

**Quantitative Methods in Education and
the Behavioral Sciences:
Issues, Research, and Teaching**

Ronald C. Serlin, Series Editor

Structural Equation Modeling: A Second Course (2005)
edited by Gregory R. Hancock and Ralph O. Mueller

Real Data Analysis (2006)
edited by Shlomo Sawilowsky

*Feng, Hamwell, Lion, & Emswiler (2006).
Advances in missing data methods
and implications for educational research.
In S. Sawilowsky (Ed.), Real data analysis
(pp. 31-78). Greenwich, CT: Information
Age Publishing, Inc.*

Real Data Analysis

edited by

Shlomo S. Sawilowsky
Wayne State University



Charlotte, North Carolina • www.infoagepub.com

Library of Congress Cataloging-in-Publication Data

Real data analysis / edited by Shlomo S. Sawilowsky.

p. cm. — (Quantitative methods in education and the behavioral sciences)

Includes bibliographical references.

ISBN-13: 978-1-59311-564-7 (pbk.)

ISBN-13: 978-1-59311-565-4 (hardcover)

1. Social sciences—Statistics—Methodology. 2. Social sciences—Research—Methodology. I. Sawilowsky, Shlomo S.

HA29 R32 2007

300.72—dc22

2006033775

ISBN 13: 978-1-59311-564-7 (pbk.)

ISBN 13: 978-1-59311-565-4 (hardcover)

ISBN 10: 1-59311-564-4 (pbk.)

ISBN 10: 1-59311-565-2 (hardcover)

Credits: The back cover image is by Student (William Sealey Gosset), "Probable error of a correlation coefficient," *Biometrika*, 1908, 6(1), p. 305, copyrighted by Oxford University Press, and reprinted with permission.

Copyright © 2007 IAP—Information Age Publishing, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the publisher.

Printed in the United States of America

CONTENTS

Preface

Shlomo S. Sawilowsky

ix

PART I: FOUNDATIONS

1. The Coevolution of Statistics and Hz

Joseph M. Hilbe

3

2. Effective Sample Size: A Crucial Concept

Thomas R. Knapp

21

3. Advances in Missing Data Methods and Implications for Educational Research

Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, and Lee H. Ehman

31

4. Methods for Simulating Real World Data for the Psycho-Educational Sciences

Todd Christopher Headrick

79

5. How and Why I Use Real, Messy Data to Investigate Theory and Inform Decision Making

Theodore Micceri

99

PART II: STATISTICAL METHODS

6. Using E-mail Messages to Help Students Prepare for a Statistics Exam

Schuyler W. Huck

107

v

*Connections on 4/23/07 by
Joanne Peng*

Here is an update of the links referred in the Peng, Harwell, Liou, and Ehman chapter (3):

[mianalyze-lr.fit(5var).sas] now at
https://oncourse.iu.edu/access/content/user/peng/Filemanager_Public_Files/articles/mianalyze-lr.fit_5var_.sas

[lr-em.fit.sas] now at
https://oncourse.iu.edu/access/content/user/peng/Filemanager_Public_Files/articles/lr-em.fit.sas

[persist.sav] now at
https://oncourse.iu.edu/access/content/user/peng/Filemanager_Public_Files/articles/persist.sav

[logregdata(peng).sav] now at
https://oncourse.iu.edu/access/content/user/peng/Filemanager_Public_Files/articles/logregdata_peng_.sav

[logregdata.por] now at
https://oncourse.iu.edu/access/content/user/peng/Filemanager_Public_Files/articles/logregdata.por

SPSS missing value module now at

http://www.spss.com/missing_value/

CHAPTER 3

ADVANCES IN MISSING DATA METHODS AND IMPLICATIONS FOR EDUCATIONAL RESEARCH

Chao-Ying Joanne Peng, Michael Harwell,
Show-Mann Liou, and Lee H. Ehman

Consider a scenario in which a freshly minted PhD showed off his bound dissertation to a fellow graduate student who was known among her peers to be a "number cruncher." As she leafed through the pages of the methods section, she quickly noticed that all missing data were coded zero, instead of being declared as missing. She also saw that all analyses in this dissertation, including factor analysis, multiple regression, discriminant function analyses, and so forth, were based on all data including the zeros. No one on the dissertation committee caught this mistake. In another scenario, a PhD candidate explaining the results of the data analyses to her dissertation committee noted that the choice of listwise deletion led to the removal of approximately one third of the data. Both the candidate and committee members failed to recognize the possibly devastating effects on inferences due to these missing data. Variations of these scenarios may occur more frequently in educational research than is commonly believed or acknowledged.

Real Data Analysis, pp. 31–78
Copyright © 2007 by Information Age Publishing
All rights of reproduction in any form reserved.

Missing data in research studies are the rule rather than the exception. For example, various literatures indicate that a 20% attrition rate has been held as a benchmark for studies related to youth, school-based programs, and clinical research (Hall, 1993; Kellam, Rebok, Ialongo, & Mayer, 1994; Mason, 1999). Many reasons contribute to data missing from research projects: subject's attrition in pretest-posttest or longitudinal designs, equipment failure, unclear instructions or intrusive questions on a survey, and poor data entry and/or record keeping in archived data—the latter was shown to introduce hidden bias in the data (Dworkin, 1987). Whatever the reason(s) for missing data, the impact on quantitative research has been a great concern to methodologists.

In survey research, missing data problems are particularly acute, with substantial amounts of missing data frequently appearing (Little & Rubin, 1987). To combat these problems, investigators spend enormous time, funding, and energy to minimize incomplete data or nonresponse among respondents (Mason, 1999). Procedures have been devised by statisticians to mitigate inferential or descriptive problems caused by missing data (Dempster, Laird & Rubin, 1977; Glasser, 1964; Rubin, 1987; Schafer, 1997). The development of these principled procedures is a unique feature of the survey research literature. For the past 2 decades, the statistical properties of these procedures have been studied and disseminated, and many have been incorporated into widely accessible software since 1990 (Cool, 2000; Kim & Curry, 1977; Mason, 1999; Raymond & Roberts, 1987; Witta, 2000). Even so, an examination of articles published in 11 education journals for the past 7 years reveals that, in studies where missing data were reported and care was taken to treat missing data, an overwhelming majority used ad hoc methods, such as listwise and pairwise deletion, to handle missing data. Ad hoc methods are characterized by their failure to take into account the mechanism that led to missing data. Yet the recent APA Task Force on Statistical Inference explicitly warned against the use of these methods:

Special issues arise in modeling when we have missing data. The two popular methods for dealing with missing data that are found in basic statistics packages—listwise and pairwise deletion of missing values—are among the worst methods available for practical applications. (Wilkinson & Task Force on Statistical Inference APA Board of Scientific Affairs, 1999, p. 598)

Newer and principled methods take into consideration conditions under which missing data occurred. In this chapter, we promote the use of three such methods, namely, two single-imputation methods and one based on multiple imputation. These methods are illustrated with a real-world data set, and the results are shown to differ from those obtained from listwise deletion. Key issues guiding the use of these methods are

also discussed. Since these methods have been implemented into statistical software (e.g., SPSS® and SAS®) and should be widely available, their use by educational researchers is expected to substantially increase. Throughout our illustration and discussion, the focus is on handling cases in which partial information is missing rather than a complete lack of information.

The remainder of this chapter is divided into nine sections: (1) Problems Caused by Missing Data and Missing Data Treatments, (2) Missing Data Methods Reported in 11 Education Journals, (3) Missing Data Mechanisms, (4) Overview of Five Ad Hoc Missing Data Methods, (5) Principled Single Imputation—FIML, (6) Principled Single-Imputation—SPSS® EM, (7) Multiple Imputation—SAS® PROC MI, (8) Importance of Statistical Assumptions, and (9) Implications and Recommendations for Educational Research.

PROBLEMS CAUSED BY MISSING DATA AND MISSING DATA TREATMENTS

Why are missing data a problem? The most serious concern is that missing data can introduce bias into estimates derived from a statistical model (Becker & Powers, 2001; Becker & Walstad, 1990; Holt, 1997; Rubin, 1987). For example, it is possible that respondents with incomplete responses might have different response profiles compared to those who responded completely. Thus, the remaining sample is no longer representative of the population from which it was randomly drawn. If the researcher chose to draw conclusions based solely on those who responded, the conclusions could be biased.

Furthermore, missing data result in a loss of information and statistical power (Anderson, Basilevsky, & Hum, 1983; Kim & Curry, 1977). The elimination of subjects with missing information on one or more variables from the statistical analysis in listwise deletion decreases the error degrees of freedom (*df*) in statistical tests such as the *t*. This decrease in turn leads to reduced statistical power and larger standard errors compared to those obtained from complete random samples (Cohen & Cohen, 1983; Cool, 2000).

Similar loss of *df* and statistical power occurs with pairwise deletion, where standard deviations, correlations and covariances are calculated on the basis of available data on each variable or variable pair (Glasser, 1964; Raymond & Roberts, 1987). As a result, the sample composition differs from variable to variable, and the population to which the results are generalized is no longer clearly defined.

Another problem with missing data is that they make common statistical methods inappropriate or difficult to apply (Rubin, 1987). For example, when missing data are present in a factorial analysis of variance, the design is unbalanced. Consequently, the standard statistical analysis that is appropriate for balanced designs is no longer appropriate under this condition. Even if data are assumed to be missing in a completely random fashion, the proper analysis is complicated. Multivariate statistical methods, as they are programmed into commercial statistical software, are applicable to complete data sets by default.

Finally, valuable resources are wasted as a result of missing data. Time and funding spent on subjects who subsequently leave a study or produce missing data represents a loss (Buu, 1999; Holt, 1997). Such loss is a particular concern in longitudinal research, large-scale assessments, high-stake studies, and surveys that ask sensitive information or target respondents who are not accustomed to responding to opinions surveys (such as the first generation Hmong immigrants). Efforts to achieve higher response rates and complete profiles from respondents require researchers to allocate additional time and resources to trace cases who failed to respond or those whose responses were incomplete or unusable. These efforts may not always pay off.

MISSING DATA METHODS REPORTED IN 11 EDUCATION JOURNALS

In order to understand how educational researchers currently deal with missing data, we reviewed the quantitative studies published in 11 education journals from 1998 to 2004. The 11 journals were *American Educational Research Journal* (AERJ), *Educational Researcher* (ER), *Journal of Counseling Psychology* (JCP), *Journal of Educational Psychology* (JEP), *Journal of Research in Science Teaching* (JRST), *Journal of Special Education* (JSE), *Journal of School Psychology* (JSP), *The Modern Language Journal* (MLJ), *Research in Higher Education* (RHE), *Journal for Research in Mathematics Education* (RME), and *Theory and Research in Social Education* (TRSE). These journals were selected because of their emphasis on research, broad coverage of research topics, relevance to subfields in education, and reputable editorial policies. We assumed that the research reported in these 11 journals reflected the mainstream topics and research methods used in educational research.

Within the review period, we identified 1,666 studies in 1,432 articles that met our criteria for quantitative research (Table 3.1). For each study, the methods, findings, and discussion sections (or their equivalents) and all summary tables reported were examined by two of the authors. Special

Table 3.1. Summary of Missing Data Methods Used in Research Published in Selected Education Journals

Journal ^a	Duration	# of Studies	Studies With or Without Missing Data ^b		Missing Data Method Used ^{c,d}										
			Complete	Missing, n	Unknown	LD	PD	MS	HD	RE	EM	MI	Unknown		
AERJ	1998-1999	29/27	2 (7%)	21 (72%)	6 (21%)	19 (90%)	1 (5%)	0	0	0	0	0	1 (5%)	1 (5%)	3 (7.5%)
ER	1998-1999	73/71	19 (26%)	40 (55%)	14 (19%)	27 (67.5%)	6 (15%)	2 (5%)	0	0	1 (2.5%)	1 (2.5%)	3 (7.5%)	0	0
JCP	1998-1999	108/73	40 (37%)	41 (38%)	2 (67%)	1 (33%)	2 (100%)	0	0	0	0	0	1 (50%)	0	0
JCP	2000-2004	3/3	0	2 (67%)	1 (33%)	2 (100%)	0	0	0	0	0	0	1 (50%)	0	0
JEP	1998-1999	147/111	47 (32%)	71 (48%)	29 (20%)	64 (90%)	3 (3%)	0	0	0	0	0	1 (3%)	8 (7%)	0
JEP	2000-2004	224/186	101 (45%)	107 (48%)	16 (7%)	83 (78%)	13 (12%)	5 (12%)	0	0	0	0	1 (50%)	0	0
JEP	1998-1999	147/111	47 (32%)	71 (48%)	29 (20%)	64 (90%)	3 (3%)	0	0	0	0	0	1 (3%)	8 (7%)	0
JEP	2000-2004	409/309	184 (45%)	183 (45%)	42 (10%)	141 (77%)	37 (20%)	0	0	0	0	0	3 (2%)	0	0
JRST	1998-1999	34/33	18 (53%)	15 (45%)	5 (15%)	16 (89%)	0	0	0	0	2 (1%)	0	3 (2%)	0	0
JRST	2000-2004	95/93	37 (39%)	47 (49%)	11 (12%)	22 (47%)	17 (36%)	0	0	0	0	0	2 (11%)	0	0
JSE	1998-1999	21/21	3 (14%)	15 (72%)	3 (14%)	13 (87%)	1 (7%)	1 (7%)	0	0	1 (2%)	0	7 (15%)	0	0
JSE	2000-2004	41/41	11 (27%)	26 (63%)	4 (10%)	15 (58%)	10 (38%)	0	0	0	0	0	1 (4%)	0	0
JSP	1998-1999	29/28	4 (14%)	16 (55%)	9 (31%)	15 (94%)	0	0	0	0	0	0	1 (6%)	0	0
JSP	2000-2004	93/89	27 (29%)	53 (57%)	13 (14%)	39 (74%)	10 (19%)	0	0	0	0	0	1 (6%)	0	0
MLJ	1998-1999	30/30	13 (43%)	13 (43%)	4 (13%)	12 (92%)	1 (8%)	0	0	0	1 (2%)	0	2 (4%)	0	0
MLJ	2000-2004	57/55	26 (46%)	23 (40%)	8 (14%)	14 (61%)	4 (17%)	0	0	0	0	0	5 (22%)	0	0
RHE	1998-1999	56/54	7 (13%)	28 (50%)	21 (37%)	24 (86%)	1 (4%)	0	0	0	0	0	2 (7%)	0	0
RHE	2000-2004	146/143	31 (21%)	76 (52%)	39 (27%)	56 (74%)	13 (17%)	0	0	0	0	0	3 (4%)	0	0
RME	1998-1999	23/19	10 (44%)	6 (26%)	7 (30%)	6 (100%)	0	0	0	0	0	0	0	0	0
RME	2000-2004	29/27	15 (52%)	10 (34%)	4 (14%)	6 (60%)	4 (40%)	0	0	0	0	0	0	0	0
TRSE	1998-1999	4/4	1 (25%)	2 (50%)	1 (25%)	2 (100%)	0	0	0	0	0	0	0	0	0
TRSE	2000-2004	12/12	3 (25%)	6 (50%)	3 (25%)	5 (83%)	1 (17%)	0	0	0	0	0	0	0	0
Total		1666/1432	592 (36%)	806 (48%)	268 (16%)	617 (77%)	129 (16%)	10 (1%)	0	0	4 (<1%)	5 (<1%)	40 (5%)	40 (5%)	0

Notes: Percentages are listed in parentheses. Details about each journal articles may be obtained from the first author. ^aJournal abbreviations: AERJ: *American Educational Research Journal*; ER: *Educational Researcher*; JCP: *Journal of Counseling Psychology*; JEP: *Journal of Educational Psychology*; JRST: *Journal of Research in Science Teaching*; JSE: *Journal of Special Education*; JSP: *Journal of School Psychology*; MLJ: *The Modern Language Journal*; RHE: *Research in Higher Education*; RME: *Journal for Research in Mathematics Education*; TRSE: *Theory and Research in Social Education*. Percentages in parentheses are based on N. ^bPercentages in parentheses are based on n. ^cMissing data method abbreviations: LD: Listwise deletion; PD: Pairwise deletion; MS: Mean substitution; HD: Hot deck; RE: Regression estimation; EM: Expectation-maximization; ML: Maximum-likelihood based; MI: Multiple imputation.

attention was given to the total sample size, the *df* of the test statistics reported, and information regarding authors' treatment of missing data. The unit of analysis was studies, not articles, as many articles presented two or more empirical studies.

Of the 1,666 studies, 592 (or 36%) did not report any missing data problem, 806 (48%) exhibited evidence of missing data, and the remaining 268 (16%) did not provide sufficient information (such as the total sample size, *df* of the statistics, means and percentages, etc.) to determine if missing data were present. Among the 806 studies that showed evidence of missing data (such as the mismatch between the sample size and error degrees of freedom), 766 (95%) explicitly or implicitly reported dealing with such a problem. Of these 766 studies, 617 (80.5%) used the listwise deletion (LD) method and 129 (16.8%) used the pairwise deletion (PD) method. The heavy reliance on listwise or pairwise deletion is probably attributable to the fact that several popular and accessible statistical software programs, such as SPSS® or SAS®, default to listwise or pairwise deletion for handling missing data.

A small number of studies (20 or 2.6%) in the review used mean substitution (MS), regression estimation (RE), the EM method, or the MI method for handling missing data (Table 3.1). A breakdown of the statistics by 1999, when the APA Task Report was published, revealed that only five studies (all published in or after 2000) used the newer, more principled EM method; only one study employed the MI method. None used the other widely recommended full information maximum likelihood (FIML) method. These results suggest that educational researchers have not as yet actively applied principled missing data methods in empirical studies, and refereed journals have not yet encouraged authors to steer away from LD or PD methods.

MISSING DATA MECHANISMS

Missing data occur to varying degrees and in various patterns (Cohen & Cohen, 1983, pp. 275-299). The impact of missing data on the validity of research findings depends on the mechanisms that led to missing data, the pattern of missing data, and the proportion of data missing (Tabachnick & Fidell, 2001, p. 58). Each is discussed below.

It has been shown that the mechanism and the pattern of missing data have greater impact on research results than does the amount of data missing (Tabachnick & Fidell, 2001, p. 58). However, both are critical issues a researcher must address before choosing an appropriate procedure to deal with missing data. According to Little and Rubin (1987), mechanisms that lead to missing data can be classified as: *missing completely at random*, *missing*

at random, and *nonignorable missing*. As defined by Little and Rubin, "if the probability of a response depends on neither the observed nor the missing value that could have been collected or recorded, the missing data are *missing completely at random*" (p. 14, italics in original).

As an example, consider the hypothetical data comprising measurements of 10 subjects on three variables (Y , X_1 , X_2) in Table 3.2A. For purposes of discussion, let us assume that Y represented posttest scores, X_1 pretest scores, and X_2 IQ scores. If the likelihood that Y is missing is unrelated to the missing value itself, nor with X_1 or X_2 , either collected or missing, then Y is said to be *missing completely at random* (abbreviated as MCAR). Under the MCAR condition, missing data can be treated as a random subsample of the potentially complete data, and the missing data mechanism capturing the reasons for missing data can be ignored for sampling-based and likelihood-based inferences (Little & Rubin, 1987, p. 15).

More formally, let Y denote a data vector composed of two parts: those completely observed and those potentially missing. In other words, $Y = (Y_{\text{observed}}, Y_{\text{missing}})$. If the probability of Y being missing does not depend on the missing value itself, but does depend on observed values of Y or other completely observed variables (X 's), then missing data are said to be *missing at random*. The *missing at random* (abbreviated as MAR) assumption states that

$$\begin{aligned} \text{Probability } (Y_{\text{missing}} | Y, X_j, j = 1 \text{ to } k) &= \\ \text{Probability } (Y_{\text{missing}} | Y_{\text{observed}}, X_j, j = 1 \text{ to } k). \end{aligned} \quad (1)$$

Equation 1 implies that the conditional probability of Y being missing, given both Y and X_j is the same as the conditional probability of missing

Table 3.2A. Hypothetical Data of 10 Subjects on Three Variables

Subject	Y (Posttest)	X_1 (Pretest)	X_2 (IQ)
31	22	17	107
32	24	20	100
33		22	110
34	25	11	104
35	24	15	99
36		18	118
37	31	27	122
38	26	25	115
39			90
10	23	19	112

Table 3.2B. Monotone Pattern of the Hypothetical Data of 10 Subjects on Three Variables

Subject	Y (Posttest)	X_1 (Pretest)	X_2 (IQ)
1	22	17	107
2	24	20	100
4	25	11	104
5	24	15	99
7	31	27	122
8	26	25	115
10	23	19	112
3		22	110
6		18	118
9			90

values on Y given observed values of Y and completely observed variables X_j . Put another way, MAR means that the distributions of Y_{missing} and Y_{observed} are the same conditional on a set of predictors X_j . Using the hypothetical data in Table 3.2A, MAR means that any student's missing score on Y (the posttest) could be related to X_1 (pretest) or X_2 (IQ) but not to the missing Y score that could have been collected.

MAR is less restrictive than MCAR; thus, MCAR is said to be a special case of MAR. Under the condition of MAR, the missing mechanism is *ignorable* for likelihood-based inferences (Little & Rubin, 1987, p.15). According to Allison (2001, p. 5):

The missing data mechanism is said to be ignorable if (a) the data are MAR and (b) the parameters that govern the missing data process are unrelated to the parameters to be estimated. Ignorability basically means that there is no need to model the missing data mechanism as part of the estimation process. However, special techniques certainly are needed to utilize the data in an efficient manner. Because it is hard to imagine real-world applications where condition (b) is not satisfied, I [i.e., Allison] treat MAR and ignorability as equivalent conditions in this book. Even in the rare situation where condition (b) is not satisfied, methods that assume ignorability work just fine, but you could do even better by modeling the missing data mechanism.

Based on this logic, MAR and ignorability will be treated as interchangeable in this paper. In contrast to missing at random, the missing data are nonignorable if the probability of missing data depends on the missing values themselves. Again, using the hypothetical data in Table 3.2A, suppose that students missed the posttest (Y) because they believed

they were poorly prepared for the test and their scores were likely to be low. In this case, the missing data on Y would be said to be nonignorable. Unlike the ignorable case, the missing data mechanism must be specified by the researcher and incorporated into the data analysis in order to produce unbiased parameter estimates, a formidable task. Understandably, nonignorable missing data in educational research are particularly likely to occur in studies that seek to gather sensitive or personal information. Unfortunately, no statistical test exists at the present to examine if this condition is met. Pilot testing of the instrument or common sense can sometimes detect this type of missing mechanism. All missing data methods presented in this paper are applicable under either the MCAR or the MAR condition.

The condition of MCAR may be examined using Little's multivariate test, which tests whether the MCAR condition is tenable for the data (Little, Roderick, & Schenker, 1995). Whether the data are consistent with the MAR condition can be examined by a simple t -test of mean differences between the group with complete data and that with missing data (Diggle, Liang, & Zeger, 1994; Kim & Curry, 1977; Tabachnick & Fidell, 2001). Both approaches are illustrated with a data set at <http://www.spss.com/SPSSBI/SPSS/mva/>. Schafer and Graham (2002) criticized the practice of creating a dummy-coded variable that indicates whether a value is present or missing because it redefines the parameters of the population. We caution readers that the results of these tests cannot be interpreted as providing definitive evidence of either MCAR or MAR.

If the pattern of missing data is monotone, then the estimation of parameters in a multivariate distribution can be simplified. A monotone missing data pattern is illustrated in Table 3.2B in which missing data are progressively more prevalent between the pretest and the posttest. For such a monotone missing data pattern, the less restrictive MAR assumption allows for a simplification of the estimation of parameters in a joint, multivariate distribution. Let's illustrate this point with three variables in Table 3.2B. For the joint, multivariate distribution of Y , X_1 , and X_2 , the probability function $f(Y, X_1, X_2)$ can be expressed as a product of the marginal probability distribution of X_1 and X_2 multiplied with the conditional probability distribution of Y given X_1 and X_2 , as in Equation 2:

$$f(Y, X_1, X_2) = f(X_1, X_2) f(Y | X_1, X_2). \quad (2)$$

By the same token, Equation 3 is also true:

$$f(X_1, X_2) = f(X_2) f(X_1 | X_2). \quad (3)$$

Parameters in $f(X_2)$ are estimated using the 10 subjects in Table 3.2B who had complete data on IQ (X_2). Parameters in the conditional distribution of the pretest (X_1) given IQ are estimated from 9 subjects for whom complete data are available. These results can be combined to estimate parameters in the joint distribution of X_1 and X_2 in Equation 3. By the same logic, parameters in the conditional distribution of posttest (Y) given the pretest (X_1) and IQ (X_2) are estimated from 7 subjects with complete information on all three variables. Results from this inference can be combined with results of Equation 3 to make inferences about parameters in the joint probability distribution of all variables, as in Equation 2. If the MAR assumption holds for missing data, the conditional distribution in both equations is often estimated using regression equations, such as regressing the pretest (X_1) on IQ (X_2) in Equation 3 or regressing the posttest (X_3) on both pretest and IQ, as in Equation 2. Thus, the problem of missing data is solved by replacing missing values with imputed scores derived from regression equations. More on this approach is presented in the next section, "Overview of Five Ad Hoc Missing Data Methods."

Regarding the question of how large a proportion of missing data can be tolerated by missing data methods, there are few general guidelines agreed on by statisticians at present. If the percentage of missing values in a random pattern from a large data set (i.e., the MCAR condition holds) is large, for example 20% or higher, there is fairly wide agreement that the problem can be very serious (Cohen & Cohen, 1983; Cool, 2000; Downey & King, 1998; Graham & Schafer, 1999; Tabachnick & Fidell, 2001). For smaller amounts of missing data, there is agreement that the problem is less serious and that different procedures for handling missing data will often produce similar findings; however, the precise cutoff of what constitutes a small amount of missing data varies across authors. For example, Downey and King (1998) used 20% as the cutoff, Brockmeier, Kromrey, and Hogarty (2002) focused on 10%, and Schafer (1997, p. 1) cited 5%. Whatever the precise cutoff, it is clear that smaller percentages of missing data are less likely to produce problems for researchers than larger percentages.

When the missing data are not MCAR, there is general agreement that some missing data methods are superior to others, but little agreement on what percentage of missing data is small enough that its effects can be ignored.

OVERVIEW OF FIVE AD HOC MISSING DATA METHODS

The history of the development of missing data methods can be divided into three periods (Schafer, 1997). In the first period, prior to 1980, most widely applied methods dealing with missing data were ad hoc. These

include LD, PD, mean substitution, simple hot-deck method, and various regression-based methods. They are easy to use, yet typically produce biased results. In the second period, roughly beginning with the publication of Little and Rubin (1987), principled methods, such as the full information maximum likelihood (FIML) and the Expectation-Maximization (EM) algorithm, began to appear. These methods are generally superior to ad hoc methods in that they are statistically efficient and produce parameter estimates with acceptable standard errors. Even though these methods are model-specific and can be difficult to implement, they are viewed as breakthroughs in the history of missing data methods.

The third period in the development of missing data methods began in the late 1980s and early 1990s; it was characterized by the introduction of multiple imputation methods to overcome limitations of single imputation methods. Computer simulation studies have shown this method to be flexible and to yield smaller standard errors than those obtained by other procedures. Even though the multiple imputation method represents the latest effort by methodologists to deal with missing data, the results of the literature review presented in Table 3.1 reveal that it has not been widely adopted by educational researchers.

In this section, we review five ad hoc methods for handling missing data due to their prevalence in statistical software. These include two described earlier (i.e., listwise and pairwise deletion), along with mean substitution, simple hot-deck, and regression. The strengths and weaknesses of each method are discussed in terms of parameter estimation and hypothesis testing, along with evidence available on their performance under realistic data conditions. Most of the evidence comes from computer simulation studies in which data are simulated according to properties specified by researchers, for example, normally-distributed data for a regression model in which particular percentages of values are missing under MCAR or MAR. Different methods for handling missing data are subsequently applied to the simulated data and various statistical indices computed. The difference between these values and known parameter values is recorded and represents bias in estimating parameters. Typically, this process is repeated many times and the bias resulted from the missing data methods studied is averaged. A summary of the features, strengths, and weaknesses of all methods is given in Table 3.3.

Listwise Deletion (LD)

As noted earlier, LD removes subjects that have missing information on one or more variables from the statistical analysis. As Kim and Curry (1977) show, 59% of the data can be lost using LD if only 10% of the data

Table 3.3. Comparison of Eight Methods for Handling Missing Data

Method	Ad Hoc Methods					Principled Methods		
	LD (Listwise Deletion)	PD (Pairwise Deletion)	MS (Mean Substitution)	HD (Simple Hot-Deck)	RE (Regression Estimation)	FIML (Full Information Maximum Likeli- hood)	SPSS' EM (EM = Expectation- Maximization)	MI (Multiple Imputation)
Features	- Discards cases with a missing value - Uses the remaining data to compute results - A default option in most statistical packages - Valid only when data are missing completely at random	- Cases with nonmissing values are used to compute means and variances - Pairs of cases with nonmissing values are used to compute correlations and covariances - Valid only when data are missing completely at random	- Missing values are substituted by means or subgroup means - Statistical analyses are based on the entire data set - Valid when data are missing completely at random	- Missing values are replaced by randomly drawn data already collected in the data set - Complete data for statistical analyses - Valid when data are missing completely at random	- Missing values are replaced from predicted values in a regression equation - The regression equation is formed from observations with complete data - Regression model is specified by researchers - Valid when data are missing at random	- Available data for each subject used to compute a likelihood and to estimate parameters - Requires specification of an imputation model containing variables believed to be predictive of missingness - Subject likelihoods are summed and maximum likelihood is used to obtain parameter estimates for the summed likelihood	- Each iteration consists of two steps: an E-step followed by an M-step - Iteratively computes maximum likelihood estimates for parameters - Iterations continue until the observed log-likelihoods produced in two consecutive iterations are almost identical - Valid when data are missing at random	- Consists of three steps: imputation, analysis and pooling - Takes into account the uncertainty multiple imputations - Solves the missing data problem at the beginning of the analysis - Valid when data are missing at random
Strength	Easy	Easy	Easy	- Easy - Does not distort the marginal distribution	Easy	- Valid when data are multivariate-normal and missing data are missing at random - In AMOS - Efficient - Makes use of all available data - Produces unbiased parameter estimates	- In SPSS MVA - Efficient	- Efficient - Flexible - Programs available from http://www.stat.psu.edu/~jls/ or SAS version 9 or later
Weakness	- Loss of information - Biased - Inefficient	- Loss of information - Varied sample sizes - Inefficient	- Biased - Inefficient	- Distorts correlations and covariances - Inefficient	- Biased - Inefficient	- Model specific or dependent - Failure to specify a reasonable imputation model or to satisfy multivariate-normality can bias estimates	- Time-consuming - Model specific or dependent	- Need to specify an imputation model as well as an analysis model - The third stage—pooling is for parameter estimates and standard errors

were eliminated randomly from each variable in a data set with five variables.

LD is the easiest and most common method for handling missing data. It was used in 77% of the studies published in 11 education journals between 1998 and 2004. This popularity is in large part due to the use of LD as the default setting for multivariate and several univariate statistical procedures in popular statistical packages such as SPSS®, SYSTAT®, and SAS®.

As long as the MCAR assumption holds for missing data, Allison (2001, p. 84) asserted that

among conventional methods for handling missing data, listwise deletion is the least problematic. Although listwise deletion may discard a substantial fraction of the data, there is no reason to expect bias unless the data are not missing completely at random. In addition, the standard errors also should be decent estimates of the true standard errors. Furthermore, if you are estimating a linear regression model, listwise deletion is quite robust to situations where there are missing data on an independent variable and the probability of missingness depends on the value of that variable. If you are estimating a logistic regression model, listwise deletion can tolerate either non-random missingness on the dependent variable or non-random missingness on the independent variables (but not both).

In other words, under MCAR, LD does not generally produce biased estimates and standard statistical procedures are applicable to the remaining data. This result is supported by several computer simulation studies (Brown, 1994; Buck, 1960; Enders, 2001; Raymond & Roberts, 1987; Roth & Switzer, 1995; Wothke, 2000). Still, the reduction in statistical power and precision of estimation typically associated with LD should not be overlooked, nor can they be compensated. An interesting but as yet unexplained finding in this literature is the evidence that LD can be safely used if the correlations among four or fewer variables are low to average and only a small proportion of missing values are present (Buck, 1960; Haitovsky, 1968; Timm, 1970).

If the missing data are MAR, the results of several computer simulation studies indicate that LD typically produces biased parameter estimates and potentially biased statistical tests because results are unrepresentative of the population sampled. Graham and Schafer (2002) reported bias results for a small study in which bivariate normally-distributed data for $n = 50$ cases were simulated. Using real data for guidance, these authors removed 73% of the cases in such a way so as to produce MAR, computed various statistics such as the sample mean, standard deviation, and correlation coefficient, and, in comparing these statistics to known values of the parameters, found substantial bias. Muthen, Kaplan, and Hollis

(1987) examined the impact of LD for data for a structural equation model and reported that, under normality, this missing data method produced biased parameter estimates when MCAR was violated.

Enders and Bandalos (2001) also reported biased parameter estimates using LD, when MCAR did not hold, and simulated data from normal distributions for a structural equation model. Furthermore, the bias increased as the percentage of missing data increased from 2% to 15%. Wothke (2000) reported similar results for growth models using normally-distributed data with large amounts of missing data (80%). Enders (2001) simulated nonnormal data for a structural equation model when MCAR did not hold and found substantial bias in parameter estimates under LD; the bias increased as the percentage of missing data increased from 5% to 25%. Little (1988) simulated normal and nonnormal data with 22% missing, and reported that LD produced biased estimates under MAR. Brockmeier et al. (2002) simulated normally-distributed nonignorable missing data, and reported that LD produced estimates whose bias tended to increase somewhat as the percentage of missing data increased from 10% to 60%.

These results reinforce the conclusion of Wilkinson and The APA Task Force on Statistical Inference (1999) that LD is in general not a satisfactory solution to the missing data problem.

Pairwise Deletion (PD)

PD retains all available data provided by a subject. If this approach is applied to data analysis, descriptive statistics and a few inferential statistics (t -, z -, and chi-square, etc.) are computed from nonmissing data on each variable (Glasser, 1964; Raymond & Robert, 1987). It is the default setting in SPSS®, SYSTAT®, and SAS® for descriptive, correlation, and regression analysis using either correlation or covariance matrices. Though PD is as easy as LD, it is not as widely used as LD. Approximately 16.8% of the studies we reviewed used PD. According to Kim and Curry (1977), PD is most attractive when there is a small number of missing cases on each variable relative to the total sample size, and a large number of variables are involved.

Compared with LD, the PD approach utilizes information obtained from partially complete observations. Its disadvantage is that the sample data change from variable to variable, creating practical problems such as the determination of sample size and degrees of freedom. It is especially problematic for multivariate statistical analyses where solutions and intermediate computations are often based on the entire raw data matrix (Rubin, 1987). Cool (2000, p.7) states "When correlations and other sta-

istics are based on different but overlapping subsamples of a larger sample, the population to which generalization is sought is no longer clear. It is possible to compute correlation matrices with mutually inconsistent correlations." Because of this problem, sample correlation or covariance matrices may not be Gramian (or semipositive definite) (Malhotra, 1987), meaning that solutions obtained from factor analysis, structural equation modeling, or other correlation/covariance based modeling or maximum-likelihood based estimation methods are not valid (Little & Rubin, 1987).

Available evidence of the effects of PD on parameter estimation are more mixed than those for LD. Hainovsky (1968) performed a computer simulation study for a linear regression model for normally-distributed data under MCAR and reported that PD was an acceptable method for handling missing data even when the percentage of missing data was quite large (90%), as long as the correlations among variables were generally low. Arbuckle (1996), Enders (1999, 2001), Enders and Bandalos (2001), Marsh (1998), and Roth and Switzer (1995) reported that PD produced relatively little bias in estimating parameters under MCAR. However, Kaplan (1995) reported that PD produced positively biased chi-square goodness-of-fit tests for normally-distributed data under MCAR. Brown (1994) also reported that PD produced biased estimates under MCAR.

Computer simulation studies by Wothke (2000), Muthen et al. (1987), and Arbuckle (1996) in structural equation modeling reported biased parameter estimates under PD when MCAR did not hold. The results of Enders (2001) for a structural equation modeling problem showed that, when MCAR did not hold, some of the estimated path coefficients were biased under PD for nonnormally-distributed data. Furthermore, the bias increased as the percentage of missing data increased. Switzer, Roth, and Switzer (1998) found similar patterns of bias for PD regardless of whether missing data were MCAR or MAR for a regression model with normally-distributed data and 10% or 20% missing data. Little (1988) reported that PD produced biased estimates under MAR. Brockmeier et al. (2002) reported that for nonignorable missing data, PD produced estimates whose bias tended to increase noticeably as the percentage of missing data increased from 10% to 60%.

These results indicated that PD cannot be counted on to produce unbiased estimates under MCAR. PD may produce biased parameter estimates, a problem that appears to worsen as the percentage of missing data increases. In conjunction with the effects of using different samples of data in the same analysis, PD is not a satisfactory solution to the missing data problem (Wilkinson & Task Force on Statistical Inference, 1999).

Mean Substitution (MS)

The MS approach "solves" the missing data problem by replacing the missing values with the mean of the variable (Wilks, 1932). This step is accomplished at the onset of data analysis. It therefore assumes that the mean of the variable is the best estimate for any observation that has missing information on that variable. In contrast to LD and PD, the MS method does not alter the sample mean of the variable and does not discard any information already collected. Although MS is available in some software programs (e.g., SPSS®), only 1.2% of the studies we surveyed used this method to treat the missing data problem.

A variation of MS is to impute the missing value with a subgroup mean. For example, if the observation with a missing value is a Republican, the mean for all Republicans is computed and inserted in place of the missing value. This procedure is not as conservative as inserting the overall mean of the variable (Tabachnick & Fidell, 2001).

Regardless of which version of MS is applied, this method has many statistical pitfalls. According to Little and Rubin (1987), the limitations include: (a) sample size is overestimated, (b) variance is underestimated, (c) correlations are negatively biased, and (d) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean. The bias introduced into the population variance, correlation, and variable distribution depends on the amount of missing data and on the actual values that are missing. Little and Rubin's recommendation is to never use the MS method.

The theoretically known effects of MS on bias have generally been reproduced in the computer simulation literature. Although Raymond and Roberts (1987) reported less bias for MS than LD when estimating correlations for normally-distributed data under MCAR, bias increased somewhat across the amounts of missing data studied (2%, 6%, 10%). The study by Downey and King (1998) investigated how two mean substitution methods (item mean versus person mean) for missing data collected from Likert scales impacted the recovery of the data and the reliability (Cronbach's alpha) of these scales. These authors used (1) a large sample of respondents ($n = 834$), (b) small numbers of items (15 and 20), and (c) well-established constructs measured by two Likert-scaled instruments. Their results suggested that both methods recovered data quite well as long as the number of respondents with missed items and the number of items with missing responses were less than 20%. As this percentage increased, the person mean substitution method tended to inflate the reliability of the Likert-scaled instrument. King, Fogg, and Downey (1998) reported similar findings for samples of less than 200.

Switzer et al. (1998) found that MS performed poorly for both MCAR and MAR for normally-distributed data, and that bias tended to increase as the percentage of missing data increased from 10% to 20%. This method, as expected, also reduced variation in the data. Timm (1970) and Kim and Curry (1977) reported that MS produced more bias in estimating covariances under MCAR than did LD, and Gleason and Staelin (1975) reported that it produced more bias than PD. Kromrey and Hines (1994) and Brockmeier et al. (2002), in studies of simulated nonignorable regression data, reported that MS produced more bias than LD or PD under MCAR for a wide range of percentages of missing data (10%–60%), and reduced the variances of variables whose missing values were imputed.

Collectively these results suggest that there are virtually no circumstances in which MS should be recommended.

Simple Hot-Deck (HD)

The HD method replaces each missing value with a randomly drawn value from the set of data values already collected on the same variable (Reilly, 1993). None of the studies in our review used this method to impute missing data. Parameters estimated by this method have larger variances, compared to those estimated by MS, but smaller variances than those obtained from complete data. The most serious deficiency of this method is the lack of theoretical work establishing its properties under MCAR or MAR (Roth, 1994). An empirical drawback is the distortion of correlations and covariances, as demonstrated in the results of Switzer et al. (1998), who found that the HD method produced similar bias values for MCAR and non-MCAR conditions, and that the amount of bias was noticeably greater for the larger percentage of missing data (20% versus 10%).

There is little evidence that HD is an appropriate method to deal with missing data, and a conservative course of action would be to avoid this method, especially when correlations or covariances are to be computed for the imputed data (see statements earlier regarding the same problem associated with PD; Little & Rubin, 1987).

Regression Estimation (RE)

The RE method imputes missing values with predicted values derived from a regression equation based on variables in the data set that contain no missing values (Buck, 1960). Variables with missing data are treated as

criterion variables and are predicted by all of the variables having complete data. Four studies reported using this method to handle missing data; they constituted less than 1% of the studies surveyed.

If the missing data exhibit a monotone pattern, as illustrated in Table 3.2B, and are assumed to be MAR—the weaker and more realistic assumption, the RE method can be used to simplify the estimation of population parameters (see earlier comments referring to this method under “Missing Data Mechanism”). Compared to other methods already discussed, the RE method is more informative because it utilizes information already existing in a data set. The AM procedure in BMDP® applies this method to estimate missing values as well as out-of-range data points (BMDP Statistical Software, 1992, pp. 959–976).

Similar to the MS and HD methods, RE has the advantage of preserving cases with missing data and, thus, maintaining the sample size. Disadvantages of RE include: (i) a regression model needs to be specified; (ii) imputed values are always perfectly predicted from the regression model, thus, correlations and covariances are inevitably inflated; (iii) it can be difficult to apply RE to multivariate data sets when more than one variable has missing values; (iv) predicted values may exceed the logical range of scores for the missing data; (v) may require large samples to produce stable estimates (Donner, 1982); (vi) can produce leptokurtic distributions (Rovine, 1994); and (vii) if good and relevant predictors of missing data are not available in the data set, predicted values are no better than the mean. In other words, RE and MS yield approximately the same results if an effective regression model cannot be identified.

To overcome limitation (ii) discussed above, statisticians have suggested a modified RE method in which imputed values have a random error added to them (Beale & Little, 1975). The random error is randomly generated from a normal distribution with a mean of 0 and a standard deviation equal to the square root of the mean square error of the regression model (Little & Rubin, 1987). The Missing Value Analysis module in SPSS® offers users this adjustment.

As stated above, the MS, HD, and RE ad hoc methods “solve” the missing data problem by imputing missing values once, thus they are referred to as single imputation methods. The single imputation approach unfortunately does not reflect the uncertainty in missing data estimates. That is, the error term in the estimation equation, of whatever form, used to impute missing values is set to zero. Furthermore, the sample size is overstated, confidence intervals for estimated parameters are too narrow, and Type I error rates are too high (Little & Rubin, 1987).

Available empirical evidence of the performance of RE indicates that it is usually, but not always, superior to LD, PD, or MS. Buck (1960) examined its performance by starting with a real dataset of 72 samples, ran-

domly removing 40% of the samples, and using RE to impute missing values. The bias associated with parameter estimates, compared to the values produced by the original 72 samples, was generally less than that produced by LD. Buck's study is not a simulation study in the usual sense, but his results are consistent with those produced by simulation studies in which computer algorithms simulated data.

Chan and Dunn (1972) performed a small simulation study in discriminant analysis with normally-distributed data, and reported that RE showed approximately the same bias as LD and PD under MCAR. Beale and Little (1974) reported that RE showed less bias than LD for normally-distributed data that were simulated for a regression analysis with the percentage of missing values of 5%, 10%, 20%, or 40% under MCAR. Timm (1970) also reported RE to show less bias than LD under MCAR. Raymond and Roberts (1987) reported that RE showed less bias than either LD or MS under MCAR. Similarly, Gleason and Staelin (1975)'s study found RE to produce less bias than MS under MCAR in a simulation study in which the percentage of missing data was fixed at 20%. Switzer et al. (1998) found that MS produced similar levels of bias under the MCAR and non-MCAR conditions, and that the amount of bias was noticeably greater for 20% missing versus 10%. Brockmeier et al. (2002) reported that RE produced noticeably less bias for nonignorable data than LD, PD, or MS.

Principled Single Imputation—FIML

FIML stands for Full Information Maximum Likelihood; it represents a principled method for estimating means and covariances based on incomplete data when the missing values are assumed to be MAR. In maximum likelihood (ML), parameter estimates are derived such that the likelihood of reproducing the data, given the parameter estimates, is maximized. The FIML method for estimating parameters in the presence of data missing at random makes extensive use of ML.

FIML has its roots in the work of Hartley and Hocking (1971). Given q groups, one for each pattern of missing data, FIML first calculates the likelihood for each of the q groups or patterns. If there are no missing data, then $q = 1$. The intent of FIML is to use the information in each missing data pattern to estimate parameters. Under the assumption of a multivariate normal population distribution, the q likelihoods are summed. The resulting summed likelihood serves as the basis for finding parameter estimates using ML. FIML is conceptually similar to the q -group method of Hartley and Hocking, except that the likelihood is cal-

culated for each case, using whatever data are available for that observation.

To illustrate how FIML estimates parameters in the presence of missing data, let's suppose that data are collected for N cases on three variables Y , X_1 , and X_2 , and the researcher's interest is in estimating the population means, variances, and covariances, denoted below in the vector μ and the matrix Σ :

$$\begin{aligned}\mu &= [\mu_1, \mu_2, \mu_3] \\ \Sigma &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}\end{aligned}\quad (4)$$

It is further assumed that X_2 is missing for some cases. FIML computes a likelihood function for each case using variables specified by the researcher, say, in a regression equation that were observed for that case. Under the assumption of multivariate normality, the case-wise likelihood functions are summed across the entire sample and the summed likelihood is maximized. If case #1 has no data for X_2 , it contributes only to the estimation of μ_1 , μ_3 , σ_{13} , σ_{31} , and σ_{33} . After N likelihoods have been computed, they are summed and ML is used to estimate the means, variances, and covariances based on the summed likelihood. These parameter estimates are unbiased and efficient in the presence of missing data under MAR and multivariate normality assumptions.

There is far less computer simulation evidence of the performance of FIML estimators than those already summarized. What is published on FIML estimators has generally suggested that this method produces unbiased estimators under more conditions than other ad hoc methods. Arbuckle (1996) examined the bias of FIML estimates for 30% missing data in a structural equation modeling problem under the multivariate normal condition. He reported that it produced less bias than both LD and PD under MCAR or MAR. Wothke (2000) reported that FIML estimates for a latent growth curve model were unbiased for both MCAR and MAR conditions, while LD and PD produced biased estimates under MAR. Enders (1999) also reported that FIML produced unbiased estimates under both MCAR and MAR for several missing data percentages (2%, 5%, 10%, 15%, 25%) in a structural equation modeling problem for normally-distributed data.

In sum, the missing data literature provides substantial guidance for choosing an appropriate method to handle missing data. In the unlikely case that the missing data are MCAR, the LD, RE, and FIML estimators

generally produce unbiased estimates, although large amounts of missing data (e.g., 60%) may require large sample sizes for unbiased estimation, for example, 500 (Arbuckle, 1996). As always, the use of LD to produce unbiased estimates under MCAR must be weighed against the potentially negative consequences of reduced statistical power and precision. Under the MAR condition, RE appears to be superior to LD or PD, and research results published so far suggest that FIML is superior to RE. This conclusion appears to hold across varying amounts of missing data.

ILLUSTRATION OF FIML USING THE AMOS SOFTWARE

In this section, we illustrate how different methods of handling missing data, FIML versus LD, can yield different parameter estimates and, potentially, different inferences. To illustrate these effects, we used a data set from 1,302 U.S. colleges and universities for the 1993-1994 school year.

Data Set

The data set was provided by U.S. News & World Report and is available at <http://portfolio.iu.edu/peng/articles/persist.sav>. For this data set, hereafter referred to as the *Persist* data, we were interested in modeling variability in student persistence, as measured by graduation rates, as a function of college-related predictors, such as tuition and institutional quality. Because the cases were colleges and universities, each institution's graduation rate (*GRADRATE*), in percentages, was the unit of analysis and served as a proxy for student persistence.

Three latent variables, often described in the literature as predictors of student persistence, appeared in the path model: student's ability, quality of institutions, and costs. These were manifested in financial cost (room and board or *RMBRD*), additional student fees (*ADDFEES*), in-state tuition (*INSTATE*), cost of books (*BOOKCOST*), selectivity (or quality) of the institution (*SELECT*), percent of faculty with PhD's (*PCTPHD*), student to faculty ratio (*STUDFACT*), percent of full time students (*PCTFULL*), expenditure per student (*PERSPEND*), and student quality and preparedness, measured by the average SAT Math and Verbal scores (*AVESAT*). The amount of missing data for these variables ranged from 0.2% to 40.3%, with the percentage of missing data for graduation rate (*GRADRATE*) equal to 7.5%. All variables thought to contribute to the missingness were included in FIML, even if some of them were not included in subsequent analyses (Schafer, 1997).

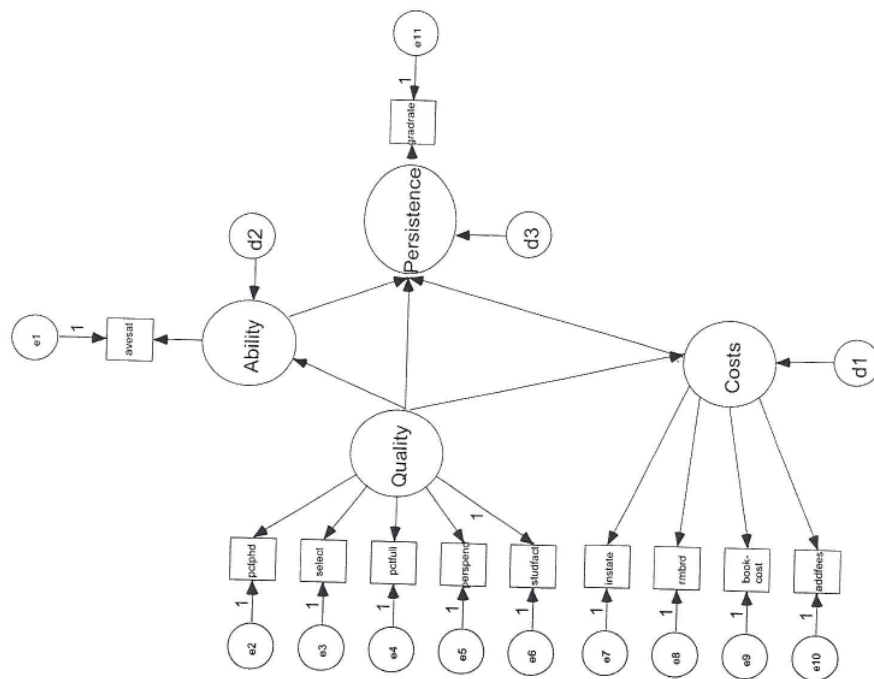


Figure 3.1. Path model fitted to graduation data.

Statistical Modeling

The AMOS software (Arbuckle, 1995) was used to fit the path model in Figure 3.1 to the *Persist* data. Both FIML and LD methods were used to handle missing data. For each method, we examined its effect on model conclusions based on estimated path coefficients and their standard errors. We did not attempt to find the best-fitting model under each method for two reasons: first, the best model might be different; second, different best models might be attributable to the vagaries of finding

Table 3.4. Estimates and Standard Errors for Path Coefficients for LD and FIML

Missing Data Method		LD		FIML	
Paths		Estimate	SE	Estimate	SE
QUALITY → ABILITY		88.87	15.36	155.78	33.71
QUALITY → COSTS		10.05	8.20	14.99	8.97
COSTS → PERSIST		-1.44	1.39	-0.57	0.32
ABILITY → PERSIST		0.06	0.01	0.01	0.009
QUALITY → PERSIST		3.43	1.09	-3.48	1.23
QUALITY → %FACULTY PhD		15.94	3.26	18.43	3.49
ABILITY → AVERAGE TOTAL SAT		1.00	—	1.00	—
QUALITY → STUDENT to FACULTY RATIO		1.00	—	1.00	—
QUALITY → AVERAGE EXPENDITURE		13.71	23.68	83.61	33.29
QUALITY → %FULLTIME		13.97	3.18	22.38	4.27
QUALITY → No. Admitted /No. Applied		0.06	0.01	-0.14	0.04
QUALITY → ADDITIONAL FEES		1.00	—	1.00	—
COSTS → COST BOOKS		2.17	1.55	1.69	0.99
COSTS → ROOM and BOARD		96.34	42.66	44.52	25.22
COSTS → OUT of STATE TUITION		38.2	76.12	152.83	86.56
PERSIST → GRADUATION RATE		1.00	—	1.00	—

Notes: Quality → Ability reflects the direct effect (path) from institutional quality to student ability; LD = listwise deletion; FIML = full information maximum likelihood; estimates of 1.00 were specified as part of the model.

these models as well as to intrinsic properties of the missing data methods. Because of these reasons, the path model was fit once by each method to the data. Table 3.4 presents the estimates and standard errors of the path coefficients using both FIML and LD.

COMPARISON OF AMOS FIML RESULTS WITH THE LD RESULTS

The results in Table 3.4 show several differences in the estimated path coefficients and their standard errors as a function of the missing data method used. FIML produced larger estimated path coefficients for eight out of twelve paths. The size of the standard errors for the two methods was equally divided, with FIML standard errors larger than LD-based val-

ues in six out of twelve paths. In two paths, the signs were reversed between the FIML- and the LD-estimated coefficients. For example, the estimated path coefficient of QUALITY on student's graduation rate (GRADRATE) using FIML was -3.48, whereas the LD-based coefficient was 3.43. The FIML coefficient (-3.48) means that an increase in institutional quality was associated with a decrease in student graduation rates, with other effects (or paths) held constant. The LD coefficient (3.43) suggested that increasing institutional quality was associated with higher student graduation rates. These results provide empirical evidence that the choice of method for handling missing data can substantially affect inferences.

COMMENTS ON FIML

Though FIML has been implemented in the AMOS software (Arbuckle, 1999), the study of FIML, as suggested above, has been primarily in structural equation modeling. Results from these studies indicate that, as expected, parameters estimated using FIML are unbiased and efficient when the data are multivariate-normal and the MAR assumption holds. However, failure to satisfy the multivariate normality assumption can dramatically bias FIML estimates (Enders, 2001).

Another drawback of FIML is that tests of goodness-of-fit are not always available. For example, in regression modeling with missing data on the dependent variable Y , the estimates of regression coefficients are unbiased and efficient under the assumption of MAR and normality. However, a test of the overall model fit to data is not possible because there is no value of N applicable to the entire sample. Similar problems with FIML emerge in structural equation modeling in which the goodness-of-fit test and several commonly prescribed measures of fit are not available. Another potential problem is that the covariance matrix may be indefinite, which can lead to significant parameter estimation difficulties, although these problems are often modest (Wothke, 2000).

A practical issue with using FIML in the presence of missing data is that variables believed to be predictive of missingness must be in the regression or structural equation model, even when these variables are not of substantive interest. In some instances, the presence of such variables may have an important impact on variables of substantive interest in the model.

PRINCIPLED SINGLE IMPUTATION—SPSS® EM

EM stands for Expectation-Maximization; it is a principled method for handling missing data. Dempster, Laird, and Rubin in their seminal

paper on this method coined the term EM (1977). According to Little and Rubin (1987),

The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters, and so forth, iterating until convergence. (p. 129)

Each iteration of the EM algorithm consists of two steps: an E (Expectation) step followed by an M (Maximization) step. In the E step, the expectation of the complete data log-likelihood is derived, given the observed data and the estimated parameters from a previous iteration. In the M step, the conditional expectation of the complete data log-likelihood is maximized. The observed data log-likelihood is increased until a stationary point is reached (Dempster et al., 1977). In other words, the algorithm continues until the observed likelihoods produced in two consecutive iterations are almost identical. Only five studies (less than 1%) we reviewed used the EM method to treat missing data problem.

The Missing Value Analysis (MVA) module available in SPSS® version 10 and beyond implements the EM algorithm indirectly as a single imputation method. First, it employs the EM algorithm to derive ML estimates of parameters based on a researcher's specification of a probability model for the data (e.g., multivariate-normal). Operationally, this involves using a series of regressions for each missing data pattern similar to Buck's RE method (Buck, 1960).

After the EM algorithm converges, MVA computes imputed values for missing data based on means and covariances estimated from the last iteration of the EM algorithm. The imputed values replace missing data to yield a complete data set, which is subsequently analyzed by the statistical method(s) of a researcher's choice. We consider the EM algorithm used in MVA to be a principled, single-imputation method for handling missing data.

Empirical evidence available thus far on the EM approach suggests that it performs well under a variety of conditions. Beale and Little (1975) reported that, for normally-distributed data, the EM approach showed bias similar to that produced by LD and RE; the bias was not affected by the amount of missing data (19%, 20%, 40%) under MCAR. Donner and Rosner (1982) simulated normally-distributed data for a regression model under MCAR. Their results revealed that parameter estimates obtained with EM showed less bias than those with LD, PD, or RE. Raymond and Roberts (1987) reported results similar to those of Donner and Rosner. Little (1988) showed that, under MAR, the EM approach yielded less bias in estimating means, correlations, and regression coefficients, but pro-

duced biased standard deviations. Brockmeier et al. (2002) reported that, for nonignorable data, the EM approach produced estimates with noticeably less bias than LD, PD, or MS, but about the same amount of bias as RE.

On the other hand, Basilevsky, Sabourin, Hum, and Anderson (1985) reported that the EM approach produced biased estimates similar in magnitude to those obtained with LD and PD. In general, though, the literature suggests that the EM approach is superior to ad hoc methods across different amounts of missing data.

ILLUSTRATION OF SPSS® EM METHOD

In this section, we describe how the SPSS® EM Method was applied to impute values for missing data for a multivariate data set. This was followed by fitting a logistic regression model to help predict adolescent's behavioral risk. Results are presented in terms of regression coefficient estimates and standard errors.

Data Set

Self-reported health behavior data were collected from 517 adolescents enrolled in two junior high schools (Grades 7 through 9) in the fall of 1988. On the day of data collection, two questionnaires were administered: the Health Behavior Questionnaire (HBQ) (Ingersoll & Orr, 1989; Resnick, Harris, & Blum, 1993) and Rosenberg's self-esteem inventory (Rosenberg, 1965). Among the 517 students, 85 did not complete all questions. Thus, the sample size with complete information was 432 (83.4% were Whites and the remaining Blacks or others), with a mean age of 13.9 years and nearly even numbers of girls ($n = 208$) and boys ($n = 224$).

The HBQ asked adolescents to indicate whether they engaged in specific risky health behaviors (Behavioral Risk Scale) or had experienced selected emotions (Emotional Risk Scale). The response scale ranged from 1 (*never*) to 4 (*about once a week*) for both scales. Cronbach's alpha reliability was 0.84 for the Behavioral Risk Scale and 0.81 for the Emotional Risk Scale (Peng & Nichols, 2003). Adolescents' self-esteem was assessed using Rosenberg's self-esteem inventory (Rosenberg, 1965). Self-esteem scores ranged from 9.79 to 73.87 with a mean of 49.97 and standard deviation of 10.09. Furthermore, among the 432 adolescents, 12.27% (or 53) indicated an intention to drop out of school; 44.68% (or 193) were from intact families, 22.69% (or 98) were from families with one

step-parent, and 32.63% (or 141) were from families headed by a single parent. The data set is hereafter referred to as the *Adolescent* data [available from [http://portfolio.iu.edu/peng/articles/logregdata\(peng\).sav](http://portfolio.iu.edu/peng/articles/logregdata(peng).sav) as an SPSS® data file or from <http://portfolio.iu.edu/peng/articles/logregdata.por> as a portable data file].

Research Question and Statistical Modeling

For the Adolescent data, we were interested in identifying adolescents at the greatest behavioral risk from their gender, intention to drop out from school, family characteristics, emotional risks, and self-esteem scores. Given this objective, the research hypothesis was stated as follows: "The likelihood that an adolescent is at high behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self-esteem." Scores on the Behavioral Risk Scale of the HBQ ranged from 40.44 to 95.21 with a mean of 47.23 and a standard deviation of 8.40, and it was decided that adolescents scoring 52 or above (i.e., more than a half of SD above the mean) were identified as being at high behavioral risk, and those with scores below 52 were not at high risk.

A logistic regression model was fit to the data using adolescents' risk level on the Behavioral Risk Scale of the HBQ as the dependent variable and gender, intention to drop out of school, type of family structure, emotional risk, and self-esteem score as covariates or the explanatory variables.

SPSS® EM Method for Imputing Missing Data

As stated earlier, out of the 517 students, 85 did not complete all questions. Most missing data occurred on Behavioral Risk Scale (77 cases) or Emotional Risk Scale (34 cases); only six adolescents did not indicate if they intended to drop out from school. To impute values for missing data, the entire data set was submitted to the MVA module in SPSS® version 11.01. The imputation model specified included gender, intention to drop from school, behavioral risk score, emotional risk score, self-esteem score, family configuration (intact family, step family, or single parent family), and living arrangement (living with both parents, father alone, mother alone, neither, father and step-mother, or mother and step-mother). These variables were predictive of the missingness for the Adolescent data (Peng & Nichols, 2003).

Once the imputation model was defined in MVA and the EM estimation method was selected, we further specified a multivariate normal dis-

tribution for all variables selected and 50 iterations for the EM iteration. The result was a complete data set with missing data replaced by imputed values obtained from the last iteration of the EM algorithm. This complete data set was subsequently analyzed by the logistic regression procedure in SAS® version 8.2.

Analysis Results

Using the criterion of 52 points on the Behavioral Risk Scale of the HBQ, 102 adolescents were identified to be at high behavioral risk while 415 were not. To entertain the research hypothesis that "the likelihood that an adolescent is at high behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self-esteem," a logistic regression model was fit to the data that yielded the following result:

$$\begin{aligned} \text{Predicted logit } (Y = \text{high behavioral risk}) = & -1.6120 + \\ & (1.1014) * \text{GENDER} + (2.5369) * \text{DROPOUT} + (0.4084) * \\ & \text{FAMILY} + (0.00809) * \text{EMOTION} + (-0.0443) * \text{ESTEEM}. \end{aligned} \quad (5)$$

According to this model, the log of the odds of an adolescent being at high behavioral risk was positively related to gender ($p < .0001$, Table 3.5), intention to drop out of school ($p < .0001$), and family structure ($p < .01$); it was negatively related to self-esteem ($p < .01$), and insignificantly related to emotional risk ($p = 0.5405$). The effectiveness of Equation 5 in explaining the data was examined using (a) overall model evaluations, (b) statistical tests of each covariate, (c) goodness-of-fit statistics, and (d) measures of association. These indices were recommended in the literature for evaluating logistic regression results (Peng, Lee, & Ingersoll, 2002; Peng, So, Stage, & St. John, 2002); and they are presented in Table 3.5.

COMPARISON OF SPSS® EM RESULTS WITH THE LD RESULTS

The same logistic regression model was fitted to 432 cases with complete data. This approach was equivalent to treating the missing data with listwise deletion. Results are presented in Table 3.6. According to Table 3.6, the LD approach led to the following logistic model:

Table 3.5. Logistic Regression Analysis of Adolescent's Self-Inflicting Behavior Risk by SAS® PROC LOGISTIC (version 8.2), $n = 517$

Predictor	β	SE β	Wald's χ^2 (df=1)	p	e^{β} (Odds Ratio)
CONSTANT	-1.6120	1.3165	1.4991	.2208	Not necessary
GENDER (boys = 1, girls = 0)	1.1014	0.2777	15.7291	<.0001	3.0083
DROPOUT (yes = 1, no = 0)	2.5369	0.3221	62.0297	<.0001	12.6404
FAMILY	0.4084	0.1505	7.3601	.0067	1.5044
EMOTION	0.0081	0.0143	0.3198	.5717	1.0081
ESTEEM	-0.044	0.0155	8.1678	.0043	0.9570

Overall Model Evaluation			
Tests	χ^2	df	p
Likelihood Ratio Test	122.04	5	<.0001
Score Test	135.07	5	<.0001
Wald Test	88.03	5	<.0001
Hosmer & Lemeshow Goodness-of-fit Test	5.54	8	.6865

Notes: Cox and Snell $R^2 = 0.2124$. Nagelkerke R^2 (Max rescaled R^2) = 0.3361. Kendall's Tau-a = 0.200. Goodman-Kruskal's Gamma = 0.627. Somers' D_{xy} = 0.625. c -statistic = 0.813.

$$\text{Predicted logit } (Y = \text{high behavioral risk}) = -1.6341 + (1.0470) * \text{GENDER} + (2.0772) * \text{DROPOUT} + (0.4628) * \text{FAMILY} + (0.0065) * \text{EMOTION} + (-0.040) * \text{ESTEEM}. \quad (6)$$

According to Equation 6, the log of the odds of an adolescent being at high behavioral risk was positively related to gender ($p < .0002$, Table 3.6), intention to drop out of school ($p < .0001$), and family structure ($p < .003$); it was negatively related to self-esteem ($p < .02$), and was not related to emotional risk ($p = .6459$). Essentially, Equation 6 uncovered the same relationship as Equation 5 did between the likelihood of high behavioral risk and the five covariates. The standard errors of the estimated regression coefficients in Tables 3.5 and 3.6 were virtually identical.

However, there were noticeable differences between these two results. First, the Wald's chi-square test of the *ESTEEM* coefficient was significant at an alpha level of 0.01 in Table 3.5 yet it was significant at an alpha level of 0.05 in Table 3.6. Second, the measures of association shown at the bot-

Table 3.6. Logistic Regression Analysis of Adolescent's Self-Inflicting Behavior Risk by SAS® PROC LOGISTIC (version 8.2), $n = 432$

Predictor	β	SE β	Wald's χ^2 (df=1)	p	e^{β} (Odds Ratio)
CONSTANT	-1.6341	1.3001	1.5797	.2088	Not necessary
GENDER (boys = 1, girls = 0)	1.0470	0.2818	13.8058	.0002	2.8491
DROPOUT (yes=1, no=0)	2.0772	0.3368	38.0385	<.0001	7.9821
FAMILY	0.4628	0.1546	8.9597	.0028	1.0648
EMOTION	0.0065	0.0142	0.2111	.6459	1.0065
ESTEEM	-0.040	0.0155	6.5224	.0107	0.9608

Overall Model Evaluation

Tests	χ^2	df	p
Likelihood Ratio Test	80.89	5	<.0001
Score Test	86.24	5	<.0001
Wald Test	63.22	5	<.0001
Hosmer & Lemeshow Goodness-of-fit Test	5.23	8	.7322

Notes: Cox and Snell $R^2 = 0.1708$. Nagelkerke R^2 (Max rescaled R^2) = 0.2675. Kendall's Tau-a = 0.189. Goodman-Kruskal's Gamma = 0.578. Somers' D_{xy} = 0.576. c -statistic = 0.788.

tom of Table 3.6 were smaller than those of Table 3.5. These differences speak to the consequences of decreasing sample size and lowering statistical power that resulted from applying the LD method.

COMMENTS ON EM AND SPSS® EM METHOD

Applications of EM-type algorithms have a long history (McLachlan & Krishnan, 1997, pp. 34-37). The popularity of the EM algorithm among statisticians is largely based on the fact that this approach allows many complex statistical problems to be reformulated as a missing data problem in a way that greatly simplifies parameter estimation (e.g., mixture models, random effects models, hierarchical linear models, unbalanced designs including repeated measures). In most applications, it is assumed that data follow a multivariate normal distribution.

In addition to taking the missing data mechanism into account, the primary advantages of the EM algorithm are simplicity and ease of com-

puting (Dempster et al., 1977; Little & Rubin, 1987). Since fewer than 1% of the studies we reviewed reported using the EM method to handle missing data, we suspected that the low frequency of usage might be partly attributable to its drawbacks. First, the M-step of the algorithm does not produce sampling variances of the parameter estimates. Several extensions of the EM algorithm have appeared in the statistical literature that estimate standard errors (Meng & Rubin, 1992; McLachlan & Krishnan, 1997, pp. 28-29). But in general these methods have not been implemented in missing data programs such as the MVA module in SPSS®. Second, all likelihood-based methods including the EM method are iterative and model-specific (or model-dependent). Because of their dependency on specific models, these methods are not available in all statistical procedures in a general-purpose statistical package, such as the SPSS® regression or the SAS® PROC GLM. Third, the convergence rate during iterations is proportional to the percent of completely observed data in a data set (Little & Rubin, 1987, p. 130). As a result, the rate of convergence can be painfully slow if the percentage of missing data is high (Buu, 1999; Little & Rubin, 1987).

Buu (1999) investigated the Type I error rate control and the precision of parameter estimation of three missing data methods: the SPSS® EM approach (i.e., MVA), the direct, one-step EM algorithm, and SAS® PROC MIXED. SAS® PROC MIXED uses either the Newton-Raphson algorithm (the default) or the Fisher scoring algorithm (when the Newton-Raphson algorithm fails to converge) to find the maximum likelihood estimates of parameters. While EM accomplishes this by iterating between the E (expectation) step and the M (maximization) step, the Newton-Raphson and the Fisher scoring algorithms do so by iteratively estimating the parameters via the computation of second derivatives of the observed data log-likelihood function (Jennrich & Schluchter, 1986; Littell, Milliken, Stroup, & Wolfinger, 1996). Buu's study was conducted in longitudinal studies using simulated and empirical data. Her results revealed that the SPSS® MVA approach rejected the null hypothesis more frequently than it should, resulting in inflated Type I error rates, though it estimated parameters accurately. These findings were obtained in repeated-measures ANOVA using both empirical data and data simulated from multivariate normal distributions. The direct, one-step EM and SAS® PROC MIXED yielded very similar and correct Type I error rates with SAS® PROC MIXED slightly outperforming the EM algorithm. Buu's study is unique in the literature as most missing data methods were evaluated in between-subjects designs rather than within-subjects or time-series studies (Roth, 1994).

MULTIPLE IMPUTATION—SAS® PROC MI

To overcome the limitations of methods that fail to take into account the uncertainty associated with imputed values, Rubin and his associates developed the multiple imputation (MI) method in the 1980s. MI is a valid method for handling missing data under the MAR condition. It is a general-purpose method, highly efficient even for small sample sizes (Graham & Schafer, 1999). Merely one study we reviewed used this method. For an introduction to MI, we recommend Schafer's 1999 paper (Schafer, 1999). MI generally consists of three steps: imputation, analysis, and pooling. In the first (imputation) step, each missing value is replaced by not one, but $m > 1$ simulated values. Imputed values are drawn from a distribution that is specified by the researcher. At the end of the imputation step, m complete data sets are created. In the second (analysis) step, each of the m complete data sets is analyzed by standard complete-data methods. Finally, results of the m analyses are integrated in the third (pooling) step to yield a final result such as an interval estimation of a population parameter, a p -value of null-hypothesis testing, or a likelihood-ratio test statistic. In the next section, we elaborate on each of these three steps applying PROC MI and PROC MIANALYZE in SAS® version 8.2 (SAS Institute, 1999) to the Adolescent data set described earlier.

Illustration of the MI Method

The research question posed to the Adolescent data set and the logistic modeling approach are described in the section titled, "Illustration of SPSS® EM Method."

Step 1—Imputation

In this step, each missing value is replaced by $m > 1$ imputed values. The model we adopted for imputation was the same logistic regression model that would be used subsequently to analyze the data, though these two models need not be identical (see the section, "Comments on the MI Method" and Schafer, 1997, pp. 139-143). The data matrix consisting of one dependent variable (the behavioral risk score) and five covariates was submitted to PROC MI in SAS® version 8.2 to impute missing data 5 times (the default). Since the missing rate for the current data was 16.44%, $m = 5$ imputations were considered sufficient (Graham & Schafer, 1999). The entire SAS® program is available from <http://portfolio.iu.edu/peng/articles/tr-em.fit.sas>. The syntax for PROC MI was as follows; the upper case commands are SAS® keywords and lower cases are user-specified variable names or SAS® data set names.


```
PROC MI DATA=risk SEED=37851 OUT=outmirisk ROUND=111111;
VAR behrisk gender dropout emotion esteem family;
```

The keyword, SEED, specifies a starting value for the random number generator whereas ROUND specifies units of rounding for imputed values. To revalidate the imputation results, a researcher can reset SEED to a different number that will result in a different set of initial estimates for parameters (i.e. regression coefficients for the Adolescent data set).

PROC MI assumes that missing data were missing at random (the MAR condition) and that the population data distribution is multivariate-normal. Both assumptions were deemed reasonable for the present data (see the section titled "Importance of Statistical Assumptions for Principled Method") on the importance of satisfying these two assumptions). At the end of the imputation, 5 data sets with $N = 517$ cases were created. Table 3.7 summarizes the increase in variability as a result of multiple imputations for the three variables that had missing data, i.e., dropout, behavioral risk, and emotional risk. The increase in variability reflected the uncertainty in the imputed values—an uncertainty not accounted for by single imputation methods. Thus, variables associated with larger amount of missing data (e.g., behavioral risk) tend to have greater uncertainty (or increase in variance) in estimated parameters than those associated with smaller amounts of missing data (e.g., dropout and emotion risk).

Step 2—Analysis

In the second (analysis) step, each of the m complete data sets is analyzed by a standard complete-data method, such as the logistic model prescribed for the Adolescent data. Consequently, each data set obtained from Step 1 was analyzed by PROC LOGISTIC in SAS® version 8.2 to determine if and how the likelihood that an adolescent was high at behavioral risk was related to his/her gender, intention to drop out of school, family structure, emotional risk, and self-esteem. The SAS® syntax for PROC LOGISTIC for these five data sets was as follows:

```
PROC LOGISTIC DATA=outmirisk;
MODEL risk=gender dropout family emotion esteem/
RSQUARE LACKFIT;
FORMAT risk br.;
BY _imputation_;
```

Again, the upper case commands are SAS® keywords and lower cases are user-specified variable names or SAS® data set names. The FORMAT statement applied labels (such as "high" and "low") to numerical codes (such as "1" and "2") of the dependent variable, *risk*. The BY statement applied the logistic regression modeling repeatedly to all imputed data

Table 3.7. Relative Increase in Variance and Fraction of Missing Information of Missing Information in the Adolescent Data Set

Variable	Relative Increase in Variance	Fraction of Missing Information
Dropout	0.0270	0.0267
Behavioral Risk	0.3195	0.2634
Emotional Risk	0.0435	0.0425

sets. The variable name, `_imputation_`, was created in the preceding PROC MI to denote each imputed (complete) data set. At the end of Step 2, five logistic regression results, similar to those presented in Tables 3.5 or 3.6, were obtained.

Step 3—Pooling

The last stage in MI is to pool results of the m analyses to yield a final result, using a formula available in Rubin (1987) and Schafer (1997). To accomplish this step, we invoked PROC MIANALYZE in SAS® 8.2 as follows: [the entire SAS® program is available from [http://portfolio.iu.edu/peng/articles/mianalyze-lr.fit\(5var\).sas](http://portfolio.iu.edu/peng/articles/mianalyze-lr.fit(5var).sas)]

```
PROC MIANALYZE DATA=outest;
VAR intercept sex drop famstruc emorisk selfest;
RUN;
```

In the program above, the keyword DATA= specifies a covariance matrix of parameter estimates (i.e., the intercept and logistic regression coefficients) obtained from PROC LOGISTIC in Step 2 above. From this covariance matrix, PROC MIANALYZE was able to yield a single (pooled) estimate of six logistic regression estimates (Table 3.8).

As shown in Table 3.8, the log of the odds of an adolescent being at high behavioral risk was linearly related to five covariates according to the following function:

$$\begin{aligned} \text{Predicted logit (Y=high behavioral RISK)} = & -1.9398 + \\ & (1.1860) * \text{GENDER} + (2.0751) * \text{DROPOUT} + (0.4343) * \\ & \text{FAMILY} + (0.0090) * \text{EMOTION} + (-0.035) * \text{ESTEEM}. \end{aligned} \quad (7)$$

The interpretation of the significance of the overall model and each covariate was similar to that of Equation 5. In other words, the predicted logit was positively related to gender ($p < .002$), intention to drop out of school ($p < .0001$), and family structure ($p < .003$); it was negatively related to self-esteem ($p < .02$), and not statistically significantly related to

Table 3.8. Pooled Logistic Regression Analysis Results of Adolescent's Self-Inflicting Behavioral Risk by SAS® PROC MI and PROC MIANALYZE (version 8.2), $n = 517$

Predictor	b	SE β	t -test (df_m) ^a	p	95% Confidence Lower Limit	95% Confidence Upper Limit
CONSTANT	-1.9398	1.2052	-1.61 (12241)	.1075	-4.3021	0.4226
GENDER (boys = 1, girls = 0)	1.1860	0.3305	3.59 (26,025)	.0014	0.5068	1.8654
DROPOUT (yes = 1, no = 0)	2.0751	0.3215	6.45 (616.45)	<.0001	1.4436	2.7065
FAMILY	0.4343	0.1450	3.00 (463.71)	.0029	0.1495	0.7192
EMOTION	0.0090	0.0131	0.69 (37846)	.4884	-0.0165	0.0346
ESTEEM	-0.035	0.0142	-2.45 (4922.2)	.0142	-0.0627	-0.0070

^aThe definition of df_m is given in Equation (12).

emotional risk ($p = .4884$). The comparison of this model to Equation 6, based on the LD method, is presented in the next section.

The pooled estimates shown in Table 3.8 are computed using formulae from Rubin (1987). For an imputed data set i , let the regression coefficient estimate for a covariate be denoted as b_i and its variance as V_i , $i = 1, \dots, m$ ($= 5$ in the present illustration). The final, pooled point estimate for the regression coefficient is the average of all b_i 's:

$$\bar{b} = \frac{1}{m} \sum_{i=1}^m b_i. \quad (8)$$

The within-imputation variance of these point estimates is given by

$$WV = \frac{1}{m} \sum_{i=1}^m V_i. \quad (9)$$

The between-imputation variance of these point estimates is given by

$$BV = \frac{1}{m-1} \sum_{i=1}^m (\bar{b}_i - \bar{b})^2 \quad (10)$$

Thus, the total variance associated with the pooled estimate $\bar{b}$ is

$$TV = WV + \left(1 + \frac{1}{m}\right) BV. \quad (11)$$

From the total variance, Rubin (1987) derived the sampling distribution (i.e., t) for the statistic $(b - \bar{b})/\sqrt{TV}$ with the df_m equal to

$$df_m = (m-1) \left[1 + \frac{WV}{(1 + m^{-1})BV} \right]^2. \quad (12)$$

The above df_m and its associated t -statistic were used by PROC MIANALYZE to compute the significance level (p) and 95% confidence intervals around estimated regression coefficients, as reported in Table 8. The level of confidence interval may be specified by researchers to a level other than 95%. Pooled estimates can also be computed by PROC MIANALYZE for a variety of statistical indices such as, means, variances, correlations, p -values, odds ratios, likelihood-ratio test statistic, and covariance matrices in both univariate and multivariate cases (Rubin, 1987; SAS Institute, 2000; Schafer, 1997).

Comparison of MI Results with the LD Result

Comparing results presented in Equation 7 and Table 3.8 to those expressed in Equation 6 and Table 3.6, one notices that both models uncovered the same relationship between the logit of the dependent variable and the five covariates. The standard errors in Tables 3.6 and 3.8 were also quite similar except for the GENDER predictor.

There was one key difference between these two results. The significance levels associated with the t -test in Table 3.8 were larger, i.e., less statistically significant, for variables GENDER, FAMILY, and ESTEEM than those associated with Wald's chi-square tests in Table 3.6. This difference illustrates how MI captures the uncertainty associated with missing values, and, consequently, leads to more valid statistical inferences in terms of both null hypothesis testing and interval estimation of the regression coefficients.

Comments on the MI Method and PROC MI in SAS®

Key to the successful implementation of the MI method is the specification of the imputation model. MI assumes that the imputation model is identical to the model that a researcher will use to analyze the data (i.e., the analysis model). In practice, though, these two models need not be identical. According to Schafer (1997), the imputation model should contain predictors that are important substantively, highly predictive of variables with missing data, and reflect special features of the sample design (e.g., probability surveys). The imputation model does not have to be conceptually meaningful. In practice, it typically contains the variables of substantive interest as well as predictors believed to reflect missingness. There is empirical evidence that the imputation model can be useful in reducing bias, even if it is mis-specified (Schafer). With the inclusion of auxiliary variables (the inclusive strategy) in an imputation model, estimation bias can be reduced and efficiency can be increased, at very minor costs (Collins, Schafer, & Kam, 2001). The inclusive strategy worked remarkably well even with liberal use of auxiliary variables and/or different missing data mechanisms (MCAR, MAR, or MNAR, i.e., missing not at random) (Collins et al.).

MI is a Bayesian approach; it is similar to the EM method in that maximum likelihood estimates of population parameters are calculated based on observed data only. Both approaches are iterative and must converge to a criterion established by the programmer or the researcher. Both methods summarize a likelihood function that has been integrated over a conditional distribution for the missing data, conditioned on the observed data and estimated parameters. The major difference between these two is that EM accomplishes this task by numerical algorithms whereas the MI method does so by Monte Carlo approaches. Specifically, PROC MI in SAS® and the NORM program written by Schafer use a data augmentation algorithm to create a Markov chain from which random draws of missing data are sampled from the conditional distribution of missing values, conditioned on the observed data and estimated parameters (SAS Institute, 2000; Schafer, 1997). These parameter estimates are themselves random draws from the posterior distribution finalized after the Markov chain becomes stationary.

The data augmentation algorithm used by PROC MI assumes multivariate normality for the complete data in order to impute values for missing data, or just enough missing values so that the imputed data sets have monotone missing patterns (SAS Institute, 2000). Violation of the multivariate normality assumption had little impact on inferences regarding the first and the second moments of the population distribution, i.e., the mean and the variance, as reported in Graham and Schafer (1999).

Table 3.9. Percent of Efficiency as a Function of Rate of Missing (γ) and the Number of Imputations (m)

m	γ				
	.1	.2	.4	.5	.9
3	97	94	88	86	79
5	98	96	95	91	86
10	99	98	96	95	93
20	100	99	98	98	96

The efficiency of a parameter estimate based on m imputations is $(1 + \gamma/m)^{-1}$, where γ is the rate of missing information (Rubin, 1987, p. 114). Table 3.9 relates the rate of missing (γ) with the number of imputations (m) for efficiency of recovery of the true parameter. It is clear from Table 3.9 that, for low rates of missing data (20% or less), no more than $m = 3$ imputations are needed to be at least 90% efficient. For higher rates of missing, $m = 5$ or 10 imputations are needed. In general, it is recommended to impute missing values 5 times before results are pooled (Schafer & Olsen, 1998). The MI procedure in SAS® defaults the number of imputations to 5.

The importance of correctly specifying the imputation model was discussed by Sinharay, Stern, and Russell (2001) based on a simulation study. In that simulation study, four aspects of data were manipulated: (1) missing data mechanism (MAR versus Missing not at random, or MNAR), (2) percent of missing data (10%, 40%, and 70%), (3) the true parameter values (the mean of a variable, say X_1 , and the correlation between X_1 and X_2 with population values specified to be 0, 0.3, 0.6, or 0.8), and (4) the imputation model (a multivariate model with 2, 3, 10, or 20 variables to impute the missing values). For each combination of these conditions, 10,000 data sets were constructed and analyzed in order to yield estimates for the mean of a variable (e.g., *income* with missing values) and correlation between *income* and another variable (e.g., *education*). Results suggested that the efficiency of MI depend on the imputation model. If the imputation model is correctly capturing the underlying association between the variable (such as *income*) with missing values and other variables with complete data (such as *education*), MI yields good results even under the MNAR condition.

Despite its strengths, there are several concerns with MI that have been documented in the literature. The first concern is that each application of MI produces slightly different imputed values and associated statistics. Consequently, results cannot always be replicated if MI is used to handle missing data. Other things being equal, this issue becomes increasingly

important as the amount of missing data increases. Second, Allison (2001) discusses difficulties with MI when the researcher's interest is in estimating interaction effects. A third concern is the limited availability of MI software in general-purpose statistical software. In addition to the two procedures (MI and MIANALYZE) implemented in SAS® version 8.2 and beyond, a free MI software NORM, written by Schafer (1997), can also be utilized to perform the multiple imputation under the multivariate normal assumption. This software is available from <http://www.stat.psu.edu/~jls/misofwa.html>, along with three additional modules also written by Schafer: CAT, MIX, and PAN. CAT performs MI for categorical data under the loglinear distribution assumption, MIX performs MI for a mixture of continuous and categorical data under the general location distribution assumption, and PAN performs MI for panel data or clustered data under a multivariate linear mixed-effects distribution assumption. Although Schafer's software allows MI to be applied to a variety of missing data problems, some educational researchers may find these programs to be less than user-friendly.

IMPORTANCE OF STATISTICAL ASSUMPTIONS FOR PRINCIPLED METHODS

The FIML, EM, and MI methods described in the preceding sections rely on two key assumptions. One is that the missing data are missing at random. To maximize the likelihood of satisfying the MAR assumption, researchers need to plan for ignorability in the design of a study (Heyting & Tolboom, 1994). In other words, if a researcher anticipates the occurrence of missing data, he/she needs to plan on collecting information from variables believed to be related to, or predictive of, missingness as a routine part of the design of a study, even though these variables may be peripheral to the research question under pursuit.

According to Graham and Schafer (1999), data-analytic results are relatively insensitive to the misspecification of the missing data model because such a model does not distort the entire data distribution, only the portion that is missing. Similarly, Schafer (1997) argued that a missing data model that explains some, but not all, of the missingness reduces the bias accordingly, and is likely to produce parameter estimates superior to those based on LD. In a small simulation study investigating the consequences of failing to satisfy the MAR assumption, Schafer (1997) showed that treating the missing data as though they were missing at random produced less biased means and standard deviations than LD. Ezzati-Rice, Johnson, Khare, Little, Rubin, and Schafer (1995) and Heijman and Basu (1996) reported similar findings. In sum, the best advice is for researchers

to plan on collecting data so as to maximize the likelihood of satisfying the ignorability condition.

The FIML, EM, and MI methods (as implemented in SAS® PROC MI and the stand-alone NORM program) also assume that complete data follow a multivariate-normal distribution. Again, there is surprisingly little literature examining effects of nonnormality on parameter estimates. While it may seem reasonable to conclude that drastic departures from this assumption will seriously bias inferences derived under these methods, some authors disagree. Graham and Schafer (1999) claimed that nonnormality had little impact on inferences derived from MI because the imputed values would still resemble the observed data in their first and second moments (i.e., the mean and the variance). Since most data analyses of interest in educational research are based on the first and second moments, it can be argued that the damage inflicted by nonnormality is usually not severe enough to significantly bias inferences. Graham and Schafer went on to argue that nonnormality is more of a threat to inferences based on parameters other than the first or second moments, for example, a variable's 95th percentile. Despite the optimism of Graham and Schafer, it seems prudent to use missing data methods cautiously when there is evidence of substantial nonnormality. And there is some evidence that parameter estimates under FIML are not robust to departures from multivariate normality (Arbuckle, 1996; Enders, 2001).

IMPLICATIONS AND RECOMMENDATIONS FOR EDUCATIONAL RESEARCH

Just as qualitative researchers in education have struggled to keep up with rapidly emerging approaches during the past 20 years, so must quantitative researchers incorporate advances provided by methodologists in their realm of inquiry. The impact of missing data on quantitative research should be a concern to educational researchers. It has been addressed in the methodological literature (e.g. Afifi & Elashoff, 1966; Becker & Powers, 2001; Becker & Walstad, 1990; Cool, 2000; Coons, 1957; Kim & Curry, 1977; Rubin, 1987; Schafer, 1999; Tirri & Silander, 1998), and substantial progress has been made in the last 2 decades of the twentieth century (Little & Rubin, 1987; Schafer, 1997; special section of *Psychological Methods*, December 2001).

However, as shown in our review of quantitative studies published in 11 education journals between 1998 and 2004, this progress has not had an impact on the way educational researchers handled missing data. In studies where missing data were reported or detected, 92.6% treated the problem with either the LD or the PD method, 5% did nothing about this

problem, and less than 1% used the principled EM or MI method. An APA task force (Wilkinson & Task Force, 1999) specifically warned against the use of LD and PD in empirical research.

Newer and more principled methods, such as the full information maximum likelihood and EM, take into account conditions under which missing data occur; hence, they should be preferred to the nonprincipled methods. The full information maximum likelihood and EM methods often produce little or no bias in parameter estimates. Nor do they artificially reduce variation in the sample across a range of missing data conditions and varying amounts of missing data. Simulation studies in general demonstrated the advantages of the principled methods extended to both the *missing completely at random* and the *missing at random* conditions. The inclusion of these methods in widely available statistical software (e.g., SPSS® and SAS®) increases their accessibility and attractiveness to educational researchers. In this chapter, we demonstrated the use of three of these principled methods for treating missing values, namely, the Full Information Maximum Likelihood (FIML) method, the SPSS® Expectation-Maximization (EM) method, and the Multiple Imputation (MI) method. Evidence of the performance of these methods, along with the ad hoc procedures, was documented.

To minimize, if not eliminate, potential bias in quantitative research findings and to ensure credible conclusions upon which we base much of our thinking, it is essential that educational researchers keep the following recommendations in mind.

First and the foremost, efforts should be exerted to collect data to the fullest extent and of the highest quality. By so doing, educational researchers keep missing data to a minimum and, therefore, reduce bias and distortion in estimating population parameters or testing pertinent hypotheses. Even though principled procedures for handling missing data have been devised, their validity depends on missing data mechanism and proper specifications of models for treating missing data. Everitt (1998, p. 9) asserted that "the percentage of missing data in a study can be considered as one indicator of the quality of the data and, therefore, the quality of the study. It is important not to be seduced into thinking that investigations that are carried out poorly can be rescued by sophisticated statistical analysis." Since no one procedure has proven to be uniformly superior to others, we concur with Anderson et al. (1983) that "the only real cure for missing data is to not have any" (p. 480).

Second, educational researchers should provide sufficient and consistent information in research reports so that readers may be able to evaluate the soundness of findings based on the methodology. Specifically, educational researchers should always report the actual sample size used in the analysis, the presence or absence of missing data, the cause(s) of

missing data, and the treatment of missing data. If results are reported in percentages, it is necessary to also include the sample size on which the percentage is based. Summary statistics such as the sample size, the number of nonresponses, the response rate, and *df* of *t*, chi-square, or *F* statistics should be presented clearly and consistently in the text and/or summary tables throughout the report.

Third, if missing data are present, researchers must determine whether the cause and pattern of the missing data will seriously impair the quality of the inferences derived and which procedure, if any, is most appropriate for handling missing data. A careful examination of factors causing missing data and the missing data pattern allows researchers to decide if and how to best deal with missing data in a study. To this end, Table 3.3 may be a good place to start.

Last, it is our hope that these recommendations will move educational researchers as well as journal editors and reviewers toward formulating new standards in research practices and editorial policies for the treatment of missing data in quantitative research. These new standards may be worded similarly to those regarding the effect size reporting. If adopted by all journal editorials, they should enhance the quality of research results and publications in education. The improved quality of research findings should in turn help to construct a knowledge base that is supported by sound methodologies.

REFERENCES

- Arbuckle, J. L. (1995) *Amos for Windows. Analysis of moment structures* (Version 3.6) [Computer Software]. Chicago: SmallWaters Corp.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-278). Mahwah, NJ: Lawrence Erlbaum.
- Alifi, A. A., & Elashoff, R. M. (1966). Missing data in multivariate statistics: I. Review of the literature. *Journal of the American Statistical Association*, 61(315), 595-604.
- Allison, P. D. (2001). *Missing data* (Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136). Thousand Oaks, CA: Sage.
- Anderson, A. B., Basilevsky, A., & Hum, D. P. J. (1983). Missing data: A review of the literature. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 415-494). San Diego: Academic Press.
- Arbuckle, J. L. (1996). Full information likelihood estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243-277). Mahwah, NJ: Erlbaum.

- Basilevsky, A., Sabourin, D., Hum, D., & Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Communications in Statistics*, 14, 371-394.
- Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, 37, 129-146.
- Becker, W. E., & Powers, J. (2001). Student performance, attrition, and class size given missing student data. *Economics of Education Review*, 20, 377-388.
- Becker, W. E., & Walstad, W. B. (1990). Data loss from pretest to posttest as a sample selection problem. *The Review of Economics and Statistics*, 72 (1), 184-188.
- BMDP Statistical Software, Inc. (1992). *BMDP statistical software manual* (Vol. 2). Los Angeles, CA: Author.
- Brockmeier, L. L., Kromrey, J. D., & Hogarty, K. Y. (2002, April). *Missing data in the multiple regression analysis: A comparison of ten missing data treatments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling*, 1, 287-316.
- Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B*, 22, 302-307.
- Buu, A. (1999). *Analysis of longitudinal data with missing values: A methodological comparison*. Unpublished doctoral dissertation, Indiana University.
- Chan, L. S., & Dunn, O. J. (1972). The treatment of missing values in discriminant analysis- I. The sampling experiment. *Journal of the American Statistical Association*, 67, 473-477.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330-351.
- Cool, A. L. (2000). *A review of methods for dealing with missing data*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX. (ERIC Document Reproduction Service No. ED 438 311)
- Coons, I. (1957, September). The analysis of covariance as a missing plot technique. *Biometrics*, 387-405.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- Diggle, P. J., Liang, K., & Zeger, S. L. (1994). *Analysis of longitudinal data*. New York: Oxford University Press.
- Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing data. *The American Statistician*, 36, 378-381.
- Donner, A., & Rosner, B. (1982). Missing value problems in multiple linear regression with two independent variables. *Communications in Statistics*, 11, 127-140.
- Downey, R. G., & King, C. V. (1998). Missing data in Likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, 125(2), 175-191.
- Dworkin, R. J. (1987). Hidden bias in the use of archival data. *Evaluation & the Health Professions*, 10(2), 173-185.
- Enders, C. K. (1999). *The relative performance of full information maximum likelihood estimation for missing data in structural equation models*. Unpublished doctoral dissertation, University of Nebraska, Lincoln, NE.
- Enders, C. K. (2001). The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6, 352-370.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 430-457.
- Everitt, B. S. (1998). Analysis of longitudinal data. *British Journal of Psychiatry*, 172, 7-10.
- Ezzai-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., Rubin, D. B., & Schafer, J. L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. In *Proceedings of the annual research conference* (pp. 257-266). Washington, DC: Bureau of the Census.
- Glasser, M. (1964). Linear regression with missing observations among the independent variables. *Journal of the American Statistical Association*, 59, 834-844.
- Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika*, 40, 229-252.
- Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for small-sample research (pp. 1-29). Thousand Oaks, CA: Sage.
- Graham, J. W., & Schafer, J. L. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.
- Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society, Series B*, 30, 67-82.
- Hall, J. (1993). *Skills training for pregnant and parenting adolescents* (NIDA Monograph No. 156: NIH Publication No. 95-3908). Rockville, MD: National Institute on Drug Abuse.
- Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics*, 27, 783-823.
- Heitjan, D. F., & Basu, S. (1996). Distinguishing between "missing at random" and "missing completely at random." *American Statistician*, 50, 207-213.
- Heyting, A., & Tolboom, J. T. B. M. (1994). Discussion of the paper by P. Diggle and M.G. Kenward. *Applied Statistics*, 43, 49-93.
- Holt, D. (1997). Missing data and nonresponse. In J. P. Keeve (Ed.) *Educational research, methodology, and measurement: An international handbook* (2nd ed., pp. 592-597). New York: Elsevier Science.
- Ingersoll, G. M., & Orr, D. P. (1989). Behavioral and emotional risk in early adolescents. *Journal of Early Adolescence*, 9, 392-408.
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805-820.

- Kaplan, D. (1995). The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics*, 20, 69-82.
- Kellam, S., Rebok, G., Ialongo, N., & Mayer, L. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry*, 35, 259-281.
- Kim, J.-O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2), 215-240.
- King, C. V., Fogg, R. J., & Downey, R. G. (1998, April). *Mean substitution for missing items: Sample size and the effectiveness of the technique*. Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Dallas, TX.
- 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.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.
- Littell, R. C., Roderick, J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. Clogg, & M. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenum.
- Little, R. J. A. (1988). Robust estimation off the mean and covariance matrix from data with missing values. *Applied Statistics*, 37, 23-38.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Malhotra, N. K. (1987). Analyzing marketing research with incomplete information on the dependent variable. *Journal of Marketing Research*, 24, 74-84.
- Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling*, 5, 22-36.
- Mason, M. J. (1999, July). A review of procedural and statistical methods for handling attrition and missing data in clinical research. *Measurement and Evaluation in Counseling and Development*, 32, 111-118.
- Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79, 103-111.
- McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
- Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431-462.
- Peng, C.-Y. J., & Nichols, R. N. (2003). Using multinomial logistic regression models to predict adolescent behavioral risk. *Journal of Modern Applied Statistical Methods*, 2(1), 177-188.
- Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3-14.
- Peng, C.-Y. J., So, T. S. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education*, 43(3), 259-293.
- Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, 47, 13-26.
- Reilly, M. (1993). Data analysis with hot deck multiple imputation. *The Statistician*, 42, 307-313.
- Resnick, M. D., Harris, L. J., & Blum R.B. (1993). The impact of caring and connectedness on adolescent health and well-being. *Journal of Pediatrics Child Health*, 29(1), 53-59.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton.
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537-560.
- Roth, P. L., & Switzer, E. S., III. (1995). A Monte Carlo analysis of missing data techniques in an HRM setting. *Journal of Management*, 21, 1003-1023.
- Rovine, M. J. (1994). Latent variable models and missing data analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 181-225). Thousand Oaks, CA: Sage.
- Rubin, D. B. (1987). *Multiple imputation for nonresponses in surveys*. New York: Wiley & Sons.
- SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8, volume 2*. Cary, NC: Author.
- SAS Institute Inc. (2000). *Proceedings of the twenty-fifth annual SAS® users group international conference*. Cary, NC: Author.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.
- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3-15.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavior Research*, 33, 545-571.
- Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4), 317-329.
- Switzer, E. S., III, Roth, P. L., & Switzer, D. M. (1998). Systematic data loss in HRM settings: A Monte Carlo analysis. *Journal of Management*, 24, 763-779.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika*, 35, 417-437.
- Tirri, H., & Silander, T. (1998, April). *Stochastic complexity based estimation of missing elements in questionnaire data*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED 431 794)

- Wilkinson, L., & Task Force on Statistical Inference APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594-604.
- Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematics Statistics*, 3, 163-195.
- Witt, E. L. (2000, April). *Effectiveness of four methods of handling missing data using samples from a national database*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 442 810)
- Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples* (pp. 219-240). Mahwah, NJ: Erlbaum.