Handling Missing Data in Research Studies of Instructional Technology.

2003-04-00


Reports - Research (143) -- Speeches/Meeting Papers (150)

*Data Analysis; *Educational Technology; Research Methodology; *Scholarly Journals

*Missing Data

Missing data is an important issue that is discussed across many fields. In order to understand the issues caused by missing data, this paper reviews the types of missing data and problems caused by missing data. Also, to understand how missing data are handled in instructional technology research, articles published in "Educational Media International," "Educational Technology Research and Development," and "Performance Improvement Quarterly" for the last 5 years are reviewed. A total of 84 quantitative research articles were identified in the 3 journals. About 42% of the reviewed studies had incomplete data sets, and in most of them, information about data completeness was clearly presented through comparisons of usable data points with the intended sample size. Overall, it was found that the awareness of missing data issues was low among the researchers in the field of instructional technology. Findings are discussed in terms of missing data mechanisms, and recommendations are presented. An appendix shows missing data methods in the three journals in table form. (Contains 19 references.) (Author/SLD)
Handling Missing Data in Research Studies of Instructional Technology

Jeong-Eun Oh
Instructional Systems Technology
School of Education
Indiana University, Bloomington
Bloomington, IN 47405
jeoh@indiana.edu

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in Chicago, IL
April 2003
Abstract

Missing data is an important issue that is discussed across many fields. In order to understand the issues caused by missing data, this paper reviews the types of missing data and problems caused by missing data. Also, to understand how missing data are handled in instructional technology research, articles published in Educational Media International, Educational Technology Research and Development and Performance Improvement Quarterly for the last five years are reviewed. Findings are discussed focused on the report of missing data and identification of missing data mechanisms, and finally, recommendations are presented.
Introduction

Missing data commonly occurs in many studies and are caused by various reasons. For instance, intended participants may refuse to join experiments or can drop out in the middle of research. Survey respondents can return their questionnaire partially or completely unanswered. And, some data points can be discarded if they are unreadable or unusable.

When encountering missing data, most researchers use listwise or pairwise deletion methods and analyze data as usual. Statistical packages also typically eliminate data units with missing scores in order to proceed to data analysis. Against such common practice, the American Psychological Association disapproved of researchers' exclusive reliance on listwise and pairwise deletion methods and asked researchers to give more attention to other alternatives (Wilkinson & APA Task Force on Statistical Inference, 1999). Similarly, Little and Rubin (1987) criticized saying that “[just excluding units] is generally inappropriate, since the investigator is usually interested in making inferences about the entire target population, rather than the portion of the target population that would provide responses to all relevant variables in the analysis. (p.4)”

Missing substantial amounts of data in non-random patterns can cause a gap in the representations between the sample and the intended target population and it can consequently threaten the validity of the study. Researchers need to approach and handle missing data with cautions following relevant principles and methods, but missing data are frequently deemed missing at random and simply discarded in practice. Such a naive approach, however, can produce bias in the analysis result and therefore, special attention is required in treating missing data. In order to best approximate the missing values and
represent the population without distortion, researchers need to identify the patterns of missing data and the mechanisms that cause the data to be missing, which should determine the way in which data are treated.

Missing data is a common issue across fields that employ empirical research studies. The literature on missing data has increased in various fields such as business (Roth, 1994) and psychology (Schafer & Graham, 2002), and it helps raise awareness of the issue among researchers (Little & Schenker, 1995; Roth, 1994; Schafer & Graham, 2002). However, more attention is requested on missing data issues to further develop active discussions within and across fields.

The purpose of this paper is to review the literature on missing data and to examine the awareness of the missing data issues among the researchers in the field of instructional technology. In order to achieve this goal, first, this paper will discuss the problems associated with missing data. Second, the patterns and mechanisms of missing data will be reviewed. Then, research studies in three leading journals in instructional technology will be examined to identify the way in which missing data were treated. Finally, findings will be discussed with suggestions for the improvements of research practice in instructional technology.

Literature Review

Issues with Missing Data

Some missing data can be ignored, and discarding it is a conventional practice. According to the Monte Carlo studies, less than 10% of data loss in a random way does not make much difference in the parameter estimates and the sample statistics (Beaton,
1997; Raymond & Roberts, 1987). However, if substantial amounts of data are missing, then several issues arise.

First, a loss of data can reduce the statistical power of estimates (Little & Rubin, 1995). Statistical power refers to the ability of statistical analysis to discover a relationship in a set of data when there is one (Trochim, 2001). Increased statistical power implies more efficiency in estimating “true” values from the data set and hence, attaining adequate statistical power is important in intervention research. Large sample is one of the factors that increase statistical power (Trochim, 2001), and a loss of considerable amounts of data decreases the power and accuracy of the data analysis, causing difficulty in detecting relationships in the data set and consequently influencing the validity of the research. By estimating reasonable effect size, alpha level and power, researchers might want to determine a reasonable sample size for a study and begin the study with a sample large enough to secure the desired level of statistical power of the analysis against missing data.

Second, missing data can bias parameter estimates and threaten the validity of inferences. When data are missing from certain parts along the sample distribution, statistical estimates can be biased in ways that are different from those that would be attained from complete data sets. If the purpose of analysis lies in descriptive measures of a variable such as central tendency or deviation, missing data can bias the estimates and the inferences towards the population unless the data are missing completely at random. No bias is deemed to exist in the estimates when data are missing completely at random, but not all studies require this condition as necessary. If interest lies in the conditional distribution of the variable given another variable such as distribution of
income given area, then, the analysis based on complete cases would be acceptable when the data are randomly missing within each category of areas. Depending on the objective of analysis, the significance of bias problem can be different, and when the validity of the study is seriously challenged by missing data, it is advised to redesign the study and conduct it again (Orme & Reis, 1991).

Third, missing data require cautions in their treatment (Little & Rubin, 1995). When data are missing completely at random, which is a rare case, does not cause a problem beyond the loss of statistical power that was discussed above. However, when data are missing at random or not missing at random, problems arise. Most statistical methods (e.g., t test, ANOVA, ANCOVA) presume complete data sets drawn from a normal distribution and simply discard incomplete data sets to proceed with data analysis. Conventional statistical software such as SPSS® or SAS® use listwise deletion as a default (Little & Rubin, 1987) since it does not require special computation procedures and can be used for any kind of statistical analysis (Allison, 2002). This highly efficient and most commonly used technique, however, can lead to biased estimates and misleading inferences when the data are not missing completely at random (Allison, 2002). Similarly, most of the imputation methods (e.g., mean imputation, hotdecking, regression imputation) also assume that the data are missing at random. Therefore, any attempt to ignore the assumptions about the missing data mechanism and proceed with imputation-based approaches can bias the estimates of coefficients in models. When data are not missing completely at random or when not missing at random, depending on the methods handling missing data, there exist possibilities of severe biases. In order to
prevent this problem, attention is called in identifying the mechanism of missing data and handling the missing data in appropriate ways.

As discussed so far, a loss of substantial amounts of data can be a serious threat to making valid inferences to the intended population, but the current research practice appears to remain still naïve and unprincipled. More attention is requested in treating missing data and the rigor should begin from identification of the patterns of missing data.

Patterns of Missing Data

An easy way to examine patterns of missing data is arranging the data in a matrix. With missing data replaced with 0, the matrix will describe the pattern of missing data such as data missing at random or data missing in special patterns. For instance, univariate missing data occur when data are missing confined to a certain variable. Unit nonresponse occurs when data are not available for analysis for a case. It is referred to monotone missing data when values of variables are missing in block and the number of missing data increase monotonically (Little & Rubin, 1989). These missing data patterns may occur by design (e.g., planned censoring) or by accident (e.g., accidental censoring, administrative loss). Whether intended or not, patterns of missing data emerge due to the mechanisms that cause missing data. Therefore, if special patterns of missing data are observed, the mechanism leading to nonresponse needs to be identified and whether it is ignorable needs to be determined.

Mechanisms of Missing Data

One of the two major criteria that determine the way by which missing data are treated is missing data mechanism (For details, see Roth, 1994). Mechanisms of missing data imply possible subsequent problems and determine the appropriateness of the way to
handle missing data (Little & Rubin, 1987; Raymond & Roberts, 1987; Roth, 1994). Therefore, it is important to understand assumptions about the missing data mechanisms. Depending on the location of the variables that cause missing, missing data mechanisms are categorized into three types: (1) missing completely at random, (2) missing at random, and (3) not missing at random. (Little & Rubin, 1987, p.14)

Data is said to be missing completely at random (MCAR) when data is missing unrelated to the values of the variable or to the values of other variables in the data set. With MCAR, the missing values are viewed as a simple random sample of all possible data values of the given variable, and it is assumed that no distinguishable difference exists between the cases with incomplete data and those with complete data sets. The set of cases with complete data are considered as a random subsample of the original set of observations (Little & Schenker, 1995), and hence no bias is deemed to reside in the results when the incomplete data set is discarded and only the cases with complete data are analyzed (Roth, 1994). MCAR data is viewed the least problematic type (Kim & Curry, 1977), but Allison (2002) advised that MCAR does not exclude the possibility of a relationship between missing data on one variable and missing data on another variable.

Another mechanism of missing data is data missing at random (MAR) (Little & Schenker, 1995). MAR occurs when data is missing unrelated to the values of the variable being measured given that the covariates are controlled (Allison, 2002; Kim & Curry, 1977). When data are MAR, missing values are considered as random samples of all possible values of the variable within each class of the covariate. MAR assumes that the probability of missing data is not related to the variable itself but related to the covariate, which differentiates between non-respondents and respondents. If the
underlying relationship between the missing data and the influence variable is identifiable, the pattern of missing data is predictable (Little & Rubin, 1987), for instance, using the sophisticated prediction-estimation algorithm (Johnson & Wichern, 1992, pp.202-207). Since the missing data mechanism is not related to the parameters to be estimated, however, missing data are ignorable if they are MAR. Most imputation methods assume MAR.

Finally, when the probability of missing data is related to the unobserved value, the data are said to be not missing at random (Little & Rubin, 1987). When data are NMAR, it is assumed that values from respondents and non-respondents systematically differ, and without information about the mechanism leading to the NMAR data, bias is unavoidable from any standard data analysis methods, which assume normal distribution of data. Therefore, when data are NMAR, researcher needs to create a nonresponse model to adjust for biases in using the collected data.

Method

Sample

In order to understand how missing data are treated in research studies of instructional technology, the researcher reviewed quantitative studies published from 1998 to 2002 in three journals in instructional technology. As a pool of quantitative studies to be examined, Educational Media International (EMI), Educational Technology Research and Development (ETR&D), and Performance Improvement Quarterly (PIQ) were selected. Both EMI and ETR&D are lead journals that are distinguished with their comprehensive contents and long publication history. Compared to these two journals, PIQ is astray from the mainstream of the field to some extent but has been one of the
major references in instructional technology as a publication from International Society for Performance Improvement, a largest professional association in the field of instructional technology. (B. A. Bichelemeyer, personal communication, June 20, 2002; M. Molenda, personal communication, June 20, 2002). Considering such indicators as publication history and recognition among professionals, it was assumed that the three journals would be able to reflect the conventional research practice in instructional technology on behalf of the other journals in the field.

Analysis

All studies employing quantitative research methods were examined, from completely quantitative research studies to studies using a mixed-method approach. As in the latter case when an article includes both quantitative and qualitative studies, only quantitative studies were considered for this study. For each study, method, results, discussion sections or their equivalents were reviewed in order to find information about how missing data were handled. Summary tables were examined as well with special attention to the total sample size, the df of the statistics reported, so as to confirm the completeness of data points. All the studies were reviewed twice with some time interval in order to ensure the accuracy of the analysis results.

Findings and Discussions

Overall Findings

A total of 84 quantitative research articles (85 studies) were identified from the three journals (19 from EMI, 33 from ETR&D, and 32 from PIQ). Of the 85 studies, 34 studies (40%) reported no missing data and 36 studies (42%) explicitly or implicitly confirmed the incomplete data sets. Fifteen studies (18%) did not provide enough
information (e.g., sample size, \( df \) of statistics estimates) for a reader to determine if missing data were there. Listwise deletion method was the method most frequently used in the studies with missing data. A summary table of these findings is presented in Appendix A.

**Reporting Missing Data**

About 42% of the reviewed studies had incomplete data sets. Among these 36 studies with missing data, thirty two studies (89%) reported both initial sample sizes and final sample sizes in the contents, and four studies (11%) provided usable data points in a summary table or degrees of freedom for statistical estimates so that readers could identify or infer the sample sizes. Fifteen studies (18%) out of the 85 reviewed studies failed to let readers know the presence of missing data, by omitting critical information such as Ns used for the mean calculations or the degrees of freedom for the statistical estimates.

In most of the studies with missing data, information about data completeness was clearly presented through comparisons of the usable data points with the intended sample size. On the other hand, it was difficult in some studies to identify the existence and amounts of missing data due to the incompleteness of or inconsistency between the information about intended observations and finally usable cases. Survey studies frequently revealed inconsistencies between the number of total responses and the numbers of item responses. Similarly, one study acknowledged only the fact that Ns varied by measure, without further information about the usable data points for each item. Another study reported different Ns for each item but not for all the items. In these cases, it was difficult to judge how substantial the losses of data were and whether they were
ignorable in the data analyses. Such failures in providing sufficient information about data completeness were mostly found in the studies using surveys with numerous items or multiple measures for data collection.

Identifying Missing Data Mechanisms

Some studies with missing data tried to provide accounts of what caused the data missing. Of the 36 studies that acknowledged the existence of missing data in one way or another, only 13 studies (36%) identified why data were lost. Most of the causes of missing data were described at the surface level such as participants’ failures to attend tests or leaving the research setting, and the data losses were considered to occur randomly, unrelated to the variables under study. Most of these studies with missing data did not cover the missing data mechanisms in detail and did not discuss possible effects of missing data on the research results. However, there were some studies that treated the missing data with more attention.

For instance, Goins & Mannix (1999) identified the differences between attained sample and intended sample and called special attention when generalizing the research findings. Hancock & Flowers (2001), Twitchell, Holton, & Trott (2000) and Ellinger, Keller, & Ellinger (2000) conducted non-response bias test and confirmed that no difference existed between the cases with complete data and those with missing data. Similarly, Watkins (2001) implied the possibility of NMAR data, by acknowledging that the sample was volunteer-based and requiring limited generalization of the research findings. Wognum (2001) insisted that by underlining the focus of the study, missing data points would not make a difference for the given purpose of the study (i.e., exploration of varieties).
From this small number of studies that tried to identify missing data mechanisms and relevant issues, missing data seemed to be an uninformed issue to the researchers in instructional technology. This impression was enhanced by some studies that ignored issues with missing data in spite of the decrease of final sample size to a less than half of the initial sample size. Such great losses of data were frequently found in survey studies and were often ignored especially when the usable sample size was still large. Regarding this problem, researchers should be noted that it is not the final sample size but the difference between the intended sample size and the final sample size that work as a criterion in determining the data representativeness.

Missing Data Methods Reported in the Journals

To report missing data methods used in the reviewed studies is beyond the intended scope of this paper but briefly summarizing, none of the studies explicitly mentioned or implied what missing data methods they used and it was difficult to understand how missing data were handled particularly when data are compared missing in different amounts by measures or experiment settings. However, based on the information given in the contents or in the tables, it was identified or inferred that all the studies with missing data used the listwise deletion method, except for one study that used the pairwise deletion method as well. Since most of the studies did not present sufficient information about the missing data mechanisms, it was hard to judge if listwise deletion was an appropriate choice. The exclusive dependence on the listwise deletion method may reflect researchers’ exclusive reliance on the conventional statistical packages such as SAS®, SPSS® in which listwise is set as a default. It may also imply the researchers’ lack of knowledge about missing data issues.
Recommendations

From the review of studies that had incomplete data but did not handle them in professional manners, it was found that the awareness of missing data issues was low among the researchers in the field of instructional technology. Researchers and journal editors are requested to pay more attention in preventing and handling missing data and journal editors may need to make journal submission guidelines more thorough with regards to research methodology.

Researchers should actively plan research for missing data not to occur and report research procedures completely. If missing data occurred, sufficient information should be given about missing data so that readers can evaluate the appropriateness of data analysis procedures and the validity of research findings. Information about missing data should include the amount of missing data, the causes of missing data, and the treatments of missing data. Any differences between intended and usable data points, the total and item response rates or different measurements should be reported and the degrees of freedom for t and F tests should be clearly given as well.

To identify any considerable differences between the respondents and the non-respondents, it would be necessary to subsample the non-respondents. This will help make inferences about the non-respondents if needed. It would be also useful to try to collect covariates that are likely to predict the missing values since it will help appropriate modeling of the missing data mechanism when adjustment for the missing values are necessary.

There is no better way to handle missing data problems than not losing any data at all (Beaton, 1997; Cool, 2000; Kim & Curry, 1977; Little & Rubin, 1987). If missing
data is unavoidable, MCAR assumption is most desired though it is practically difficult to meet. Researchers should make every efforts to design the survey or experiment and the data collection protocols carefully to collect data as completely as possible. This careful planning and implementation of research will help researchers to achieve MAR data or collect information helpful creating a model of missing data, if necessary, that best approximates the missing values.

This paper tried to cover issues related to missing data and assess the level of awareness of the issues among the researchers in the filed of instructional technology. Due to the limited time and scope, this paper covered only part of the issues but continuous interest and more research are encouraged regarding other issues including missing data handling methods.
References


*Educational research, methodology, and measurement: An international
handbook* (2nd ed.). Cambridge, UK: Cambridge University Press.


*Biometrics, 46*, 143-155.

the Annual Meeting of the Southwest Educational Research Association, Dallas,
TX. (ERIC Document Reproduction Service No. ED 438 311).

integration: An evaluation of three strategic approaches for performance


Hancock, D. R., & Flowers, C. P. (2001). Comparing social desirability responding on
world wide web and paper-administered surveys. *Educational Technology
Research and Development, 49*(1), 5-13.


## Appendix A. Missing Data Methods Reported in the Three Journals

<table>
<thead>
<tr>
<th>Journal</th>
<th>Articles</th>
<th>Studies</th>
<th>Data</th>
<th>Methods</th>
<th>Reasons for Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Complete</td>
<td>Unknown</td>
<td>Missing</td>
</tr>
<tr>
<td>Educational Media International</td>
<td>19</td>
<td>19</td>
<td>10 (53%)</td>
<td>7 (37%)</td>
<td>2 (10%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LD</td>
<td>PD</td>
<td>Discussed and Treated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>1 (50%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not Discussed and Treated</td>
<td></td>
<td>1 (50%)</td>
</tr>
<tr>
<td>Educational Technology Research and Development</td>
<td>33</td>
<td>34</td>
<td>17 (50%)</td>
<td>6 (18%)</td>
<td>11 (32%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LD</td>
<td>PD</td>
<td>Discussed and Treated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>1</td>
<td>7 (64%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not Discussed and Treated</td>
<td></td>
<td>4 (36%)</td>
</tr>
<tr>
<td>Performance Improvement Quarterly</td>
<td>32</td>
<td>32</td>
<td>7 (22%)</td>
<td>2 (6%)</td>
<td>23 (72%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LD</td>
<td>PD</td>
<td>Discussed and Treated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>23</td>
<td>0</td>
<td>8 (35%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not Discussed and Treated</td>
<td></td>
<td>15 (65%)</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>85</td>
<td>34 (40%)</td>
<td>15 (18%)</td>
<td>36 (42%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LD</td>
<td>PD</td>
<td>Discussed and Treated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>1</td>
<td>16 (44%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not Discussed and Treated</td>
<td></td>
<td>20 (56%)</td>
</tr>
</tbody>
</table>
I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: 

Organization/Address: INDIANA UNIVERSITY
DEPARTMENT OF INSTRUCTIONAL SYSTEMS TECHNOLOGY
W.W. WRIGHT EDUCATION BUILDING
201 NORTH ROSE AVENUE
BLOOMINGTON, IN 47405-1086

Printed Name/Position/Title: JEONG-EUN OH
Telephone: 812-333-0486
E-Mail Address: jeooh@indiana.edu
Date: June 7, 2003
III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

<table>
<thead>
<tr>
<th>Publisher/Distributor:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Address:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Price:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

<table>
<thead>
<tr>
<th>Name:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Address:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org