DOCUMENT RESUME

ED 426 096 TM 029 326

AUTHOR	Helms, LuAnn Sherbeck
TITLE	Dealing with Missing Data: How To Write an SPSS Syntax File.
PUB DATE	1999-01-00
NOTE	19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computer Software; Heuristics; *Social Science Research; *Statistical Analysis; Surveys
IDENTIFIERS	Mean (Statistics); *Missing Data; *Statistical Package for the Social Sciences

ABSTRACT

Dealing with missing data is frequently necessary in social sciences research. There are many statistical methods for dealing with missing data. This paper focuses on four of the more basic methods: (1) listwise deletion; (2) pairwise deletion; (3) total mean substitution; and (4) variable mean substitution. Each of the methods is discussed, and the discussion includes instruction on how to perform the technique using the Statistical Package for the Social Sciences (SPSS). Information is also provided on how to write an SPSS syntax file to perform variable mean substitutions. A heuristic data set of 11 individual surveys, each of which contains 20 questions, is used to make the discussion concrete and optimally useful. Three appendixes illustrate determining the variables to delete, deleting some cases and determining means, and performing the mean substitution. (Contains four tables and seven references.) (Author/SLD)

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

ED 426 096

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

LuAnn
Sherbeck-Helms

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
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 OERI position or policy.

Dealing with Missing Data:

How to write an SPSS Syntax File

LuAnn Sherbeck Helms

Texas A&M University

TM029326

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January, 1999.

Abstract

Dealing with missing data is frequently necessary in social sciences research. There are many statistical methods for dealing with missing data. This paper will focus on four of the more basic methods: listwise deletion, pairwise deletion, total mean substitution, and variable mean substitution. Each of these methods will be discussed including instruction of how to perform them using SPSS. Also, information will be provided on how to write a SPSS syntax file to perform variable mean substitution. A heuristic data set of 11 individual surveys (ID) each with 20 questions (V1 to V20) will be utilized so as to make the discussion concrete and optimally useful.

Dealing with Missing Data:

How to write an SPSS Syntax File

Dealing with missing data is frequently necessary in social sciences research. There are many possible causes of missing data. Individuals may inadvertently skip or miss items on a survey or they may drop out of a study before all the data are collected. As Kromrey and Hines (1994) pointed out, we never ignore missing data-- we either decide not to use it or we change it.

There are many statistical methods for dealing with missing data (Little & Rubin, 1987). The present paper will focus on four of the more basic methods: listwise deletion, pairwise deletion, total mean substitution, and variable mean substitution. These methods have been compared to one another in research literature, often with contradictory results, using both real and simulated data (Huberty & Julian, 1995; Kaiser, 1994; Kromrey & Hines, 1994; Raymond & Roberts, 1987). However, according to Kromrey and Hines (1994) very little has been presented in the literature or in software users' manuals to guide the applied researcher. Therefore, in this paper each of these methods will be discussed and step-by-step instruction will be given on how to write a Statistical Package for the Social Sciences (SPSS) syntax file to perform variable mean substitution. A simple heuristic data set of 11 individual surveys (ID) each with 20 questions (V1 to V20) will be used so as to make the discussion concrete.

Step One: Input Data

When inputting data, either in Word or in the SPSS data file, leave a blank space for data that are missing. Do not put a 0 for missing data, because the computer will interpret this not as missing data but as data with a value of 0.

INSERT TABLE 1 HERE.

The Table 1 data set shows that person 1 is missing data for variable 1, person 2 is missing data for variable 2, and so on. Person 11 is missing the last 60% of their data. This data matrix has been chosen for heuristic purposes. It is not likely many researchers will be missing data in this same pattern.

Step Two: Choose a Deletion Method

Listwise Deletion

Using listwise deletion will delete any row in the data matrix that contains any missing data whatsoever. Listwise is the default in SPSS. Every row in the example data set, Table 1, contains one at least one piece of missing data. If we were to run a regression with this data, even though we are only missing about 10%, of the possible data, SPSS would throw out all of our data and tell us that we have no valid cases. According to Delucchi (1994), there are two problems with listwise deletion. First, listwise deletion may increase Type II error. By decreasing our sample size our ability to find statistically significant effects would be decreased. Second, listwise deletion may decrease our ability to generalize our results. If data were not missing at random then the data that are left may only be representative of a subpopulation of our original sample.

Pairwise Deletion

When pairwise deletion is used all available data are used to compute variances and means, while all available pairs of data are employed to compute covariances (Raymond & Roberts, 1987). Using pairwise also decreases sample size and may increase Type II error.

Although pairwise deletion also tends to increase Type II error and to decrease the representativeness of the remaining sample, these problems tend not to be as severe with this option as with listwise deletion. For example, for the Table 1 data, listwise deletion removes every case from the analysis involving all 20 variables, but pairwise deletion would not delete all 11 cases of data, and some analyses would now be possible.

For example, the mean of variable 1 would be computed as the mean of the scores of persons 2 through 11. The mean of variable 2 would be computed as the mean of the scores of persons 1 and persons 3 through 11. If variables 1 and 2 were correlated, the correlation coefficients would be computed using only persons who had no missing data on either variable (i.e., the nine cases 3 through 11).

Of course, these examples highlight a separate problem that may occur when pairwise deletion is employed. Various statistics may all be calculated on different subsamples of persons, and so results may not be directly comparable, or at least must be compared with considerable caution.

Deletion Before Substitution

Before performing a substitution method on the missing data, there may be entire rows of data that should first be deleted, due to excessive amounts of missing data. In Table 1, persons 1 through 10 answered all but one question on the survey. We may not wish to delete these cases from our data set, since each of these 10 persons is only missing 5% of the possible data (i.e., one out of 20 scores). On the other hand, person 11 did not answer more than half the questions on his survey. In this case, we may decide not to include Person 11 in any statistical analysis.

One helpful vehicle to determine whom, if anyone, to delete from the data set is the SPSS "COUNT" command. This procedure is illustrated in Appendix A. In the example SPSS creates a new variable, "MISSING", that is a count of the missing data for each person (here the scores for "MISSING" on the Table 1 data would be 1 for persons 1 through 10, and 12 for person 11). After the variable, "MISSING", is created, the FREQUENCIES procedure can be run on this count variable; the results from the analysis here are displayed in Table 2. These results are analyzed to identify gaps in the number of missing data points for the study participants, and then to select a criterion for deleting cases with excessive missing data.

For example, lets say that 90% of the sample had 0 missing data, 4% had 1 missing score out of 55 possible scores, 3% had 2 missing scores, 1% had 3 missing scores, 1% had 4 missing scores, and 1% had 12 missing scores. Here a reasonable decision might be to delete all cases with more than 4 missing data points, (a) because 12 of our 55 scores is deemed excessive, and (b) the gap between 4 and 12 missing data points suggests that this may be a reasonable criterion. Of course, judgment must be exercised in making these decisions, and different researchers may reasonably disagree as regards these choices.

Step Three: Decide on What Substitution Method, if Any, to Use.

Although there are many different types of substitution methods (see Little & Rubin, 1987) this paper will discuss mean substitution. There are two different types of mean substitution: total mean and variable mean substitution.

Total Mean Substitution

In total mean substitution all missing data are replaced with the mean of all the variables. It is the easy substitution method in SPSS. It is included as an option, using the term MEANSUB, on some of the analyses (e.g., regression). For this data the total mean for the variables equaled 3.35. To use total mean substitution for this data set we can add the command 'SET BLANKS = 3.35.'

Although total mean substitution is easy it is not usually the preferred method of substitution. There are two reasons it may not be preferred. First, using the total mean can decrease the reliability of some variables. The total mean for the study across all variables and all people may be unduly influenced by outlying people and outlying variables, because the mean is most influenced by atypicality. For example, let's say 8 of 10 people surveyed responded to question 10 with a one, on a scale of one to seven, and two people left the item blank. If you substituted all missing values with the total mean of 3.35, then you would get a mean of 1.43 for question 10. If you would have used the variable mean for the substitution you would have gotten a mean of 1.00 for question 10, which may be a better representation of the participants' responses. Second, by using the total mean you could decrease your chances of getting statistically significant differences between variable means. By substituting the total means you could decrease the differences between the variable means. For example, a mean of 1.43 for question 10 is closer to the total mean of 3.35 then a mean of 1.00 would be. As the variables become closer to the total mean the range of variable means across items would get smaller.

Variable Mean Substitution

In variable mean substitution missing data on a given item are replaced with the mean derived from the valid cases providing data on that single item. Using variable mean substitution in SPSS requires a multi-stage substitution process. An example of this process, the related syntax will be outlined. The complete syntax files are listed in Appendices A through C.

Step 1: Read your data and set blanks to -99999. In social sciences research it is rare to have a number this large or a negative. By changing missing data to -99999 it will be easier to determine whether or not the substitution process has worked, because if this value is erroneously treated as a legitimate score, the effect on all descriptive statistics (e.g., means) will be so dramatic that the error will be obvious. Here are the SPSS commands that accomplish this substitution for all missing data:

```
SET BLANKS = -99999 PRINTBACK = LISTING
DATA LIST FILE='A:\MISSINGDATA.DAT' FIXED RECORDS=1
/ID 1-2 V1 TO V20 4-23 .
```

Step 2: Delete persons with too much missing data. First, choose a variable name (e.g., “MISSING”) and then count the number of missing data for each person Table 2 illustrates the output from this analysis. Then decide how much missing data is acceptable, as described previously. For this example only people who are missing less than two pieces of data are considered valid cases.

```
COUNT MISSING=V1 TO V20(-99999).
FREQUENCIES VARIABLES=MISSING .
SELECT IF (MISSING LT 2) .
```

Step 3. Find the means for each item based on the valid cases. First, delete any cases with excessive missing data, using the “SELECT IF” command. Next, declare that scores of -99999 are actually markers for missing data, and not valid scores, by using the “MISSING VALUES” command. Table 3 presents these results for the heuristic data. This process is illustrated both below and in Appendix B.

```
MISSING VALUES ID TO V20(-99999) .
DESCRIPTIVES VARIABLES=ALL .
```

Step 4. Replace missing data with the variable means. The SPSS syntax for this last step is illustrated in Appendix C. First, go back to the syntax statement "MISSING VALUES ID TO V20 (-99999)" and in front of this statement type the word "COMMENT". This is necessary for the “IF” statements to work. Then insert a series of “IF” statements to substitute the variable means for cases with some (but not excessive) missing data.

```
IF (V1 LT -1) V1 =3.11 .
IF (V2 LT -1) V2 =3.33 .
.
.
.
IF (V20 LT -1) V20 =4.60 .
```

To see the final output add the statement DESCRIPTIVES VARIABLES=ALL. Then rerun the syntax file.

Table 4 consists of the final output for this data. Notice when you compare Tables 3 and 4 that the N for all the variables is now 10, the mean for each variable did not change, and the standard deviations for the items that used mean substitution are slightly smaller. Since correlation can be expressed as: $r_{XY} = COV_{XY} / (SD_X * SD_Y)$ seeing that mean substitution decreases standard deviation some researchers may assume that using mean substitution increases correlation coefficients. But, Walsh (1996) has shown that mean substitution tends to attenuate correlation coefficients. The numerator of this formula can be expressed as: $COV_{XY} = (SUM\ (X_i - Xbar)\ (Y_i - Ybar))/\ n-1$. According to Walsh (1996), when we substitute the mean for a given missing score then that person's deviation score will necessarily be zero. Thus, for any item that we estimate missing data, the numerator of the COV (the product of that person's deviation scores) will be 0, making that mean substitution push the correlation closer to 0. Therefore, we would not want to use mean substitution with too many cases, especially when the correlation coefficients are already fairly close to 0.

Conclusion

This paper briefly discussed four basic methods for dealing with missing data: listwise deletion, pairwise deletion, total mean substitution, and variable mean substitution. A step-by-step plan for developing a SPSS syntax file for replacing missing data using variable means substitution was also provided. It is important to note that SPSS as well was other statistical computer packages are under constant revision and updating. Many programs are in development, or have been developed, that provide researchers with the ability to easily replace missing data by using a choice of several data replacement methods. Of course, it will be important for the researcher to understand the possible effects each method may have on their results, as discussed here as regards variable mean substitution.

References

Delucchi, K. L. (1994). Methods for the analysis of binary outcome results in the presence of missing data. Journal of Consulting and Clinical Psychology, 62, 569-575.

Huberty, C. J., & Julian, M. W. (1995). An ad hoc analysis strategy with missing data. Journal of Experimental Education, 63, 333-342.

Kaiser, J. (1994, August). The estimation of correlation matrix for data having missing values. Paper presented at the Islamic countries conference on statistical sciences, Lahore, Pakistan. (ERIC Document Reproduction Service No. ED 380 473)

Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. Educational and Psychological Measurement, 54, 573-593.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. Educational and Psychological Measurement, 47, 13-25.

Walsh, B. D. (1996). A note on factors that attenuate the correlation coefficient and its analogs. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 21-32). Greenwich, CT: JAI Press.

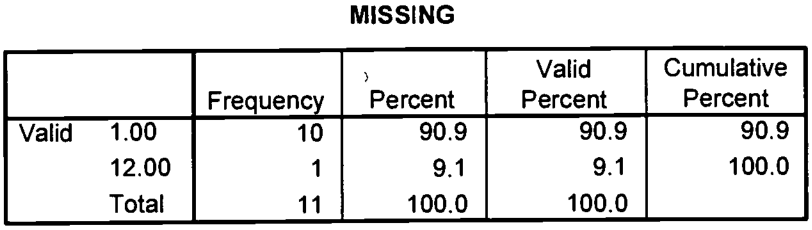
Table 1

The data set as it would appear in Word. File name: A:\MISSINGDATA.DAT

```
01  7465241312322345364
02 3 3564234132463427 35
03 53 342345134344 21424
04 421 3453414724363322
05 2244 234512423453245
06 12352 12612134423434
07 345453 3413443233456
08 3423243 314213634235
09 32336433 12453443244
10 443354341 3344544347
11 63427534
```

Table 2

SPSS Frequencies table for missing data.



MISSING

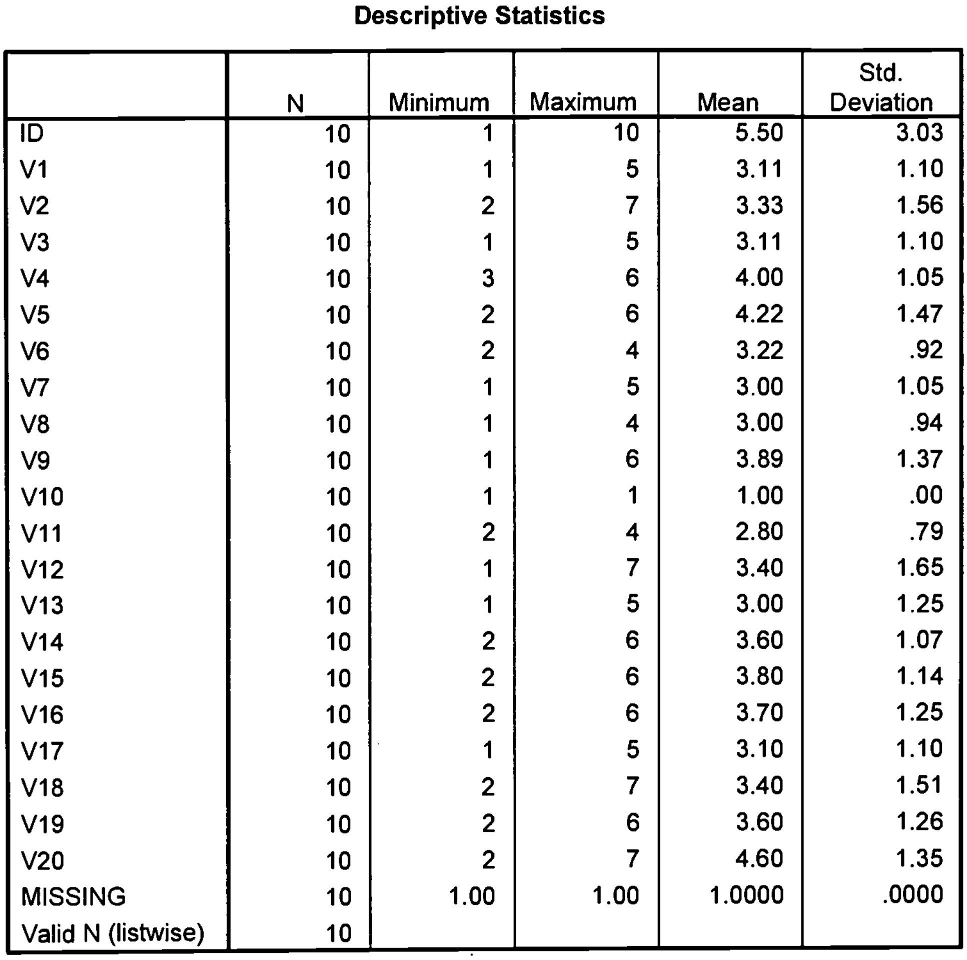
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	10	90.9	90.9	90.9
	12.00	1	9.1	9.1	100.0
	Total	11	100.0	100.0	

Table 3
SPSS descriptive statistics chart used to find each variable means.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ID	10	1	10	5.50	3.03
V1	9	1	5	3.11	1.17
V2	9	2	7	3.33	1.66
V3	9	1	5	3.11	1.17
V4	9	3	6	4.00	1.12
V5	9	2	6	4.22	1.56
V6	9	2	4	3.22	.97
V7	9	1	5	3.00	1.12
V8	9	1	4	3.00	1.00
V9	9	1	6	3.89	1.45
V10	9	1	1	1.00	.00
V11	10	2	4	2.80	.79
V12	10	1	7	3.40	1.65
V13	10	1	5	3.00	1.25
V14	10	2	6	3.60	1.07
V15	10	2	6	3.80	1.14
V16	10	2	6	3.70	1.25
V17	10	1	5	3.10	1.10
V18	10	2	7	3.40	1.51
V19	10	2	6	3.60	1.26
V20	10	2	7	4.60	1.35
MISSING	10	1.00	1.00	1.0000	.0000
Valid N (listwise)	0				

Table 4
The final SPSS descriptive statistics chart with missing data substituted by the variable means.



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ID	10	1	10	5.50	3.03
V1	10	1	5	3.11	1.10
V2	10	2	7	3.33	1.56
V3	10	1	5	3.11	1.10
V4	10	3	6	4.00	1.05
V5	10	2	6	4.22	1.47
V6	10	2	4	3.22	.92
V7	10	1	5	3.00	1.05
V8	10	1	4	3.00	.94
V9	10	1	6	3.89	1.37
V10	10	1	1	1.00	.00
V11	10	2	4	2.80	.79
V12	10	1	7	3.40	1.65
V13	10	1	5	3.00	1.25
V14	10	2	6	3.60	1.07
V15	10	2	6	3.80	1.14
V16	10	2	6	3.70	1.25
V17	10	1	5	3.10	1.10
V18	10	2	7	3.40	1.51
V19	10	2	6	3.60	1.26
V20	10	2	7	4.60	1.35
MISSING	10	1.00	1.00	1.0000	.0000
Valid N (listwise)	10				

Appendix A

Using the COUNT Command to Determine Who to Delete

```
TITLE 'MISSINGDATA.SPS'.
SET BLANKS = -99999 PRINTBACK = LISTING
DATA LIST
FILE='A:\MISSINGDATA.DAT' FIXED RECORDS=1
/ID 1-2 V1 TO V20 4-23 .
SUBTITLE 'NUMBER OF MISSING VALUES PER PERSON'.
COUNT MISSING=V1 TO V20(-99999).
FREQUENCIES VARIABLES=MISSING .
```

Appendix B

Deletion of Some Cases and Determining Means

```
TITLE 'MISSINGDATA.SPS'.
SET BLANKS = -99999 PRINTBACK = LISTING
DATA LIST
FILE='A:\MISSINGDATA.DAT' FIXED RECORDS=1
/ID 1-2 V1 TO V20 4-23 .
SUBTITLE 'NUMBER OF MISSING VALUES PER PERSON'.
COUNT MISSING=V1 TO V20(-99999).
FREQUENCIES VARIABLES=MISSING .
SUBTITLE 'DELETE PERSONS WITH TOO MUCH MISSING DATA'.
SELECT IF (MISSING LT 2) .
MISSING VALUES ID TO V20(-99999) .
EXECUTE .
SUBTITLE 'FIND MEANS FOR EACH VARIABLE BASED ON THE NUMBER OF
VALID CASES'.
DESCRIPTIVES VARIABLES=ALL .
```

Appendix C

Perform the Mean Substitution

```
TITLE 'MISSINGDATA.SPS'.
SET BLANKS = -99999 PRINTBACK = LISTING
DATA LIST
FILE='A:\MISSINGDATA.DAT' FIXED RECORDS=1
/ID 1-2 V1 TO V20 4-23 .
SUBTITLE 'NUMBER OF MISSING VALUES PER PERSON'.
COUNT MISSING=V1 TO V20(-99999).
FREQUENCIES VARIABLES=MISSING .
SUBTITLE 'DELETE PERSONS WITH TOO MUCH MISSING DATA'.
SELECT IF (MISSING LT 2) .
Comment MISSING VALUES ID TO V20(-99999) .
EXECUTE .
SUBTITLE 'FIND MEANS FOR EACH VARIABLE BASED ON THE NUMBER OF
VALID CASES'.
DESCRIPTIVES VARIABLES=ALL .

IF (V1 LT -1) V1 =3.11 .
IF (V2 LT -1) V2 =3.33 .
IF (V3 LT -1) V3 =3.11 .
IF (V4 LT -1) V4 =4.00 .
IF (V5 LT -1) V5 =4.22 .
IF (V6 LT -1) V6 =3.22 .
IF (V7 LT -1) V7 =3.00 .
IF (V8 LT -1) V8 =3.00 .
IF (V9 LT -1) V9 =3.89 .
IF (V10 LT -1) V10 =1.00 .
IF (V11 LT -1) V11 =2.80 .
IF (V12 LT -1) V12 =3.40 .
IF (V13 LT -1) V13 =3.00 .
IF (V14 LT -1) V14 =3.60 .
IF (V15 LT -1) V15 =3.80 .
IF (V16 LT -1) V16 =3.70 .
IF (V17 LT -1) V17 =3.10 .
IF (V18 LT -1) V18 =3.40 .
IF (V19 LT -1) V19 =3.60 .
IF (V20 LT -1) V20 =4.60 .
EXECUTE.
MISSING VALUES ID TO V20(-99999) .
DESCRIPTIVES VARIABLES=ALL.
```

TM029326

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

ERIC®

I. DOCUMENT IDENTIFICATION:

Title: Dealing with missing data: How to write a SPSS syntax file	
Author(s): LuAnn Sherbeck Helms	
Corporate Source:	Publication Date: 1/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LuAnn Sherbeck Helms
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: X LuA Helms	Position: RES ASSOCIATE
Printed Name: LuAnn Sherbeck Helms	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1831
	Date: 11/9/98

ERIC Full Text Provided by ERIC