class. This study has now been replicated in a more rigorous design using two teachers. Preliminary data analyses support the conclusion made in the Mayer and Rojas study.

A threat to external validity is the inability to generalize since subjects are not randomly chosen and because data are collapsed into a single measure for each group. This is a problem not unique to the intensive time-series designs. In fact most, if not all, classroom research designs suffer from the same threat to external validity. The advantage of the intensive time-series design is that a great deal of information is obtained about each group. This large quantity of information allows generalizations to be made which then can be tested in subsequent studies. If replicated in a second and third study, a great deal of confidence can be placed in the generalization. This is not true in designs where only one or two data points are obtained. The reason for this will become clearer later in this handbook when data from one of our pilot studies are discussed.

Early in the development of the design, it was felt that the best demonstration of validity of data collected through the design would be if the data were found to be consistent with that generated by traditional methods. In the Mayer and Lewis (1979) study, the positive effects of field trips and the negative effects of examinations on attitudes were demonstrated repeatedly in an operant intensive time-series design. In the Mayer and Kozlow (1980) study the attempt was successfully made to replicate a learning curve during instruction on a topic. Each of the subsequent studies has measured a learning curve occurring during instruction. These results confirm the validity of the design for collecting achievement data. Another aspect of valid data is whether they discriminate between two groups that should

indeed be different on the criteria being measured. Farnsworth (1981) was able to demonstrate this when studying the differential effects of a unit on plate tectonics on the achievement of children with formal cognitive tendencies and those with concrete cognitive tendencies. Her results demonstrated the precision of the design. They have now been replicated with two other teachers in different schools.

There were a number of problems to be overcome in the development of the design. Of these, the time necessary for testing was a major concern. To be practical, only a few minutes of class time could be taken. This meant that to measure achievement objective items such as multiple choice items had to be used. In the Mayer and Kozlow study, results from a three-item instrument were compared with those from a one-item instrument. The three-item format used the same three items for all students on a given day. This was done in an attempt to provide individual student data as well as class data. The one-item format seemed to provide more reliable results. Using this format each student received a different item. Therefore, in a class of 30, 30 items would be used from an item pool each day. The conclusions were that this latter method was most useful. The one-item format is now used for both achievement and attitude testing. Attitude items are in the semantic differential format. Such instruments can be responded to rapidly by students.

What is the theoretical basis for using the single-item-per-subject data gathering technique? Lord (1962) pioneered the

use of matrix sampling in which a subset of items from an item pool is given to a subset of the tested population. Each population subset gets a different item subset so that the entire population and the entire item pool is used. In a sense, the one-item one-student technique we use is matrix sampling taken to its ultimate. However, in Lord's studies the technique begins to break down with less than five items and respondents. Therefore, matrix sampling cannot provide the theoretical basis for our technique. The recent incorporation of Rasch (1960) item calibration procedures within the framework of intensive time-series design has provided the link with evaluation theory which could not be provided by matrix sampling theory. Essentially, the Rasch procedures allow the researcher to state that each daily measurement of group performance was made utilizing the same metric or measure. The Rasch procedures adjust daily group scores for item difficulty, sample size, and variance in item difficulty. The Rasch methods, therefore, if used to calibrate our data, provides us with theoretical justification of the one-item data gathering technique as an accurate metric.

Although there does seem to be a sound theoretical basis for using this technique, our major thrust, however, has been to provide an empirical basis for justifying the measurement procedures. If it works, then it must be an appropriate technique. That is why our studies have attempted to replicate the results obtained by traditional and theoretically sound designs.

INSTRUMENT DEVELOPMENT

Development of Item Pools

Two types of items have been used in constructing daily instruments in studies to date. Achievement has been assessed through the use of multiple choice items and attitudes through the use of semantic differential items. In each case a pool of items has been developed, validated, and tested prior to their use in the daily instruments. Examples of items from a recent study can be found in Appendix A.

Achievement items were generated in reference to the objectives and specific content of the instructional unit. In the pilot studies this has been a unit on plate tectonics. Items were developed that assess learning at two levels, knowledge and understanding, by three or four individuals familiar with the unit. The items were then edited and culled for duplication. An effort was also made to assure that each of the unit objectives was proportionally represented in the item pool at each of the two levels of learning. The resulting item bank and a set of unit objectives were then submitted to a group of individuals familiar with the subject of plate tectonics. They were asked to match items to objectives to determine content validity. These individuals were also asked to comment on the clarity and accuracy of each item. Items were further refined on the basis of this information and submitted to several science educators to categorize them as assessing knowledge or understanding as defined by Bloom's taxonomy. The resulting item pool was then

assembled as a test and administered to a population of students who had completed instruction in the unit on plate tectonics. Item analysis was performed and poor items either revised or eliminated. A final check was made to assure proportional representation of levels of learning and unit objectives among the items. The remaining items, then, form the final item pool from which items are randomly selected for daily instruments.

Thus far, student attitudes toward five different concepts have been identified as potentially affected by use of the design and/or the instructional unit. They are: teacher, science, science class, plate tectonics, and this short test. Adjective pairs have been selected from a variety of sources for each of the concepts with an attempt to represent the three dimensions of potency, evaluation, and understanding. A scale for each concept has been constructed. The scales were validated by a group of science education faculty and graduate students. Weightings were assigned. The items were then tested with a population of junior high school students. Factor analyses were conducted on the resulting data and items examined for grouping into appropriate dimensions. Each scale was refined based on this data. The remaining items of each scale then became subsets of an attitude item pool from which items are drawn to construct daily attitude instruments.

The result of the item development process is an achievement item pool of 75 plate tectonics items, evenly distributed between knowledge and understanding levels, and five subgroups of attitude concepts, each with about 15 items, representing the

adjective pairs assigned to each of the five concepts. In the most recent study, reported in part as an example in this paper, three of the subgroups were used, science class, plate tectonics, and these short tests.

Generating Daily Instruments

The administration of the daily testing regime utilized in intensive time-series designs is a formidable task. For each subject, a set of items must be drawn for each day of the study. The assignment of items should meet the following criteria:

1) Each subject receives one item from each item pool on each day of the study.
2) Assignment of such items is random.
3) Random assignment is made with the following constraints:
   a) Each group in a study receives the same set of items.
   b) No subject is to receive an item for a second time until the entire item pool is exhausted.
   c) Within any given group, no two subjects are to receive the same item on the same day.

Through the use of coding schemes to identify items and computer programs which generate random number lists such as PROC PLAN in the SAS statistical package, it is fairly easy to construct an item assignment plan. Once such a plan is generated, several approaches can be used to prepare the daily instruments for individual subjects.

The most time consuming procedure has involved the duplication of each item and the hand collation of these items into daily instruments. A relatively small study involving the

measurement of the responses of 100 students on two variables lasting for 40 days would involve the hand sorting and collating of 8000 sheets of paper. In the most ambitious intensive time-series study to date (the one from which examples were drawn for this manual), approximately 500 subjects were utilized, the study lasted for a period of approximately 70 days, and each subject responded to two items on each day. This study entailed the hand sorting and collating of over 70,000 individual sheets of paper. This process, needless to say, took literally hundreds of man-hours. It also caused the researchers involved to examine other methods of generating and administering the daily instruments.

The daily instruments consist of one multiple choice achievement item and one semantic differential item. Each student within a group will have different items on successive days and no two students within a group will have the same items on a given day. In the most recent study a folder was prepared for each student in a class. On the inside front cover of the folder a mark-sense answer sheet was taped. The daily instrument was inserted in the folder before class by the teacher. At the end of each class the folders were handed out and students responded to their instruments by recording their answers on the answer sheet.

Two alternative methods are being developed to reduce this staggering workload. A FORTRAN computer program has been developed and refined which inputs the items in an item pool, the number of days in a study, the number of groups in the study, and

the number of subjects in each group. From this input the computer generates daily items, in packet form, for each subject. No handling of individual items is necessary and no human error can enter into the assignment of items. The major drawback of this program is its inability to generate graphics. Diagrams cannot automatically be included with the question. If diagrams are necessary, they must be provided by some alternative method such as the preparation of wall charts or diagram folders which could be referred to by the subjects.

A second method now being investigated is the use of a microcomputer based testing system. This system would randomly select items, present those items to subjects, solicit subject responses, and record those responses. To conduct the research without causing a great deal of interruption in the class flow, a minimum of one computer station for every three or four subjects in a class would be necessary. In addition, storing subject responses, and the items and the programming necessary to conduct the testing requires a large amount of common high speed storage. For this reason, the system envisioned would more than likely take the form of either a multi-user system with one main computer or a network system of microcomputers utilizing one common storage device. The use of any such system is appealing in that it not only eliminates the use and manipulation of papers, but also in that such a system automatically records all subject responses in a manner that allows such data to be examined in an ongoing fashion.

Constructing Multiple Item Instruments

The multiple item instruments use the entire pool of items. Since there are a large number of achievement items used in these studies (75 in the most recent one) it was necessary to construct two forms so that students would not experience undue fatigue in responding to the items. In the sample study items were randomly selected to appear on both forms. The remaining items were randomly assigned to one of the forms giving two of 42 and 43 items each. The common items are used to check the equivalence of the two groups of students taking each of the forms.

An attitude multiple item instrument was also constructed by assembling all adjective pairs under their respective concepts. This yielded an instrument of three concepts each having 15 adjective pairs. It was given on the same day to all students of a test population.

ADMINISTRATION AND DATA COLLECTION

Collecting Daily Data

The collection of student responses to daily items and the subsequent coding of those responses can also be a difficult task. In all but the most recent intensive time-series studies, students responded to items directly on the item sheets. These responses then were scored and transferred to some suitable format for computer entry. While this procedure is satisfactory

for studies of limited scope, in large studies such coding procedures are prohibitive.

In the study used for illustration here, coding was facilitated by the use of mark sense answer sheets. Students responded to each item by coding on the answer sheet an item identification code given on the question sheet and their response to that item. To reduce the possibility of errors, an overprint was used to clearly indicate where each code was to be placed. The use of the mark sense sheets did allow for the direct entry of data into the computer and computer scoring of all student responses. Unfortunately, errors were made by subjects at times both in coding the item identification code and in the placement of their responses. Although these errors were generally easily identified, their correction proved to be a long and tedious process. An example of the mark sense sheet used can be found in Appendix A.

Both alternative techniques described for generating daily items also will reduce the work necessary to code subject responses. The FORTRAN program mentioned automatically assigns subject and item codes for those items responded to by each subject for each day of the study. In addition, all items on the daily instruments are numbered consecutively from day one of the study. When using the mark sense answer sheet the subject need only place the answer to a given item by the item number on the sheet, no overprinting is necessary and the chance for error is greatly reduced. To allow for the subsequent scoring of the items, a file is generated by the program which contains the

codes for the assigned items for each subject for each day of the study. A scoring program would use this file and an item key to score each subject's responses.

The most promising technique for collecting data is the use of the microcomputer based system described earlier. Such a system would allow the collection of data while reducing the possibility of human error. In addition, data would be available for immediate examination. This would allow researchers to conduct ongoing analyses and would also allow teachers immediate feedback as to the effectiveness of a day's teaching.

Multiple Item Instruments

The administration of the multiple item instruments is far simpler than the administration of the daily instruments. The only question which needs to be answered, in this case, is on what day are such instruments to be administered?

Generally, if the study incorporates an intervention, the achievement multiple item instruments should be administered on the final day of the intervention period. This allows for the collection of data immediately following the time when all students have been exposed to all information tested by the item pool. The attitude instrument would best be given during the follow-up.

The administration of the multiple item instruments is conducted in the same manner as would be employed with any such test. It is important, however, to allow enough time on the day such instruments are used so that subjects may also respond to

the intensive time-series items for that day. The availablity of both time-series data and multiple item instrument data for at least one day of the study is essential in the validation procedures utilized with the design.

ANALYSIS

The analysis of data collected during an intensive time-series study consists of two phases. The first phase is that of analyzing the data obtained through the administration of multiple item instruments generated from the item pools utilized in the study. Such analysis is performed to verify the reliability and validity of the instrumentation. Also it provides data which allow for the daily calibration of data collected during the course of the study. The second phase is the analysis of the daily data collected during the study. Hypothesis testing is conducted on the basis of these analyses. A flowchart of the basic analysis procedures is presented in Table 1.

Whole Instrument Analysis

Before starting research using an intensive time-series design, the item pools undergo extensive examination as to their reliability and validity. Item pools so deveoped are used not only as the source of the single items to which subjects respond on a daily basis, but also as the source for the multiple item

Table 1
Flow Chart of Procedures

Phase I

Multiple Item Instrument Analysis and Calibration

1. Item Analysis and Calibration
2. Factor Analysis
3. Elimination of Items Shown To Fit Poorly
4. Item Analysis and Recalibration
5. Determination of Instrument Reliability
6. Joint Calibration of Multiple Item Instruments If Multiple Forms are Used
7. Analysis Procedures to Test Hypotheses and Establish Concurrent Validity

Phase II

Daily Data Analysis

1. Generation of Daily Group Scores based on item calibration
2. Graphing Daily Scores
3. Regressions of Variables on Day For Each Group and Stage
4. Comparisons of Regressions By Group and Stage
6. Correlational Analysis of All Variables
7. Multiple Regressions With Autocorrelation for Each Group and Stage, Using Day as a Trend Variable
8. Graph Multiple Regression With Autocorrelation Results
9. Incorporate Modified Trend and Dummy Variables in Multiple Regression With Autocorrelation
10. Graph and Examine Final Model Results

instruments. These instruments, generally given in the later part of an intensive time-series study, are used to: 1) verify item quality, 2) calibrate items, and 3) examine the validity of item pools. In addition, the data from multiple item instruments may be used in testing hypotheses dealing with general group differences.

The first step is to conduct an item analysis. In studies to date both traditional methods of item analysis and Rasch (1960) item calibration techniques have been used. Item analysis generally can be seen as providing the following information:

1) Measures of item discrimination, item fit, and point biserial item correlation. These can be used to verify the quality of individual items.
2) Measures of item difficulty. This can be used to check on equivalency of items assigned from day to day during the testing period.
3) Chi square goodness of fit tests. This allows the equality of several alternate instrument forms to be assessed.

Based on the measures of item quality and difficulty, the researcher may drop poor items from the item pool and recalibrate the remaining items to identify the best possible pool of items for subsequent analysis. Of the procedures to date, the Rasch method seems to provide item analysis information most amenable for use in subsequent analysis procedures.

Factor analysis of the item pools should also be conducted. If the item pool has been shown to display an underlying factor

structure prior to its use in the intensive time-series design, this structure may be verified. Items can, at the researcher's discretion, be added, deleted, or weighted on the basis of factor structure to insure the best possible measurement of group characteristics.

When low quality items have been identified and removed from the item pool and factor structure determined, the next step is to determine the reliability of the multiple item instruments. The reliability measures derived at this stage are assumed to be estimates of the item pool reliabilities and as such should be in agreement with reliability estimates obtained when the item pools were developed and tested.

Subsequent analysis of the data obtained from the multiple item instruments provides information as to the nature of group differences. This information not only aids in hypothesis testing, but also can be utilized to support the validity of the intensive time-series design. The presence of group differences on multiple item instruments which parallel those obtained by more traditional designs provides a form of concurrent validity. In addition, through the use of procedures such as the t-test it is possible to compare daily group scores obtained on the day on which the multiple item instrument was given with the results of the multiple item instrument itself. If it is found that the daily group score on a given criterion is not significantly different from the mean of the group on a multiple item instrument, the validity of the daily measure is supported.

Multiple item instruments have been utilized in all of the

intensive time-series design studies completed so far. In the Mayer and Kozlow (1980) study, multiple item instruments were used to establish the reliability of the item pool and as a means of providing information that allowed the researchers to examine the effect of item difficulty on daily scores. Mayer and Rojas (1982) expanded the use of such instrumentation to include the testing of hypotheses dealing with group differences. Farnsworth (1981) not only incorporated hypothesis testing based on the multiple item instruments, but also used item difficulties based on such instruments as a means of standardizing daily group scores.

In the study being referred to in this paper, the results from the item analyses of the multiple item instruments were used in generating daily scores from the single-item-per-subject responses. In addition, they provided estimates of reliability for the item pool and allowed examination of the validity of the daily measures. The reliability of the instruments used in this study are presented in Table 2.

One of the most difficult problems encountered in the use of intensive time-series designs is establishing the validity of the daily measures of group performance. It is assumed that a measure generated on any given day of a study by pooling individual subject responses to single items accurately reflects total group performance on that day. One of the key elements that permits such an assumption is that daily measures of group performance from individual raw scores can be based on a similar metric. That is to say, on each day of the study, the same ruler

Table 2

Reliabilities of Multiple Item Instruments

Instrument	n[1]	Reliability[2]
Plate Tectonics Achievement Form A	247	.87
Plate Tectonics Achievement Form B[3]	239	.87
Attitude Toward Today's Science Class	361	.92
Attitude Toward Plate Tectonics	123	.88

NOTES:

(1) n on which reliability is based is derived from the total study - not solely the subjects used in this discussion.

(2) Cronbach's alpha as computed by the Hoyt analysis of variance procedure.

(3) Reliability as determined after removal of two items of poor fit as determined by Rasch Method.

can be used to gauge group performance. It is for this reason the Rasch (1960) item calibration procedures are recommended in the analysis of multiple item instrument data. A detailed examination of the heuristic arguments for the use of the Rasch procedures is included in Appendix B.

A study by Monk (1983) has indicated that more traditional item analysis techniques can be used in the item calibration process with no significant differences in the results when they are compared with results from data calibrated with the Rasch techniques. The researcher can confidently use the normal item analysis techniques in the "every day" processing of intensive time series data so long as non-random trends are not present in the difficulties of daily item sets. The effects of such trends have not been fully investigated. Therefore, if such trends are discovered, the use of Rasch procedures would be advised.

Analysis of Daily Data

By employing appropriate calibration techniques, daily group measures should be generated for each group on each variable being assessed. These calibrated scores can then be subjected to a series of analyses, the final goal of which is to generate a model or profile of daily fluctuations in group performance. Depending on the study, these analyses can proceed in several directions. For illustration, the analysis procedures employed in a current study will be examined. This study has the following characteristics:

1) Students are blocked into two groups based on a measure of cognitive (Piagetian), level forming a

group with formal tendencies and a group with concrete tendencies.

2) The primary dependent variable is achievement on items designed to measure attainment of plate tectonics concepts. The item pool for this purpose consisted of 75 items.

3) Two other dependent variables measured are:

   a) Attitude Toward Plate Tectonics, i.e. attitude toward the content being taught, and

   b) Attitude Toward Today's Science Class, i.e. attitude toward the daily classroom situation.

4) The study consisted of three stages:

   a) Baseline, an 18 day period of time prior to the introduction of plate tectonics unit,

   b) Intervention, the 25 days during which plate tectonics was being taught, and

   c) Followup, a period of 12 days following the end of the plate tectonics unit.

The temporal sequence for the study can be represented in modified Campbell and Stanley (1966) notation as follows:

$O_1\ O_2\ \ldots\ O_{18}\quad IO_{19}\ IO_{20}\ \ldots\ XIO_{43}\quad O_{44}\ O_{45}\ \ldots\ YO_{55}$

Baseline — Intervention — Followup

where:

O = observation

IO = observation during intervention

XIO = observation and multiple item achievement testing

YO = observation and multiple item attitude testing

This study was selected as an example because it exhibited all levels of data analysis that have been utilized in intensive time-series design. There is a blocking variable (cognitive

tendency), a primary dependent variable (achievement), two explanatory variables (the two attitude measures), and the presence of an intervention (the teaching of plate tectonics to the classes).

As this study was descriptive in nature, the following working hypotheses were used in examining the data:

1) Within cognitive groups, achievement differences will be observed in the slopes and levels of the regression lines generated in baseline, intervention, and followup periods.

2) There will be differences observed between the levels of and slopes of regression lines generated for the two cognitive level groups for the baseline, intervention, and followup periods.

3) Attitudes will account for a significant amount of the daily variance in achievement, especially attitude toward plate tectonics.

To begin the analysis procedure for the daily data, the first step is the generation of a series of graphs for each variable for each group. For ease of visual interpretation, data for each cognitive tendency group on each variable were placed on a single graph. Graphs of the calibrated data derived for each day of the study are provided in Figure 1 for the formal cognitive tendency group and in Figure 2 for the concrete cognitive tendency group.

Initial examination of these graphs allows the researcher to identify relationships that appear consistent with the working hypotheses. To gain a better feel for the daily data, however, ordinary least squares regressions are also done. In the case of the study in question, the most appropriate procedures were to regress each variable by the day of the study for each group and

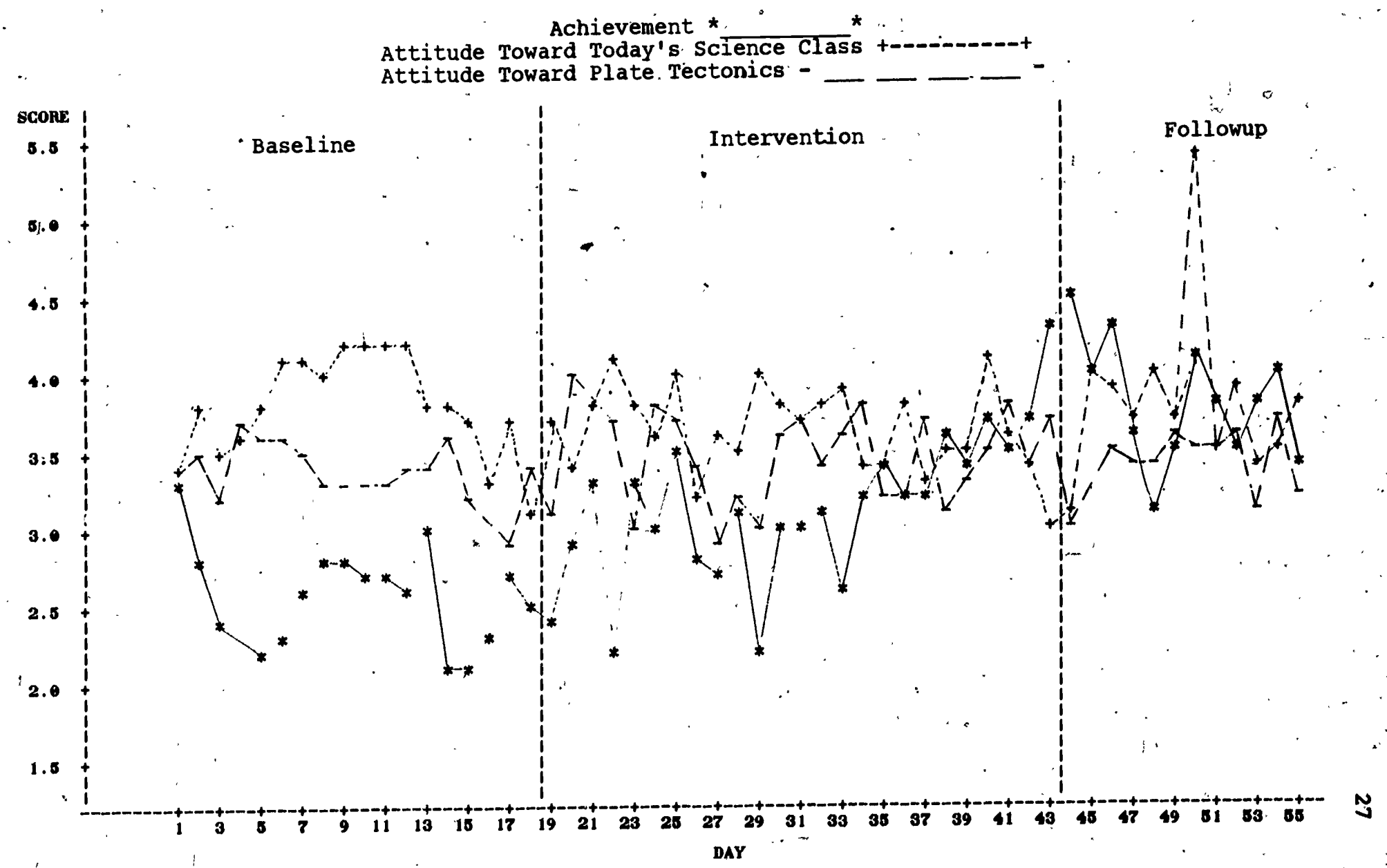


Figure 1
Plot of Formal Group Daily Measures

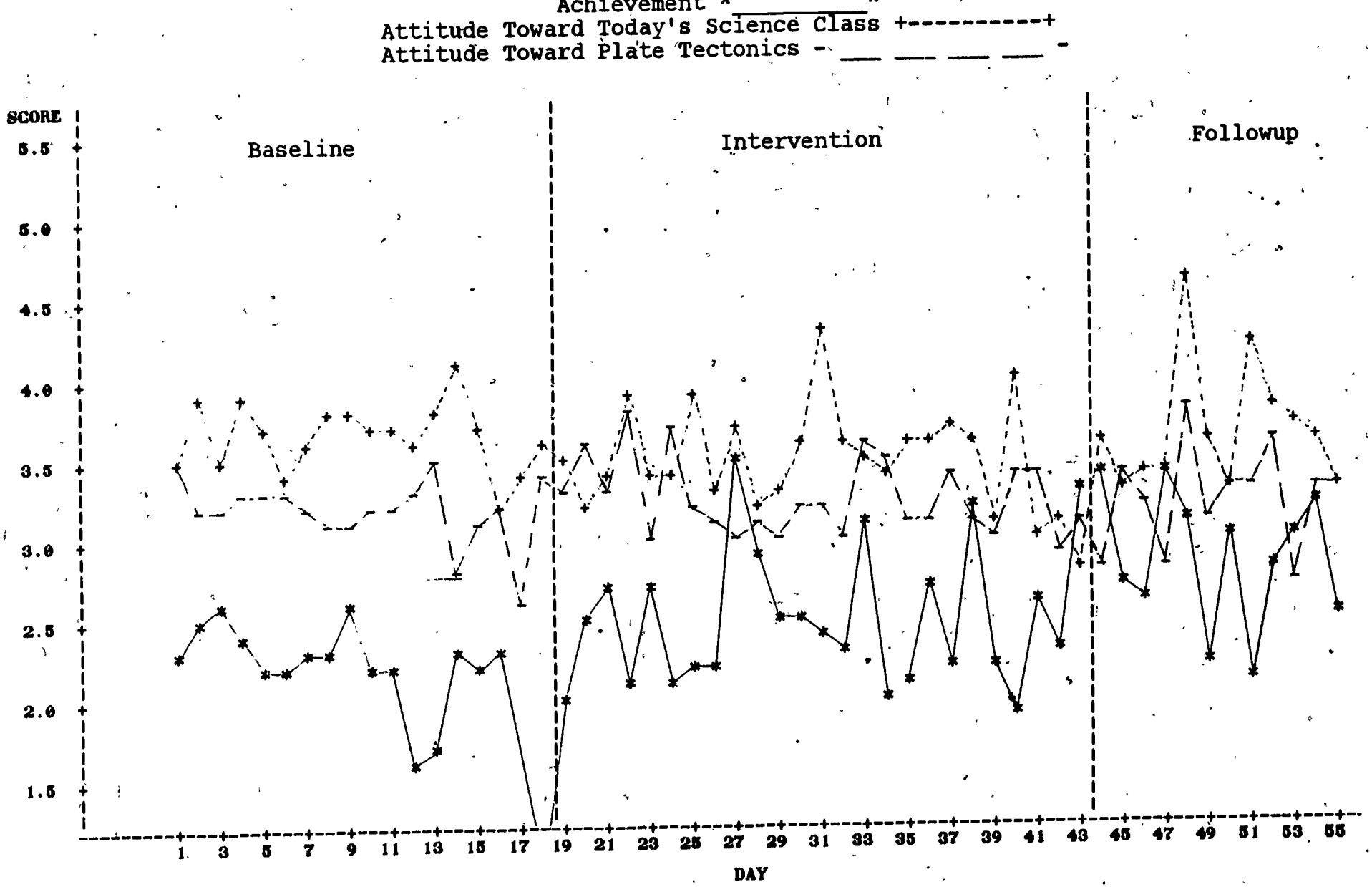


Figure 2
Plot of Concrete Group Daily Measures

stage of the study. This entailed eighteen (18) regressions:

Formal Cognitive Tendency Group

Achievement on Day

Regression 1: Baseline

Regression 2: Intervention

Regression 3: Followup

Attitude Toward Today's Science Class on Day

Regression 4: Baseline

Regression 5: Intervention

Regression 6: Followup

Attitude Toward Plate Tectonics on Day

Regression 7: Baseline

Regression 8: Intervention

Regression 9: Followup

Concrete Cognitive Tendency Group

Above procedures repeated

Results from this set of regressions performed on the sample study data are found in Table 3. Only two were found to be significant, the regression of score on day for the formal group during intervention and the regression of score on day for the concrete group during baseline.

At this point in the analysis procedure, it becomes possible to state some intermediate conclusions as to the nature of temporal trends observed in the daily data. It is also possible to examine hypotheses dealing with: 1) expected differences

Achievement	FORMAL	CONCRETE
Baseline	F .13 p .72 R SQUARED .01	F 11.48 p .01 R SQUARED .42
Intervention	F 18.30 p .0003 R SQUARED .4431	F .45 p .51 R SQUARED .02
Followup	F 2.71 p .13 R SQUARED .21	F .62 p .45 R SQUARED .06

Today's Science Class	FORMAL	CONCRETE
Baseline	F .29 p .60 R SQUARED .02	F .37 p .55 R SQUARED .02
Intervention	F 2.35 p .14 R SQUARED .09	F 1.22 p .27 R SQUARED .05
Followup	F .00 p .99 R SQUARED .00	F .01 p .99 R SQUARED .007

Plate Tectonics	FORMAL	CONCRETE
Baseline	F 3.27 p .09 R SQUARED .1698	F 2.58 p .13 R SQUARED .14
Intervention	F .00 p .95 R SQUARED .00	f 2.43 p .13 R SQUARED .10
Followup	F .00 p .97 R SQUARED .00	F .09 p .76 R SQUARED .01

Table 3
Results of Simple Regressions

between cognitive tendency groups on the achievement and attitude measures, and 2) differences between trends and levels of achievement and attitude measures, regressed on day, within cognitive tendency groups. Several approaches can be taken in detecting such differences. Neter and Wasserman (1974, pp. 160-167) describe procedures for comparing the slopes and levels of regression lines based on regression techniques. Another approach uses analysis of variance procedures to highlight differences due to group and stage. Differences in slope, however, cannot be assessed with such procedures. Concern about violating the assumptions of normal theory tests could also dictate that various nonparametric techniques be employed. In the case of the sample study, the ANOVA procedures were used. Results from these analyses presented in Table 4, indicate significant differences between groups and, in the case of both the achievement measure and the measure of attitude toward today's science class, significant differences were present between baseline, intervention, and followup periods within groups.

The next step in the analysis of the daily data generated by an intensive time-series study is the examination of correlations between dependent variables. This procedure is used to identify patterns in correlations due to stage of the study and cognitive level. In the sample study correlations were performed between achievement scores and each set of attitude scores and between attitude scores. Again 18 analyses were used.

These analyses provide simple comparisons of all factors in

the study, giving an indication of the relationships of the various variables, and the degree to which they fit the hypotheses. In addition, inter-relationships of dependent variables can be investigated. Upon completion of such assessments it is then possible to approach the next step of the analysis procedure, the development of an explanatory model.

The modeling procedure incorporates multiple regression coupled with autocorrelation. The reason for using such a procedure is that a score on a given day is partly the product of the conditions on that day (material taught, attitudes, teacher effectiveness, temperature, humidity, etc.) and partly the product of carry-over effects from previous days (autocorrelated effects).

In most intensive time-series studies to date, autoregressive components have been included in the analysis models. Work by Mayer and Rojas (1982), and Farnsworth (1981) has indicated that benefits can be derived from incorporating autocorrelation in the modeling of intensive time-series data. In general, when autocorrelation has been incorporated, the variance in the data accounted for by the models has increased.

The incorporation of predictor variables such as attitude in the modeling procedure is a recent addition to the intensive time-series design. In studies published to date, no attempt was made to synthesize all measured variables into one model.

To incorporate autocorrelation and multiple regression in one modeling process, the researcher must have access to an appropriate computer autoregression program. In addition a

Table 4

Analysis of Variance of Results Achievement, Attitude Toward Plate Tectonics and Attitude Toward Today's Science Class By Group and Stage of Study

Achievement

Source	df	MSE	F
Group	1	11.6169	53.64*
Stage	2	7.6751	35.44*
Group * Stage	2	.0465	2.15
Error	104		

p .001

Attitude Toward Today's Science Class

Source	df	MSE	F
Group	1	.7433	6.27
Stage	2	.4311	3.64
Group * Stage	2	.0008	.01
Error	104	.1186	

Attitude Toward Plate Tectonics

Source	df	MSE	F
Group	1	1.1504	16.67*
Stage	2	.0464	.67
Group * Stage	2	.0198	.75
Error	104	.0690	

decision must be made not only as to what variables will be regressors in the model but also as to what lags will be included in the autocorrelation model. Lags designate the period of the autocorrelation. A model of lag one (1), for example, indicates that each daily score is the result of the effects on that day and a portion of the effects of the previous day. For the purposes of intensive time-series designs, models incorporating lags of one (1) and five (5) seem most appropriate. Such models will be sensitive to both day to day effects and also such effects as may be caused by day of the week. The results of a modeling procedure using autocorrelation and multiple regression are presented in Table 5. These results were obtained by separately conducting the modeling procedures for each group and each stage in the study. As can be seen by examining the results graphed as Figures 3 and 4, the explanatory power of the modeling procedure can be rather significant.

Table 5
Partial Results from Multiple
Regressions with Autocorrelation - Achievement
Regressed by Day (Trend) and Attitudes

Formal

R Squared		Partials	
	Source	df	t
(Baseline) .4514	Day	1	-1.88*
	Today's Science Class	1	0.15
	Plate Tectonics	1	-3.29
(Intervention) .4649	Day	1	3.11*
	Today's Science Class	1	-1.60
	Plate Tectonics	1	1.06
(Followup) .4349	Day	1	-2.36*
	Today's Science Class	1	-0.42
	Plate Tectonics	1	0.18

**p .01
*p .05

Concrete

R Squared		Partials	
	Source	df	t
(Baseline) .5397	Day	1	-3.77*
	Today's Science Class	1	0.45
	Plate Tectonics	1	-1.15
(Intervention) .3574	Day	1	-0.89*
	Today's Science Class	1	-0.69
	Plate Tectonics	1	-3.20
(Followup) .3148	Day	1	0.06*
	Today's Science Class	1	0.42
	Plate Tectonics	1	-1.77

p .01

Figure 3
Plot of Concrete Group Multiple Regression Autocorrelation
Achievement Score *_____*
Predicted Score + ---------- +
Residual R _____ R

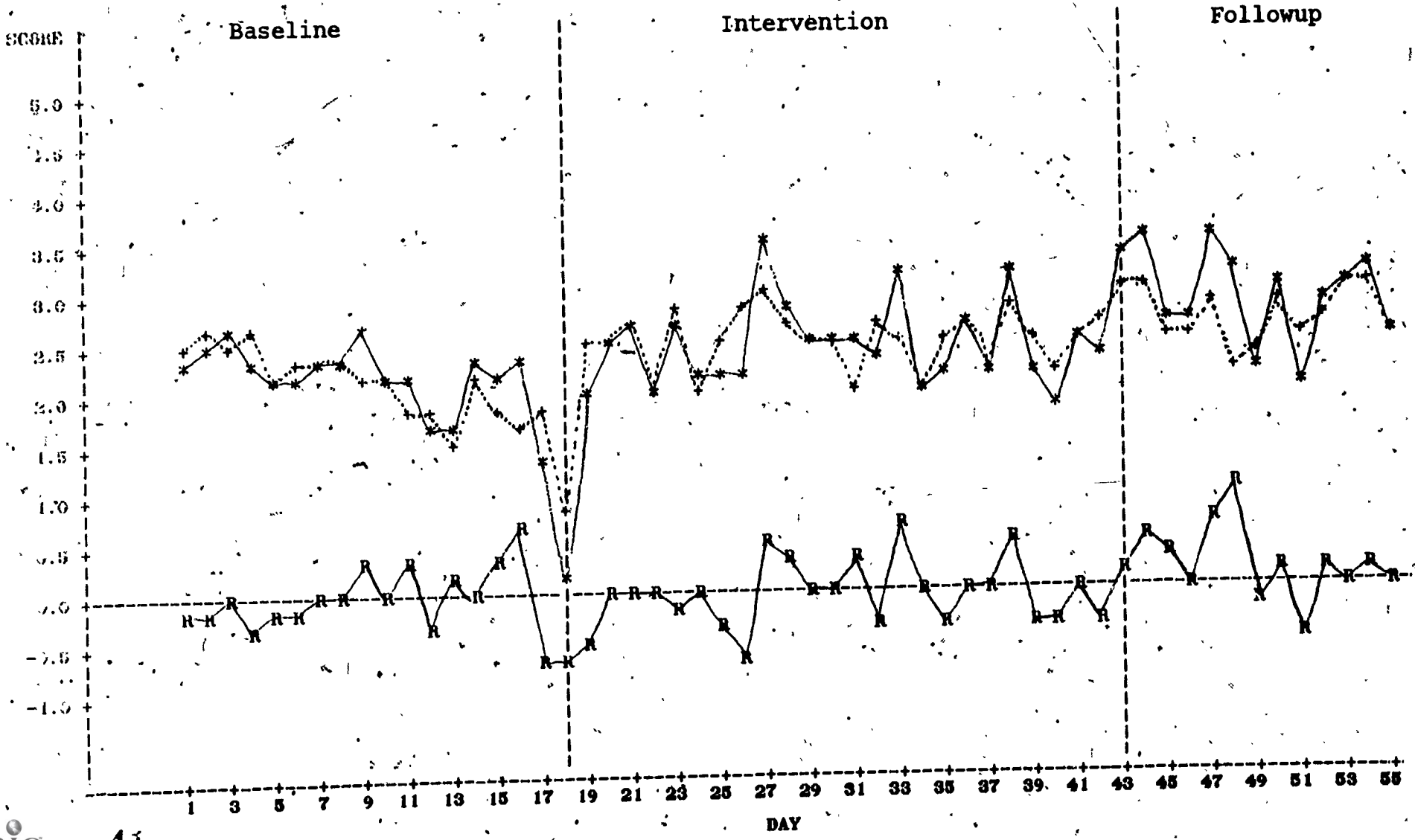
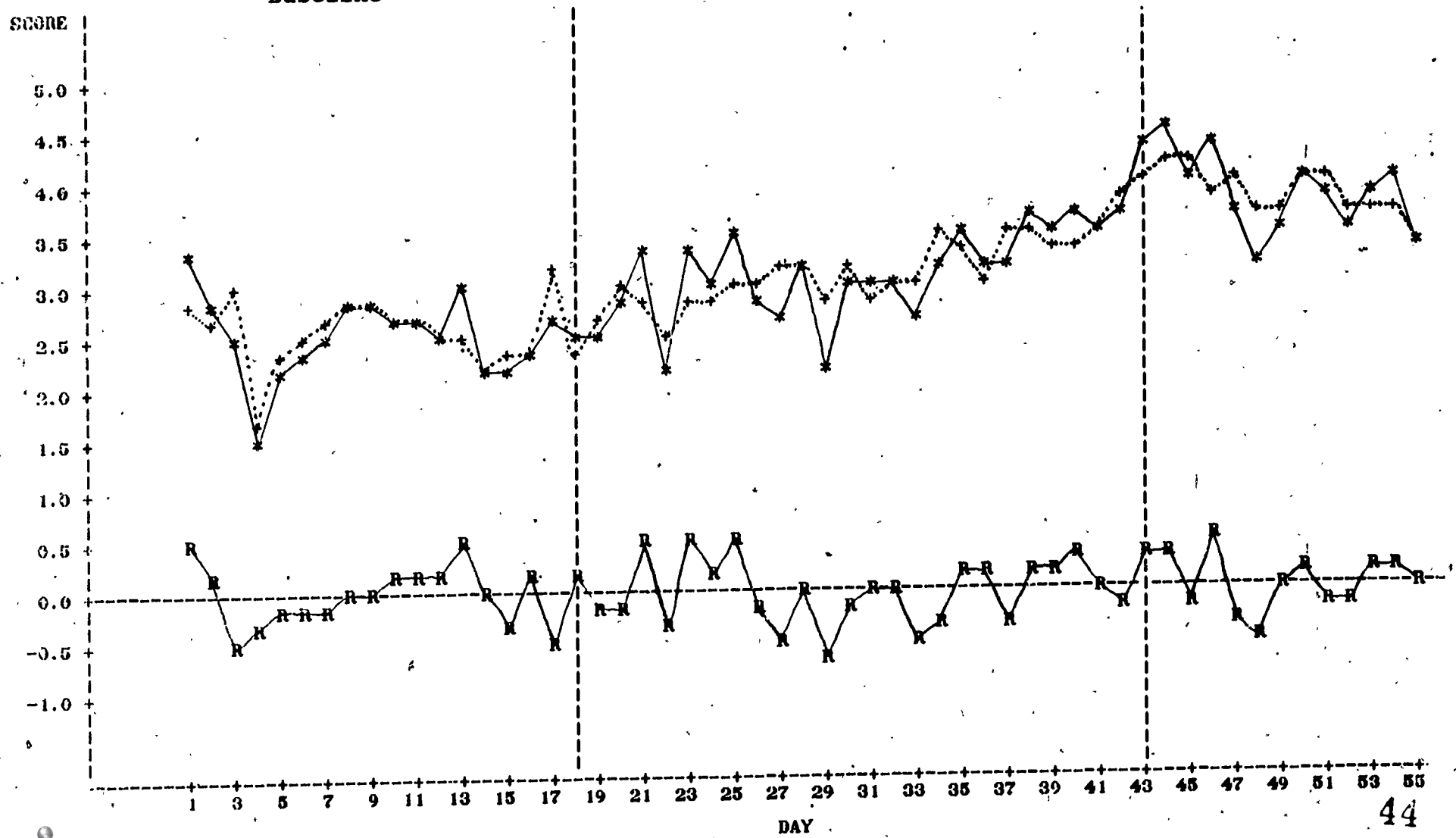


Figure 4

Plot of Formal Group Multiple Regression With Autocorrelation



From the careful examination of the results obtained through the use of the above procedures, a reasonable description of the effects of the predictive variables (i.e. cognitive level, stage of the study, and the two attitude measures) on achievement can be made. In any intensive time-series experiment such descriptive statements are the prime goal.

The final procedure in the intensive time-series analysis scheme is the development of an overall model. This is conducted in the following manner. First, a dummy coding scheme is used to code for the stage of the study. Since the study examined included three stages, two dummy variables are needed. Next, through the examination of the general trends in the baseline, intervention, and followup, a decision is made as to what type of trend to incorporate in the model. In the case of the study being examined here, the overall trend incorporated a downward slope during baseline and followup, and upward slope during intervention. In this completed model, the incorporation of the trend and dummy variables in the multiple regression with correlation, if conducted separately for each group, can produce some striking results. In the case of the sample data, the model, when applied separately to each cognitive tendency group, accounts for over 40% of the variance observed in the knowledge scores of the concrete group and over 60% of the variance observed in the knowledge scores of the formal group. Graphs of the scores predicted by the final models for the concrete and formal groups are presented in Figures 5 and 6.

Upon completion of the final modeling procedures, the

the scores predicted by the final models for the concrete and formal groups are presented in Figures 4 and 5.

Upon completion of the final modeling procedures, the research should be able to address all hypotheses made prior to the start of the study.

Final Conclusions

The approach taken to the analysis of intensive time-series data may best be summarized as a descriptive modeling approach. Final conclusions from the modeling procedures depend upon a thorough examination of both simple and complex models which may be generated by many means. The above procedures represent a synthesis and revision of procedures utilized by Mayer and Lewis (1979), Mayer and Kozlow (1980), Mayer and Rojas (1982), and Farnsworth (1981). These procedures are in no way the only approach that may be utilized. In any approach taken, however, the general flow seen here should be followed. That is: 1) standardize the daily scores in an effort to generate uniform measures of performance, 2) start the analyses procedures by looking for the simplest interactions in the data, and 3) generate more complex models in the best possible manner as needed to explain the effects observed.

Figure 5

Plot of Concrete Group Multiple Regression with Autocorrelation of Achievement Incorporating Dummy Coding For Stage and a Trend Variable

Achievement Score *_____*
Predicted Score + ---------- +
Residual R _____ R

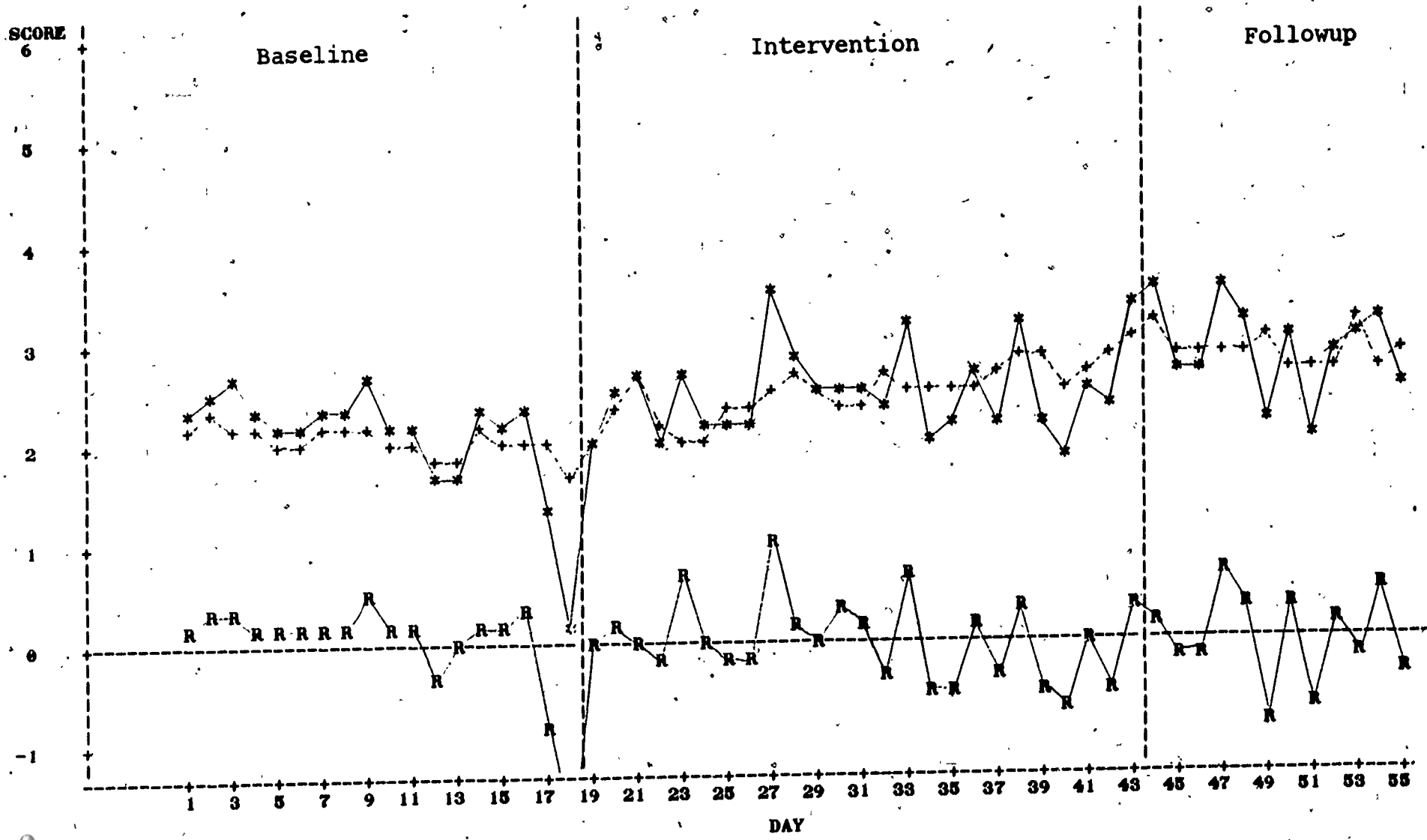
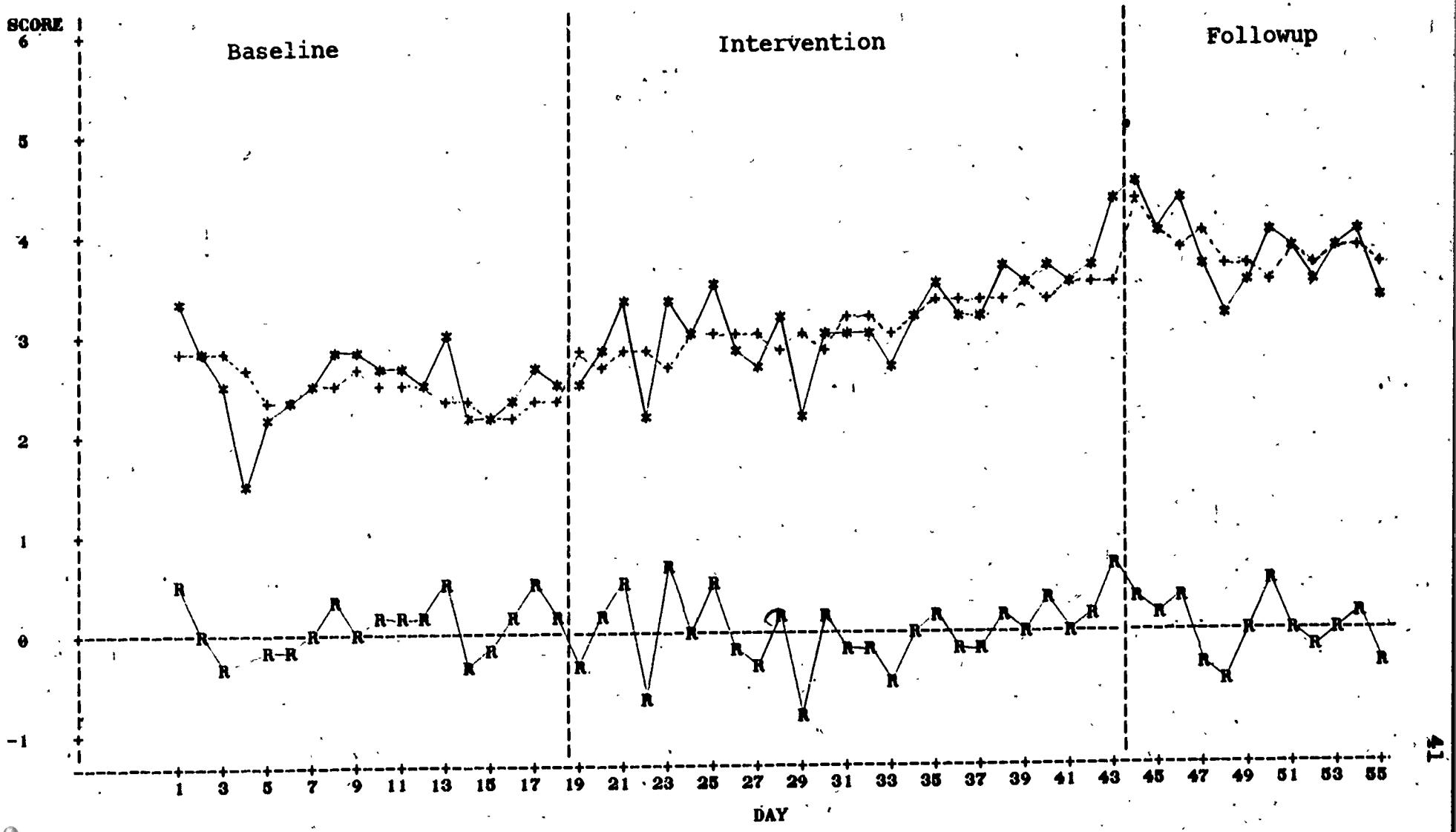


Figure 6

Plot of Formal Group Multiple Regression with Autocorrelation of Achievement
Incorporating Dummy Coding For Stage and a Trend Variable

Achievement Score *_____*
Predicted Score + ---------- +
Residual R _____ R



REFERENCES

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally and Company, 1966.

Farnsworth, C. H. "Using an intensive time-series desgin to develop profiles of daily achievement and attitudes of eigth-grade earth-science students at different cognitive levels during the study of plate tectonics". Unpublished doctoral dissertation. The Ohio State University. Columbus, Ohio, 1981.

Galss, G. V., Willson, V. L., and Gottman, J. M. Design and Analysis of Time-Series Experiments. Boulder, Colorado: Colorado Association Press, 1975.

Lord, F. M. Estimating norms by item sampling. Journal of Educational and Psychological Measurement, 1962, 22, 259-267.

Mayer, V. J. and Kozlow, M. J. An evaluation of a time-series single-subject design used in an intensive study of concept understanding. Journal of Research in Science Teaching, 1980, 17, 455-461.

Monk, J. S. An examination of methods used to generate daily group scores from single item per subject data collected in intensive time-series designs, Journal of Research in Science Teaching. In Press, 1983.

Mayer, V. J. and Lewis, D. K. An evaluation of a time-series single-subject design. Journal of Research in Science Teaching, 1979, 16, 137-144.

Rasch, G., Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark: Neilsen and Lydiche, 1960.

APPENDIX A

SIDE 1

GENERAL PURPOSE — NCS — ANSWER SHEET

SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last First M I)

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

BIRTH DATE — MO DA YR

IDENTIFICATION NUMBER — A B C D E F G H I J

SPECIAL CODES — TC CN #L P

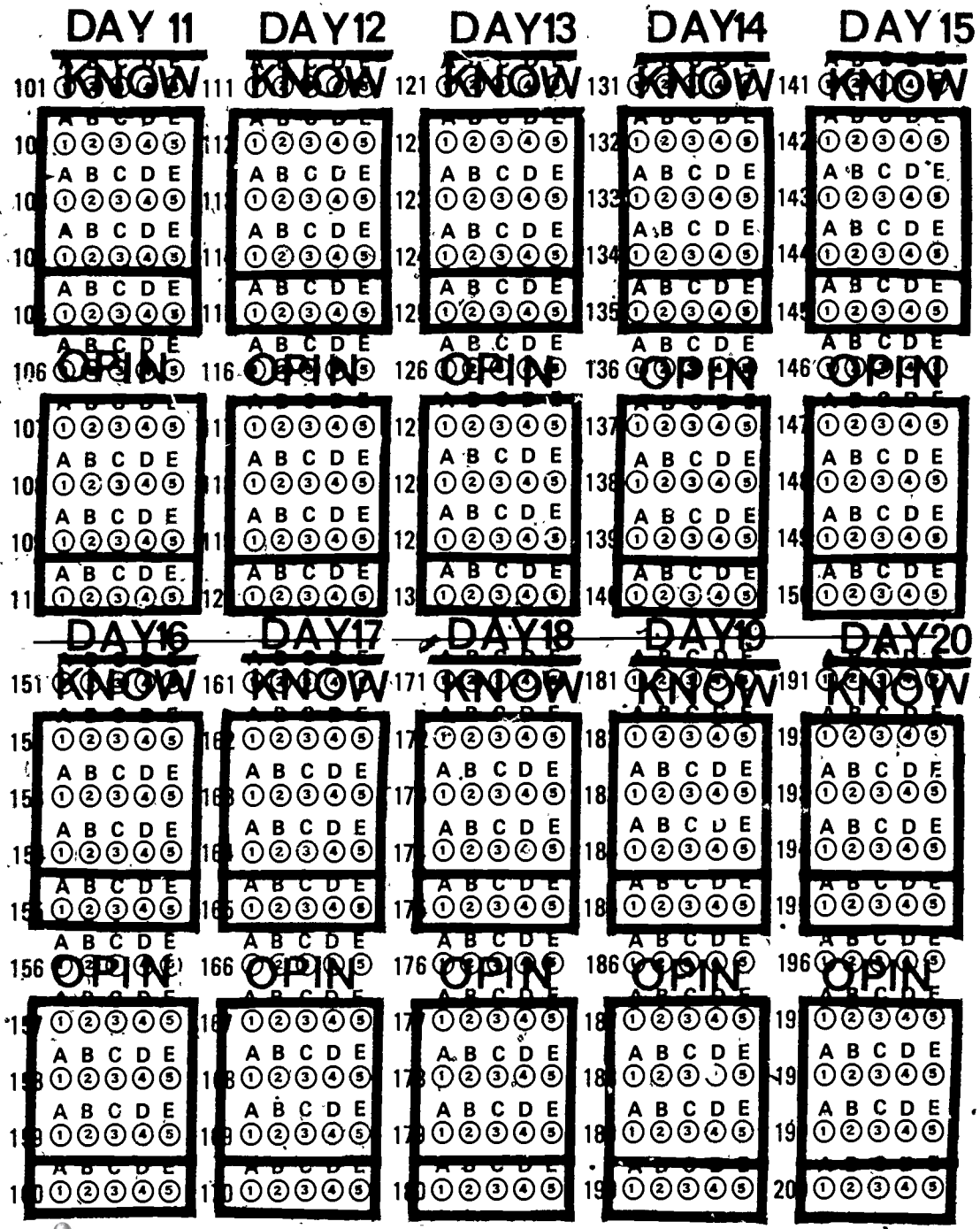
DAY 1 — DAY 2 — DAY 3 — DAY 4 — DAY 5

DAY 6 — DAY 7 — DAY 8 — DAY 9 — DAY 10

KNOW

OPIN

SIDE 2



GENERAL PURPOSE

NCS

ANSWER SHEET

EXAMPLES	IMPORTANT DIRECTIONS FOR MARKING ANSWERS
WRONG 1 ① ⊗ ③ ④ ⑤	• Use black lead pencil only (No. 2½ or softer)
WRONG 2 ① ② ✓ ④ ⑤	• Do NOT use ink or ballpoint pens
WRONG 3 ① ② ③ ● ⑤	• Make heavy black marks that fill the circle completely
RIGHT 4 ① ② ③ ● ⑤	• Erase cleanly any answer you wish to change
	• Make no stray marks on the answer sheet

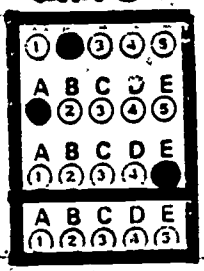
DO NOT WRITE IN THIS SPACE

ns-Optic F 4521 151413121110

© 1977 by National Computer Systems, Inc.

Sample Items from the Plate Tectonics Achievement Item Pool

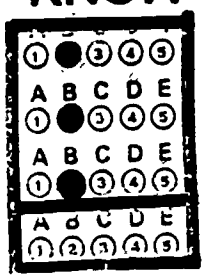
DAY KNOW



According to the theory of plate tectonics, the upward movement of the mantle material causes

(A) mid-ocean ridges.
(B) continental shelves.
(C) sea-floor trenches.
(D) abyssal plains.

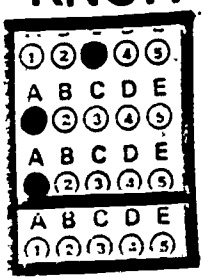
DAY KNOW



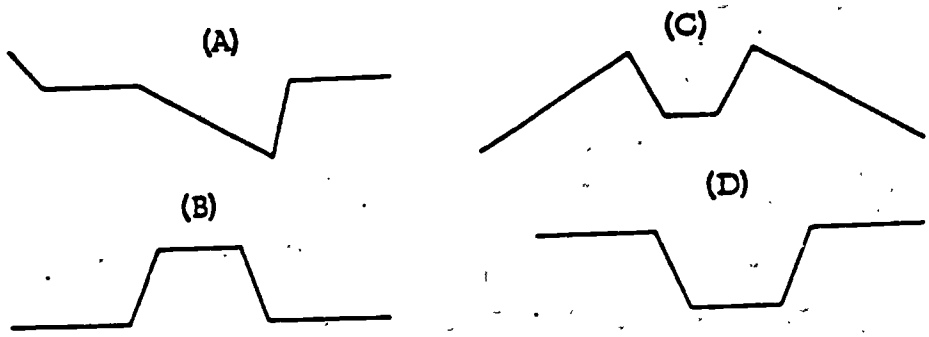
Subducting ocean floor plates usually move

(A) down under the continent.
(B) even with and alongside the continent.
(C) over the continental plate.
(D) out under the ocean.

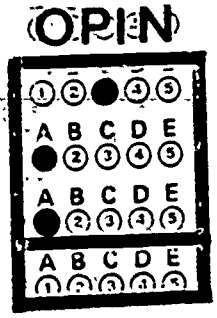
DAY KNOW



Which one of the following diagrams most clearly resembles a bathymetric or topographic profile drawn across a mid-ocean ridge?



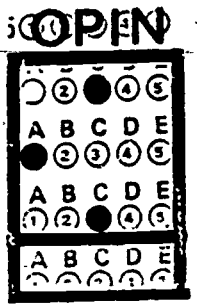
Sample Items from the
Plate Tectonics Attitude Item Pool



Choose the letter that best tells how you feel about

PLATE TECTONICS.

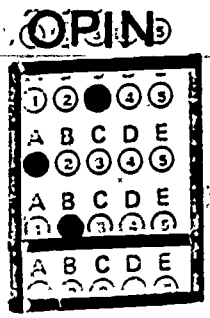
VALUABLE A : B : C : D : E WORTHLESS



Choose the letter that best tells how you feel about

PLATE TECTONICS.

MYSTERIOUS A : B : C : D : E UNDERSTANDABLE

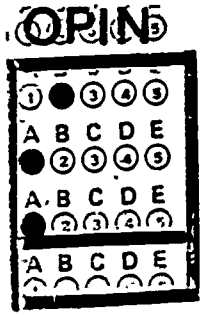


Choose the letter that best tells how you feel about

PLATE TECTONICS.

BORING A : B : C : D : E EXCITING

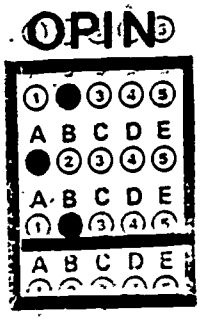
Sample Items from the Today's Science Class Attitude Item Pool



Choose the letter that best tells how you feel about

TODAY'S SCIENCE CLASS.

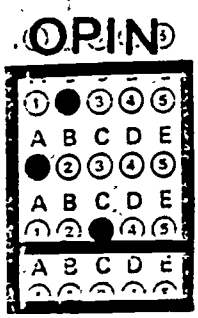
VALUABLE A : B : C : D : E WORTHLESS



Choose the letter that best tells how you feel about

TODAY'S SCIENCE CLASS.

BORING A : B : C : D : E EXCITING



Choose the letter that best tells how you feel about

TODAY'S SCIENCE CLASS.

IRRITATING A : B : C : D : E PLEASANT

APPENDIX B

Rasch Item Calibration

The Rasch (1960) method of item calibration allows the researcher to state that, on each day, the measure of group performance generated from raw subject scores, is scaled, or if you will, standardized, in the same manner. To understand the validity of this statement it is necessary to examine, at least briefly, the nature of the Rasch method.

The Rasch method of calibrating items is based on the hypothesis that the probability that a person will respond correctly to a given item on a test instrument can be expressed as a function of two factors. These factors are:

B - the person's ability (free from influences of item difficulty)

and

d - the difficulty of the item (free from person ability)

These factors can be related by looking at the difference between the two factors (B-d). It is this difference which subsequently governs the individual's probability of getting an item correct. Both B and d can range from zero to infinity, therefore, the difference (B-d) can range between plus and minus infinity. To allow probabilities to assume their correct range, than, the Rasch probability of obtaining a correct score is stated as (Wright, 1977):

$$P = \frac{e^{(B-d)}}{1+e^{(B-d)}}$$

At the point where B equals d, it can be seen that the individual has a 50/50 chance of answering the item correctly. As the difference between B and d becomes negative, it is less probable that the indvidual can answer the item correctly. As the difference becomes positive, a correct answer is more probable. It is this relationship that allows the researcher to arrive at both "item free person measurements" (estimates of B) and "person free item difficulties" (estimates of d) (Wright 1977). Additionally, once individual item difficulties are known it is possible to generate, ability measures from any subset of items of known difficulty. It is this characteristic of Rasch difficulties that makes the use of that method of item calibration so appealing. Once items are calibrated from the multiple item instruments, the difficulties obtained for each of the items can then be utilized with any subtest of items to generate measures of performance which are on a uniform scale. The procedure to generate such measures incorporates four factors. They are: 1) the number of items to which the individual was exposed, 2) the score on each item, 3) the mean difficulty of the items taken, and 4) an expansion factor incorporating the variance of the difficulties of the items on the subtest. The formula for obtaining subject measures is (Wright 1977):

$$B = \bar{d} + 1 + \frac{Var_{Dif}}{248} \ln \frac{C}{T - C}$$

where:

B = estimated measure (ability, attitude etc.)

$\bar{d}$ = mean difficulty of items on the subtest

Var_{diff} = variance of the difficulties on the subtest

$(1+Var_{diff}/2.98)$=expansion factor for variance

C = number of items correct on the subtest

T = total number of items on the subtest

To utilize the above formula for the calculation of measures of group performance from the single-item-per-subject data collected in an intensive time series design, the single items must be treated as if they were items on a single instrument given to a single subject. The single instrument would constitute a subtest generated from the previously calibrated item pool and the single subject would represent the "average" member of the group being studied.

This process of creating a synthetic subtest may be justified heuristically in the following manner. Recall the assumption that the probability of an individual getting an item correct is based on the difference between the measure of person ability (B) and the measure of item difficulty (d) in such a manner that B-d determines the probability of a correct response

to the item. Since items are randomly assigned to subjects on each day of the study, the probability that an individual will receive an item for which there is a 50/50 chance of success or greater can be seen as a function of the distribution of item difficulties in the item pool. A person of high ability has, on any given day, a high probability of getting an item correct. A person of low ability has a low probability of getting an item for which there is a high probability of success. When considering an entire group, the best estimator of the probaibility of an individual getting an item correct can be seen as:

$$P=\frac{e\ (\bar{B}\ \ -\bar{d}\ )}{1\ +\ e^{\ (B\ -d\ )}}$$

where:

$\bar{B}$ = mean group ability

d = mean item pool difficulty

The score obtained on a given instrument is a function of the probability of success on the items of that instrument and the ability of the individual responding to that instrument. It follows that a measure of ability based on the responses of a group of persons each responding to a single randomly selected item on an instrument should accurately reflect the ability of the group. In essence, it is this measure of group ability that the Rasch technique provides.

While Rasch measures of ability and difficulty are not, in theory, that difficult to understand, they are log or logit scores. In this respect, the use of Rasch measures does require

some adjustment on the part of the researcher. The logit nature of the scores, however, does not detract from their usefulness, especially, since Rasch procedures have been developed to calibrate items scored both in a dichotomous (right - wrong) and in a polychotomous (Likert scales for instance) manner.

References on the Rasch Method

Mead, Ronald, Assessing the fit of data to the Rasch model, Paper presented at the American Educational Research Association Annual Meeting. San Francisco, 1976.

Perline, R., Wright, B.D., and Wainer, H. The Rasch model as additive conjoint measurement. Research Memorandum Number 25, Statistical Laboratory, Department of Education The University of Chicago, 1977.

Rasch, G., Probabilistic Models for Some Intelligence and Atttainment Tests. Copenhagen, Denmark: Neilsen and Lydiche, 1960.

Waner, H. and Wright, B.D. Robust estimation of ability in the Rasch model. Psychometrika, 1980, 43, 373-392.

Wright, Benjamin D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, Benjamin D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1968, 85-101.

Wright, B.D. and Mead, R.J., BICAL: Calibrating rating scales with the Rasch model. Research Memorandum Number 23, Statistical Laboratory, Department of Education, The University of Chicago, 1976.