# Combining SAS Data Sets in Educational Research

Leping Liu, Towson University, Baltimore, MD
George C. J. Fernandez, University of Nevada, Reno, NV

## ABSTRACT

In the field of education, research data were usually collected from several sites. Sometimes, the observations lasted several years. And during the process of research, the researchers might consider to add more variables to analyze the relationships or differences. This caused one problem to the data analysis: the data sets – from different sites, different years, or including different variables – needed to be combined or merged into one set. This paper will present an educational study in which several SAS data sets were combined or merged to investigate the influences of several attitude variables on students' computer achievements. The processes and some problems of combining/merging SAS data sets will be discussed.

## INTRODUCTION

In educational research, sometimes the researchers or educators collect data from more than one site (e.g., from several classrooms, schools, regions, or states) or from more than one period of time (e.g., from several school semesters). Sometimes they need to adjust the study by adding more variables or observations. Therefore in their study they usually need to deal with more than one set of data; and one task in their data analysis is to combine these data sets for data analysis. Although combining data sets can be implemented in many ways with different software, SAS system provides a convenient way to do so. The current study lasted two years. The research data were collected during six school semesters, and four more variables were added into the original data sets. The procedures of combining data sets and the processes of data analysis will be presented.

## THE CURRENT STUDY

While technology has been becoming an important component of classroom teaching and learning (Becker, 1991; Liu & Johnson, 1998; Liu 1999), a major task in teacher education is to help schoolteachers learn and use technology (Maddux, Johnson, & Willis, 1997; Liu, & Johnson, 1998). Can teacher education students learn computer skills effectively and obtain necessary basic computer knowledge for their future teaching career in the college setting? Furthermore, what variables, concerned with individual's learning and the learning environment, will influence their computer achievement?

Research findings suggest that students' attitudes toward computers have been one of the major factors relating to their success in learning and using them (Francis & Evans, 1995; Liu & Johnson, 1998; Liu 1999). Computer attitude studies have investigated four attitude variables: (1) enjoyment—the extent to which students like or are interested in learning and working with a computer (King & Bond, 1996); (2) motivation—the degree to which students are willing to learn and use a computer (Kellenberger, 1996); (3) importance—the extent to which students feel learning and using computer as important (Corston & Colman, 1996); and (4) anxiety—the degree to which students feel nervous or uncomfortable in learning and using computer (Cooper & Stone, 1996). These variables were found to be related to students' computer learning. In studying the factors related to these attitude variables, some environmental variables were discussed, such as computer access, convenient help, and computer requirement (Kellenberger, 1996; Liu & Johnson, 1998). These environmental variables were believed to be associated with positive attitudes toward computers.

The purpose of this study was to examine the relationships between (a) the four attitude variables and computer achievement, (b) the three environmental variables and each of the four attitude variables. And finally, another important variable, time spent on learning or using computer, was investigated in relation to computer achievement.

## SUBJECTS AND INSTRUMENTS

The population from which the sample of this study was selected was the total 1061 undergraduate teacher education students enrolled in the College of Education at the University of Nevada, Reno, from 1996 to 1997. Subjects in this study were 609 of the total 1061 teacher education students who were

enrolled in a basic computer technology course. the course was required for all teacher education students in the College of education. The subjects' ages ranged from 18 to 51 (average age 23.26), including 381 females and 288 males. Around 90% had no previous computer skills, beyond using a word processor.

Four instruments were used to collect the data. The first instrument used in this study was derived form Aiken's (1979) Likert-type questionnaire consisting of 24 statements to measure the four attitude variables; we named it internal-variable questionnaire. The second instrument included only three statements to measure the environment variables, named external-variable questionnaire. The third instrument was the combination of the first two, including 27 statements; named achievement-variable questionnaire. Time measurement was another separate instrument, as the fourth questionnaire.

## PROCESSES OF DATA COLLECTION

The data of this study was collected during six school semesters (spring, summer, and fall in 1996 and 1997). Students in the six semesters were specified as subject Group 1 to Group 6 correspondingly. First, the internal-variable questionnaire was distributed to Group 1; data of the four attitude variables were collected and SAS data was created as Data Set 1, measuring enjoyment, motivation, importance, and freedom from anxiety. During the semester, it was found that the issues students frequently talked about were computer access, convenient computer helper, and the computer requirement (required by other courses). Therefore, the second instrument, environmental questionnaire was distributed to the same group. The responses from this questionnaire were created as Data Set 2, measuring the three environmental variables.

Next, in the following five semesters, the third instrument with the measurement of the seven variables (four attitude variables and three environmental variables) was distributed to Groups 2 to 6. The data sets created for the five groups were Data Set 3 to 7 correspondingly.

Finally, the fourth instrument was used for all six groups to measure the average time (minutes per week) students spend on learning or working with a computer. The responses were put into Data Set 8. Achievement is the final scores as the response variable in this study. The variables in the eight data sets are listed in Table 1.

**Table 1. Eight Data Sets in This Study**

| Data Set | Group | Enjoyment | Motivation | Importance | F. Anxiety | Access | Helper | Require | Time | Achievement |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | √ | √ | √ | √ | | | | | √ |
| 2 | 1 | | | | | √ | √ | √ | | |
| 3 | 2 | √ | √ | √ | √ | √ | √ | √ | | √ |
| 4 | 3 | √ | √ | √ | √ | √ | √ | √ | | √ |
| 5 | 4 | √ | √ | √ | √ | √ | √ | √ | | √ |
| 6 | 5 | √ | √ | √ | √ | √ | √ | √ | | √ |
| 7 | 6 | √ | √ | √ | √ | √ | √ | √ | | √ |
| 8 | 1- 6 | | | | | | | | √ | |

## METHODS OF COMBINING DATA SETS

As listed in Table 1, we had eight SAS data sets. When performing data analysis, first we needed to combine these data sets. Basically, there are six methods for combining SAS data sets (SAS, 1995):

1. Concatenating
2. Interleaving

3. One-to-one reading
4. One-to-one merging
5. Match merging
6. Updating

In combining the above data sets, we used the methods of *concatenating* and *one-to-one merging*.

Concatenating appends the observations from one data set to another data set. For example, if we use the method of concatenating to combine two data sets, Data 1 and Data 2, first Data 1 is read sequentially until all observations have been processed, then Data 2 is continually read after Data 1 (see Table 2). In this case, we combine different observations (subjects) with the common variables.

**Table 2. Concatenating Data Sets**

| Data Set | ID | X1 | X2 |
|---|---|---|---|
| Data 1 | 01 | | |
| | 02 | | |
| | 03 | | |
| | 04 | | |
| Data 2 | 011 | | |
| | 012 | | |
| | 013 | | |
| | 014 | | |

One-to-one merging combines observations from two or more data sets by creating observations that contain all of the variables from each contributing data set. Observations are combined based on their relative position in each data set, that is, the first observation in one data set with the first in the other, and so on. Even when there is no common variable in two data sets, we can merge them into one data set. In this case, we combine the same observations (subjects) with different variables. The DATA step reads all observations from all data sets (as in Table 3).

**Table 3. One-To-One Merging Data Sets**

| | Data 1 | | Data 2 | |
|---|---|---|---|---|
| ID | X1 | X2 | Y1 | Y2 |
| 01 | | | | |
| 02 | | | | |
| 03 | | | | |
| 04 | | | | |

## PROCEDURES OF COMBINING DATA

As shown in Table 1, the eight data sets were created separately. They were combined with the following procedures:

1. Data Set 1 was combined with Data Set 2, using the one-to-one merging method with MERGE statement. Data Set 1 contains four variables, and Data Set 2 contains three other variables observed form the same objects as in Data Set 1. After the merging, the new data set contains seven variables for the same subjects, group 1 subjects.
2. The new data set from procedure 1 was combined with Data Set 3, 4, 5, 6, and 7, using concatenating method with SET statement. Data Sets3 through 7 contain the seven variables for Groups 2 through 6 correspondingly. After the concatenating, the new data set contains seven variables for the six groups of subjects.
3. Finally, the new data set from procedure 2 was combined with Data Set 8, using one-to-one merging method with MERGE statement. Data Set 8 contains only one variable with the data for all subjects of the six groups. Therefore, the new data set contains eight variables for all subjects of the six groups. This data set was used for the data analysis.

## DATA ANAYSIS AND RESULTS

According to the purpose of this study, multiple regression was performed to examine the predictive relationships among all the variables—internal and external variables, time, and computer achievement. Several multiple regression analyses and model selections were performed. A SAS macro MULTREG (Fernandez, 1996) was employed for checking the assumptions and performing the standard multiple regression analyses. Liu and Fernandez (1998) presented some findings from the data analyses:

1. Computer achievement could be predicted from a linear combination of the four attitude variables—enjoyment, motivation, importance, and freedom form anxiety.
2. Computer achievement could be predicted from a quadratic function of time spent on computer learning.
3. Time spent on computer learning could be predicted from a quadratic combination of the four attitude variables.

Besides these findings, the data analyses in this study provided some more implications of the relationships among these variables, emphasizing on the relationships between the internal attitude variables and the external environmental variables. The three environmental variables—computer access, computer helper, and computer requirement

—were regressed to each of the four attitude variables. The results show:

1. Attitude variable *enjoyment* could be predicted from a linear combination of access, helper, and requirement variables (see Table 4).
2. Attitude variable *motivation* could be predicted from a quadratic combination of the three environmental variables. The interaction between computer access and computer requirement was found to be significantly related to students' motivation (see Table 5).
3. Attitude variable *importance* could be predicted from a quadratic combination of the three environmental variables. The interactions between (a) computer access and computer requirement, and (b) computer access and computer helpers were found to be significantly related to how students think learning and using computer as important (see Table 6).
4. Attitude variable *freedom from anxiety* could be predicted from a linear combination of access, helper, and requirement variables (see Table7).

## CONCLUSIONS

Combining the four new findings from the current data analyses with the three findings from previous data analyses (Liu & Fernandez, 1998), the conclusions would be interesting:

1. Computer achievement is influenced by the four internal attitude variables (the previous finding).
2. Each computer attitude variable is influenced by the three external environmental variables (the current findings).

Therefore, we can conclude that computer achievement is influenced directly by the internal attitude variables, and indirectly by the external environmental variables. The findings of these direct/indirect relationships have not been discussed or suggested by any other authors; hence, they are important for contributing to the literature in the field of computer achievement studies and computer education for teacher education. This study also raises some further research questions in this field, such as, for some quadratic relationships found in this study, where is the turning-point from increasing to decreasing? And why does this happen? Or why does this turning point exist?

We may conduct further study to investigate these issues. That means, more data sets will be used for the study. If we want to examine or compare some possible reasons or variables, we will need to combine the data sets again. SAS system does provide the convenient methods to do so.

## RFERENCES


Aiken, L. R. (1979). Attitudes toward mathematics and science in Iranian middle schools. *School Science and Mathematics, 79*(3), 229-234.

Becker, H. J. (1991). How computers are used in United States schools: basic data from the 1989 I. E. A. computers in Education Survey. *Journal of Educational Computing Research, 7* (4), 385-406.

Cooper, J., & Stone, J. (1996). Gender, computer-assisted learning, and anxiety: With a little help from a friend. *Journal of Educational Computing Research, 15*(1), 67-91.

Corston, R., & Colman, A. M. (1996). Gender and social facilitation effects on computer competence and attitudes toward computers. *Journal of Educational Computing Research, 14*(2), 171-183.

Fernandez, G. C. J. (1996). *Linear and non-linear regression.* Lectures in APST705 course offered by University of Nevada, Reno, Fall semester, 1996.

Francis, L. J., & Evans, T. E. (1995). The reliability and validity of the Bath County computer attitude scale. *Journal of Educational Computing Research, 12*(2), 135-146.

Kellenberger, D. W. (1996). Preservice teachers' perceived computer self-efficacy based on achievement and value beliefs within a motivational framework. *Journal or Research on Computing in Education, 29*(2), 124-140.

Liu, L., & Fernandez, G. C. J. (1998). Multiple regression and model selection on developing a computer achievement model. In K. Lebouton, S. Gupta, & T. Cermack (Eds.), *Let your SAS Knowledge Grow* (pp.228-233). Oceanside, CA: WUSS.

Liu, L., & Johnson, L. (1998). A computer achievement model: Computer attitude and computer achievement. *Computers in the Schools, 14*(3-4), 33-53.

Liu, L. (1999). The influence of different teaching methods on student attitudes toward technology. *Technology and Teacher Education Annual 1999* (pp. 1400-1405). Charlottesville, VA: AACE.

Maddux, C. d., Johnson, D. L., & Willis, J. W. (1997). *Educational computing: Learning with tomorrow's technology.* Needham Heights, MA: Allyn & Bacon.

SAS Institute. (1995). *Combining and modifying SAS data sets: Examples.* Cary, NC: Author.

## ABOUT THE AUTHORS

Dr. **LEPING LIU** is Assistant Professor, Towson University, Maryland. She has been a SAS user for four years. Her SAS programming areas of expertise include statistical data analysis for educational and psychological research data, application design and development, data presentation, and data management. She currently uses SAS for Windows for statistical analysis and graphics.

Dr. **GEORGE C. J. FERNANDEZ** has ten years of experience teaching introductory statistics, advanced design and analysis, regression models, and SAS programming classes at the University of Nevada, Reno. He has 19 years of SAS experience in BASE, STAT, GRAPH, LAB, INSIGHT, and ETS. He currently uses SAS for Windows for statistical analysis and graphics.

**Table 4. Multiple Regression Output for "Enjoyment"**

Dependent variable: Y Enjoyment

Analysis of Variance

| Source | DF | SS | MS | F | P>F |
|---|---|---|---|---|---|
| Model | 3 | 1371.34955 | 457.116254 | 54.214 | 0.0001 |
| Error | 605 | 5101.13964 | 8.43164 | | |
| C Total | 608 | 6472.48919 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 2.90373 | R-square | 0.2119 |
| Dep Mean | 21.83471 | Adj. R-sq | 0.2080 |
| C.V. | 13.29868 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter = 0 | P > \|T\| |
|---|---|---|---|---|---|
| Intercept | 1 | 14.551559 | 0.59664188 | 24.389 | 0.0001 |
| Access | 1 | 1.052430 | 0.12481488 | 8.432 | 0.0001 |
| Help | 1 | 0.317823 | 0.11105543 | 2.862 | 0.0001 |
| Req | 1 | 0.565343 | 0.10820121 | 5.225 | 0.0001 |

**Table 5. Multiple Regression Output for "Motivation"**

Dependent variable: Y Motivation

Analysis of Variance

| Source | DF | SS | MS | F | P>F |
|---|---|---|---|---|---|
| Model | 5 | 950.03551 | 190.00710 | 20.211 | 0.0001 |
| Error | 603 | 5668.98671 | 9.40130 | | |
| C Total | 608 | 6619.02222 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 3.06615 | R-square | 0.1435 |
| Dep Mean | 22.22606 | Adj. R-sq | 0.1364 |
| C. V. | 13.79513 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter = 0 | P > \|T\| |
|---|---|---|---|---|---|
| Intercept | 1 | 16.903982 | 1.66076807 | 10.178 | 0.0001 |
| Access | 1 | -0.972750 | 0.71715841 | -1.356 | 0.0355 |
| Help | 1 | 0.436019 | 0.09934193 | 3.791 | 0.0002 |
| Req | 1 | 1.553812 | 0.37445879 | 4.149 | 0.0001 |
| Access*Access | 1 | 0.370308 | 0.09934193 | 3.728 | 0.0002 |
| Acces*Req | 1 | -0.282078 | 0.09056803 | -3.115 | 0.0019 |

**Table 6. Multiple Regression Output for "Importance"**

Dependent variable: Y Importance

Analysis of Variance

| Source | DF | SS | MS | F | P>F |
|---|---|---|---|---|---|
| Model | 5 | 509.77203 | 101.95411 | 12.341 | 0.0001 |
| Error | 603 | 4981.54115 | 8.26126 | | |
| C Total | 608 | 5491.31319 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 2.87424 | R-square | 0.0928 |
| Dep Mean | 24.47069 | Adj. R-sq | 0.0853 |
| C. V. | 11.74565 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter = 0 | P > \|T\| |
|---|---|---|---|---|---|
| Intercept | 1 | 26.857503 | 1.62148190 | 16.564 | 0.0001 |
| Access | 1 | -1.353858 | 0.68802502 | -1.968 | 0.0496 |
| Req | 1 | -1.354658 | 0.65758369 | -2.060 | 0.0398 |
| Access*Access | 1 | 0.387959 | 0.09572953 | 4.053 | 0.0001 |
| Req*Req | 1 | 0.371607 | 0.08794409 | 4.225 | 0.0002 |
| Acces*Req | 1 | -0.282677 | 0.08403521 | -3.364 | 0.0008 |

**Table 7. Multiple Regression Output for "Freedom from Anxiety"**

Dependent variable: Y Freedom from Anxiety

Analysis of Variance

| Source | DF | SS | MS | F | P>F |
|---|---|---|---|---|---|
| Model | 3 | 5645.45202 | 41881.81734 | 139.117 | 0.0001 |
| Error | 605 | 8108.24911 | 13.52107 | | |
| C Total | 608 | 13825.70113 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 3.67710 | R-square | 0.4083 |
| Dep Mean | 20.21010 | Adj. R-sq | 0.4054 |
| C.V. | 138.19437 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter = 0 | P > \|T\| |
|---|---|---|---|---|---|
| Intercept | 1 | 5.976191 | 0.71811642 | 8.322 | 0.0001 |
| Access | 1 | 1.730925 | 0.15565416 | 11.120 | 0.0001 |
| Help | 1 | 0.777364 | 0.14052917 | 5.532 | 0.0001 |
| Req | 1 | 1.330760 | 0.13423544 | 9.914 | 0.0001 |