Data Preparation

Measurement Scales

- Data analysis involves the partitioning, identification, and measurement of variation in a set of variables, either among themselves or between a dependent variable and one or more independent variables.

- The key word is measurement because the researcher cannot partition or identify variation unless it can be measured.

- Measurement is important in accurately representing the concept of interest and is instrumental in the selection of the appropriate multivariate method of analysis.

Measurement Scales

- Understanding the different types of measurement scales is important for two reasons.
  - First, the researcher must identify the measurement scale of each variable used.
  - Second, the measurement scale is critical in determining which multivariate techniques are the most applicable to the data, with considerations made for both independent and dependent variables.

- The metric or non metric properties of independent and dependent variables are the determining factors in selecting the appropriate technique.

Measurement Scales

- There are two basic kinds of data:
  - Nonmetric (Qualitative)
  - Metric (Quantitative)

Measurement Scales
“Nonmetric”

- Nonmetric data are attributes, characteristics, or categorical properties that identify or describe a subject.
- Nonmetric data describe differences in type or kind by indicating the presence or absence of a characteristic or property.
- Many properties are discrete in that by having a particular feature, all other features are excluded, for example, if one is male, one cannot be female. There is no “amount” of gender, just the state of being male or female.

Measurement Scales
“Metric”

- Metric data measurements are made so that subjects may be identified as differing in amount or degree.
- Metrically measured variables reflect relative quantity or degree.
- Metric measurements are appropriate for cases involving amount, such as the level of satisfaction to a job.

Nonmetric Measurement Scales

- Nonmetric measurement can be made with either:
  - Nominal Scale
  - Ordinal Scale

Nonmetric Measurement Scales "Nominal"

- Measurement with a nominal scale assigns numbers used to label or identify subjects or objects.
- Nominal scales, also known as categorical scales, provide the number of occurrences in each class or category of the variable being studied.
- The numbers or symbols assigned to the objects have no quantitative meaning beyond indicating the presence or absence of the attribute or characteristic under investigation.
- Examples of nominally scaled data include an individual's gender, religion, or political party. The researcher might assign numbers to each category, for example, 2 for females and 1 for males. These numbers only represent categories or classes and do not imply amounts of an attribute or characteristic.

Nonmetric Measurement Scales "Ordinal"

- Ordinal scales are the next higher level of measurement precision.
- Variables can be ordered or ranked with ordinal scales in relation to the amount of the attribute.
- Every subclass can be compared with another in terms of a "greater than" or "less than" relationship.
- For example, different levels of an individual consumer's satisfaction with several new products can be illustrated on an ordinal scale.
- Numbers utilized in ordinal scales are nonquantitative because they indicate only relative positions in an ordered series. There is no measure of how much satisfaction the consumer receives in absolute terms, nor does the researcher know the exact difference of satisfaction.
- Many scales in the behavioral science fall into this ordinal category.

Metric Measurement Scales

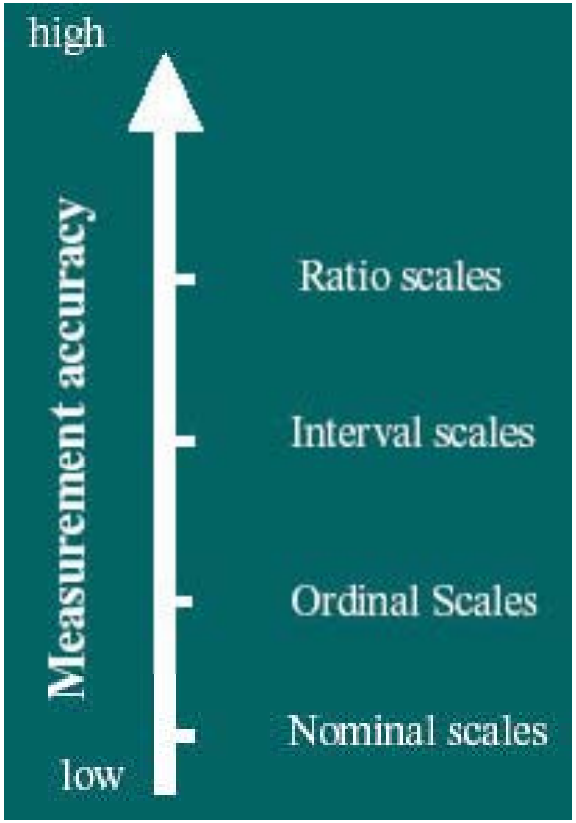
- Metric measurement can be made with either:
  - Interval Scale
  - Ratio Scale

Metric Measurement Scales

- Interval scales and ratio scales provide the highest level of measurement precision.
- These two scales have constant units of measurement, so differences between any two adjacent points on any part of the scale are equal.
- The only real difference between interval and ratio scales is that interval scales have an arbitrary zero point, whereas ratio scales have an absolute zero point. The most familiar interval scale is Celsius temperature scale and ratio one is weight.

Measurement Accuracy

- Ratio scales represent the highest form of measurement precision because they possess the advantages of all lower scales plus an absolute zero point.



What's the most important?

- The trick is NOT the computation, but the selection of reliable and valid measurement, use of appropriate technique and the appropriate interpretation of the results.

Examining Your Data

- ***OBJECTIVES***
- Select the appropriate graphical method to examine the characteristics of the data or relationships of interest.
- Understand the different types of missing data processes.
- Assess the type and potential impact of missing data.
- Explain the advantages and disadvantages of the approaches available for dealing with missing data and outliers.

Examining the Data

- By examining the data before the application of a multivariate technique, the researcher gains several critical insights into the characteristics of the data.

- First and foremost, the researcher attains a basic understanding of the data and relationships between variables. Multivariate techniques place great demands on the researcher to understand and interpret results based on relationships that are ever increasing in complexity.

- Second, multivariate techniques demand much more from the data they are to analyze. The statistical power of the multivariate techniques requires larger data sets and more complex assumptions than encountered with univariate analyses.

- The analytical sophistication needed to ensure that the statistical requirements are met has forced the researcher to use a series of data examination techniques that in many instances match the complexity of the multivariate techniques.

Separate Phases of Data Examination

1) A *graphical examination* of the nature of the variables in the analysis and the relationships that form the basis of multivariate analysis;

2) An evaluation process for understanding the impact that *missing data* and *outliers*, can have on the analysis, plus alternatives for retaining cases with missing data in the analysis.
These cases may misrepresent the relationships by their uniqueness on one or more of the variables under study.

Graphical Examination of the Data

- The use of multivariate techniques places an increased load on the researcher to understand, evaluate, and interpret the more complex results.

- One aid in these tasks is a thorough understanding of the basic characteristics of the underlying data and relationships.

- When univariate analyses are considered, the level *of* understanding is fairly simple. But as the researcher moves to more complex multivariate analyses, the need and level of understanding increase dramatically.

Graphical Examination of the Data

- The Nature of the Variable: Examining the Shape of the Distribution
- Examining the Relationship between Variables
- Examining Group Differences

The Nature of the Variable
Examining the Shape of the Distribution

- The starting point for understanding the nature of any variable is to characterize the shape of its distribution.
- Although a number of statistical measures are available, but many times the researcher can gain an adequate perspective on the variable through a histogram.
- A histogram is a graphical representation of a single variable that represents the frequency of occurrences (data values) within data categories.
- The frequencies are plotted to examine the shape of the distribution of responses.

Examining the Relationship between Variables

- Whereas examining the distribution of a variable is essential, many times examining relationships between two or more variables is also interested in.

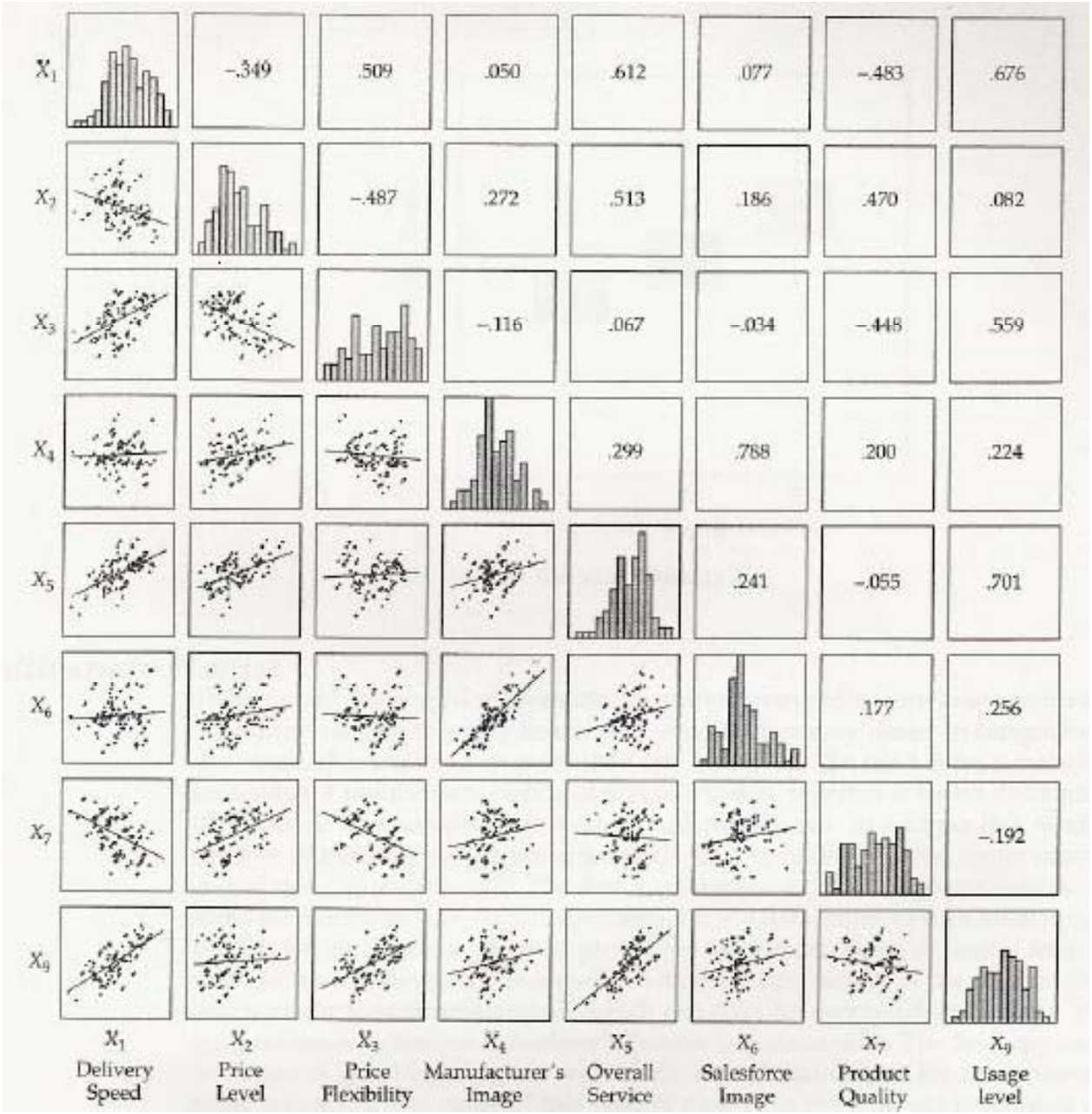
- The most popular method for examining bivariate relationships is the scatterplot, a graph of data points based on two variables. One variable defines the horizontal axis and the other variable defines the vertical axis. The points in the graph represent the corresponding joint values of the variables for any given case.

Examining the Relationship between Variables

- The pattern of points represents the relationship between variables. A strong organization of points along a straight line characterizes a linear relationship or correlation. A curved set of points may denote a nonlinear relationship, which can be accommodated in many ways. Or there may be only a seemingly random pattern of points, indicating no relationship.

- One format of *scatterplot* particularly suited to multivariate techniques is the *scatterplot matrix*, in which the *scatterplots* are represented for all combinations of variables in the lower portion of the matrix. The diagonal contains histograms of the variables. Included in the upper portion of the matrix are the corresponding correlations so that the reader can assess the correlation represented in each *scatterplot*.

Example


X1
-.349
.509
.050
.612
.077
-.483
.676
X7
-.487
.272
.513
.186
.470
.082
X3
-.116
.067
-.034
-.448
.559
X4
.299
.788
.200
.224
X5
.241
-.055
.701
X6
.177
.256
X7
-.192
X9
X1 Delivery Speed
X2 Price Level
X3 Price Flexibility
X4 Manufacturer's Image
X5 Overall Service
X6 Salesforce Image
X7 Product Quality
X9 Usage level

Scatterplot Matrix

- Figure presents the scatterplots for a set of variables (Xl, X2, X3, X4, X5, X6, X7, and X9). For example, the scatterplot in the bottom left corner (Xl versus X9) represents a correlation of .676. The points are closely aligned around a straight line, indicative of a high correlation.

- The scatterplot in the leftmost column, third from the top (Xl versus X4) demonstrates the opposite, an almost total lack of relationship as evidenced by the widely dispersed pattern of points and the correlation .050.

- Scatterplot matrices and individual scatterplots are now available in all popular statistical programs.

Examining Group Differences

- The researcher is also faced with understanding the extent and character of differences between two or more groups for one or more metric variables.

- In these cases, the researcher needs to understand how the values are distributed for each group and if there are sufficient differences between the groups to support statistical significance.

Missing Data

- Missing data are a fact of life in multivariate analysis; in fact, rarely does the researcher avoid some form of missing data problem. For this reason, the researcher's challenge is to address the issues raised by missing data that affect the *generalizability* of the results.

- To do so, the researcher's primary concern is to determine *the reasons underlying the missing data.* This need to focus on the reasons for missing data comes from the fact that the researcher must understand the processes leading to the missing data in order to select the *appropriate course of action*.

- A missing data process is any systematic event external to the respondent (such as data entry errors or data collection problems) or action on the part of the respondent (such as refusal to answer) that leads to missing values.

- The effects of some missing data processes are known and directly accommodated in the research plan. But others, particularly those based on actions by the respondent, are rarely known. When the missing data processes are unknown, the researcher attempts to identify any patterns in the missing data that would characterize the missing data process.

Missing Data

- In doing so, the researcher asks such questions as:
  - ✓ Are the missing data scattered randomly throughout the observations or are distinct patterns identifiable? and
  - ✓ How common are the missing data?

- Any statistical results based on these data would be biased to the extent that the variables included in the analysis are influenced by the missing data process. The concern for understanding the missing data processes is similar to the need to understand the causes of nonresponse in the data collection process. For example,
  - ✓ Are those individuals who did not respond different from those who did?
  - ✓ If so, do these differences have any impact on the analysis, the results, or their interpretation? Concerns similar to these also arise from missing responses for individual variables.

- The different types of missing data processes, methods to identify the nature of the missing data processes, and available remedies for accommodating missing data into multivariate analyses should be considered.

Understanding the Reasons Leading to Missing Data

- Before any missing data remedy can be implemented, the researcher must first diagnose and understand the missing data processes underlying the missing data.

- Sometimes these processes are under the control of the researcher and can be identified. In such instances, the missing data are termed *ignorable*, which means that specific remedies for missing data are not needed.

Ignorable Missing Data Process

- One example of an ignorable missing data process is the "missing data" of those observations in a population that are not included when taking a sample. The purpose of multivariate techniques is to generalize from the sample observations to the entire population, which is really an attempt to overcome the missing data of observations not in the sample. The researcher makes this missing data ignorable by using probability sampling to select respondents. Thus, the "missing data" of the *nonsampled* observations is ignorable.

- Another instance of ignorable missing data occurs when the data are censored. *Non Available data* are observations not complete because of their stage in the missing data process. A typical example is an analysis of the causes of death. Respondents who are still living cannot provide complete information (i.e., cause or time of death) and are thus censored.

The Reasons Missing Data Occurs

- Missing data can occur for many reasons and in many situations. One type of missing data process that may occur in any situation is due to procedural factors, such as *errors in data entry* that create invalid codes, failure to complete the entire questionnaire, or even the morbidity of the respondent. In these situations, the researcher has little control over the missing data processes, but some remedies may be applicable if the missing data are found to be random.

- Another type of missing data process occurs when the response is inapplicable, such as questions regarding the years of marriage for adults who have never been married. Again, the analyses can be specifically formulated to accommodate these respondents.

- Other missing data processes may be less easily identified and accommodated. Most often these are related directly to the respondent. One example is the refusal to respond to certain questions. This is common in questions of a sensitive nature (e.g., income) or when the respondent has no opinion or insufficient knowledge to answer the question. The researcher should anticipate these problems and attempt to minimize them in the research design and data collection stages of the research. However, they still may occur, and the researcher must now deal with the resulting missing data.

Example of Missing Data

Case ID	V_1	V_2	V_3	V_4	V_5	Missing Data by Case: Number	Missing Data by Case: Percent
1	1.3	9.9	6.7	3.0	2.6	0	0
2	4.1	5.7			2.9	2	40
3		9.9		3.0		3	60
4	.9	8.6		2.1	1.8	1	20
5	.4	8.3		1.2	1.7	1	20
6	1.5	6.7	4.8		2.5	1	20
7	.2	8.8	4.5	3.0	2.4	0	0
8	2.1	8.0	3.0	3.8	1.4	0	0
9	1.8	7.6		3.2	2.5	1	20
10	4.5	8.0		3.3	2.2	1	20
11	2.5	9.2		3.3	3.9	1	20
12	4.5	6.4	5.3	3.0	2.5	0	9
13					2.7	4	80
14	2.8	6.1	6.4		3.8	1	20
15	3.7			3.0		3	60
16	1.6	6.4	5.0		2.1	1	20
17	.5	9.2		3.3	2.8	1	20
18	2.8	5.2	5.0		2.7	1	20
19	2.2	6.7		2.6	2.9	1	20
20	1.8	9.0	5.0	2.2	3.0	0	0
MISSING DATA BY VARIABLE						TOTAL MISSING VALUES	
Number	2	2	11	6	2	Number: 23	
Percent	10	10	55	30	10	Percent: 23	

Descriptive Statistics for the Observation

	Number of Cases with Valid Data	Mean	Standard Deviation	Missing Data	
				Number	Percent
X_1 Delivery speed	45	4.0133	.9664	19	29.7
X_2 Price level	54	1.8963	.8589	10	15.6
X_3 Price flexibility	50	8.1300	1.3194	14	21.9
X_4 Manufacturer image	60	5.1467	1.1877	4	6.3
X_5 Overall service	59	2.8390	.7541	5	7.8
X_6 Salesforce image	63	2.6016	.7192	1	1.6
X_7 Product quality	60	6.7900	1.6751	4	6.3
X_9 Usage level	60	45.9667	9.4204	4	6.3
X_{10} Satisfaction level	60	4.7983	.8194	4	6.3

Graphical Display of Missing Data

Case	Number of Missing Values	Variables: Missing Data X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}
202	2	S		S						
203	2		S					S		
204	3	S		S						S
205	1			S						
207	3	S		S						S
213	2		S	S						
216	2	S				S				
218	2	S				S				
219	2							S	S	
220	1		S							
221	3	S		S				S		
222	2			S		S				
224	3	S	S						S	
225	2			S	S					
227	2		S						S	
228	2	S			S					
229	1					S				
231	1							S		
232	2	S	S							
235	2						S			S
237	1		S							
238	1	S								
240	1	S								
241	2			S		S				
244	1								S	
246	1				S					
248	2	S	S							
249	1		S							
250	2	S		S						
253	1	S								
255	2	S		S						
256	1	S								
257	2		S	S						
259	1	S								
260	1	S								
267	2			S	S					
268	1									S
269	2	S		S						

Missing Data Patterns

Number of Cases	Missing Data Patterns[a]								
	X_6	X_{10}	X_4	X_7	X_9	X_5	X_2	X_3	X_1
26									
1								X	
4								X	X
6									X
1			X						X
1			X						
2			X					X	
2						X		X	
1						X			
2						X			X
2							X		X
3							X		
2							X	X	
1				X			X		
1				X					
1				X	X				
1					X				
1					X		X		
1					X		X		X
1		X							
1	X	X							
2		X						X	X
1				X				X	X

Diagnosing the Randomness of the Missing Data Process

- To decide whether a remedy for missing data can be applied, the researcher must first ascertain the degree of randomness present in the missing data. Some methods are available for this diagnosis like assessing the missing data process of a single variable Y by forming two groups-observations with missing data for Y and those with valid values of Y.

- Statistical tests are then performed to determine whether significant differences exist between the two groups on other variables of interest. Significant differences indicate the possibility of a nonrandom missing data process.

- Let us use the example of household income and gender. We would first form two groups of respondents, those with missing data on the household income question and those who answered the question. We would then compare the percentages of gender for each group. If one gender (e.g., males) was found in greater proportion in the missing data group, we would suspect a nonrandom missing data process. The researcher should examine a number of variables to see whether any consistent pattern emerges. Remember that some differences will occur by chance, but any series of differences may indicate an underlying nonrandom pattern.

Approaches for Dealing with Missing Data

1) Use of Observations with Complete Data Only

- The simplest and most direct approach for dealing with missing data is to include only those observations with complete data, also known as the complete case approach. This method is available in all statistical programs and is the default method in many programs.

2) Delete Case(s) and/or Variable(s)

- Another simple remedy for missing data is to delete the offending case(s) and/or variable(s). In this approach, the researcher determines the extent of missing data on each case and variable and then deletes the case(s) or variable(s) with excessive levels. In many cases where a nonrandom pattern of missing data is present, this may be the most efficient solution. The researcher may find that the missing data are concentrated in a small subset of cases and/or variables.

Imputation Methods

3) Imputation methods

- A third category of remedies for handling missing data is through one of the many **imputation** methods. Imputation is the process of estimating the missing value based on valid values of other variables and/or cases in the sample. The objective is to employ known relationships that can be identified in the valid values of the sample to assist in estimating the missing values. However, the researcher should carefully consider the use of imputation in each instance because of its potential impact on the analysis.

3-1) Using All Available Information as the Imputation Technique

- The first type of imputation method does not actually replace the missing data, but instead imputes the distribution characteristics (e.g., means or standard deviations) or relationships (e.g., correlations) from all available valid values.

Imputation Methods

3-2) The Replacement of Missing Data

- The second form of imputation involves replacing missing values with estimated values based on other information available in the sample. There are many options, varying from the direct substitution of values to estimation processes based on relationships among the variables. *Case Substitution* In this method, observations with missing data are replaced by choosing another nonsampled observation. A common example is to replace a sampled household that cannot be contacted or that has extensive missing data with another household not in the sample, preferably very similar to the original observation. This method is most widely used to replace observations with complete missing data, although it can be used to replace observations with lesser amounts of missing data as well.

The Replacement of Missing Data

- *Mean Substitution*

✓ One of the more widely used methods, mean substitution replaces the missing values for a variable with the mean value of that variable based on all valid responses. In this manner, the valid sample responses are used to calculate the replacement value. The rationale of this approach is that the mean is the best single replacement value. This approach, although it is used extensively, has three disadvantages.

- *External Source Imputation*

✓ In this method, the researcher substitutes a constant value derived from external sources or previous research for the missing values. This is similar in nature to the mean substitution method, differing only in the source of the substitution value. This imputation method has the same disadvantages as the mean substitution method, and the researcher must be sure that the replacement value from an external source is more valid than an internally generated value, such as the mean.

The Replacement of Missing Data

- *Regression Imputation*

✓ In this method, regression analysis is used to predict the missing values of a variable based on its relationship to other variables in the data set.

- *Multiple Imputation*

✓ The final imputation method is actually a combination of several methods. In this approach, two or more methods of imputation are used to derive a composite estimate-usually the mean of the various estimates-for the missing value. The rationale of this approach is that the use of multiple approaches minimizes the specific concerns with any single method and the composite will be the best possible estimate.