Africa Institute of Project Management Studies

(AIPMS)

**Course:** Monitoring & Evaluation

**Level:** Postgraduate Diploma

**Course Assignment:** Module 5: Assignment

Assignment presented to Lecturer: Charles Kaduwa PHD

Prepared by Student: David Ambu George

**Adm No:** AIPMS/263/2019

**Place:** Nairobi - Kenya

30th Oct 2019

**Module 5 Questions:**

**Q1.** Explain the difference between data collection and data capture (10mrks)

**Data collection** is any process whose purpose is to acquire or assist in the acquisition of data. Collection is achieved by requesting and obtaining pertinent data from individuals or organizations via an appropriate vehicle. The data is either provided directly by respondent (self-enumeration) or an interviewer. Data collection also includes the extraction of information from administrative sources which may require asking the respondents permission to link to administrative records,

While **data capture** refers to any process that coverts the information provided by a respondent into electronic format. This is conversion either automated or involves staff keying the collected data (keyers). The key staff to data capture includes; receipt of forms, editing, querying, imputation, coding, conversions, verification and validation.

**Q 2:** Explain the benefits of correctly interpreting data in an M&E process. (5 mrks)

**Data interpretation**: is the process of making sense of the information. It allows us ask: what does this information tell me about the program?

Once we transformed data into information by summarizing them with tables, graphs, or narrative, we need to interpret the data. That is, we need to consider the relevance of the findings to our program – the potential reasons for the findings – and possible next step.

Therefore, interpreting data correctly will adds meaning to our M & E process by making connections and comparisons to program, secondly, when data is correctly interpreted it helps us to conduct further research, and thirdly, good interpretation of data also helps to measure the performance of a program, hence resulted to a smooth decision making.

**Q3.** Explain the main concerns for a data analyst while undertaking the task of data analysis. (10 mrks)

**Data requirement:**

The data are necessary as inputs to the analysis which is specified based upon the requirements of those directing the analysis and customers (who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g. age & income) may be specified and obtained. Data may be numerical or categorical (i.e. a text table for numbers).

**Data Collection:**

Data are collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloaded from online sources, or reading documents.

**Data Processing:**

Data initially must be processed or organized for analysis. For instance, these may involve placing data into rows & columns in a table format (i.e structured data) for further analysis, such as within a spreadsheet or statistical software.

**Data cleaning:**

Once processed and organized, the data may be incomplete contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data entered and stared. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, reduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques.

**Exploratory data analysis:**

Once the data are cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to be begin understanding the message contained in the data. The process of exploration may result into additional data cleaning or additional request for data, so those activities may be iterative in nature. Descriptive statistics, such as the average or median, may be generated to help understand the data. Data visualization may also be used to examine the data in a graphical format to obtain additional insight regarding the messages within the data.

**Data Product:**

A data product is a computers application that takes data inputs and generates output, feeding them back into the environment. It may be based on a model or algorithm. An example is an application that analyses data about customer purchasing history & recommends other purchases the customer might enjoy.

**Communication:**

Once the data are analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.

When determining how to communicate the results, the analyst may consider data visualization techniques to help clearly and efficiently communicate the message to the audience.

**Q4.** Describe key measures that are mandatory for data quality assurance at program level and explain the value of data quality assurance. (15 mrks).

**Data quality:** refers to the state of qualitative or quantitative pieces of information. There are many definitions of data quality but data is generally considered high quality if it is “fit for (its) intended uses in operations, decision making and planning.

**Data quality assurance:** is the process of data profiling to discover inconsistencies and other anomalies in the data as well as performing data cleansing activities (e.g. removing outliers missing data interpolation) to improve the data quality.

**This question explains the below key measures and values that are mandatory for data quality assurance at program level:**

**Accuracy**

Refers to business transactions or status changes as they happen in real time. Accuracy should be measured through source documentation (i.e., from the business interactions), but if not available, then through confirmation techniques of an independent nature. It will indicate whether data is void of significant errors.

A typical metric to measure accuracy is the ratio of data to errors, that tracks the amount of known errors (like a missing, an incomplete or a redundant entry) relatively to the data set. This ratio should of course increase over time, proving that the quality of your data gets better. There is no specific ratio of data to errors, as it very much depends on the size and nature of your data set – but the higher the better of course. On the example below, we see that the data to error rate is just below the target of 95% of accuracy:

**Consistency**

Strictly speaking, consistency specifies that two data values pulled from separate data sets should not conflict with each other. However, consistency does not automatically imply correctness.

An example of consistency is for instance a rule that will verify that the sum of employee in each department of a company does not exceed the total number of employee in that organization.

**Completeness**

Completeness will indicate if there is enough information to draw conclusions. Completeness can be measured by determining whether or not each data entry is a “full” data entry. All available data entry fields must be complete, and sets of data records should not be missing any pertinent information.

For instance, a simple quality metric you can use is the number of empty values within a data set: in an inventory/warehousing context, that means that each line of item refers to a product and each of them must have a product identifier. Until that product identifier is filled, the line item is not valid. You should then monitor that metric over time with the goal to reduce it.

**Integrity**

Also known as data validation, integrity refers to the structural testing of data to ensure that the data complies with procedures. This means there are no unintended data errors, and it corresponds to its appropriate designation (e.g., date, month and year).

Here, it all comes down to the data transformation error rate. The metric you want to use tracks how many data transformation operations fail relatively to the whole – or in other words, how often the process of taking data stored in one format and converting it to a different one is not successfully performed. On our example below, the transformation error rate is represented over time:

**Timeliness**

Timeliness corresponds to the expectation for availability and accessibility of information. In other words, it measures the time between when data is expected and the moment when it is readily available for use.

A metric to evaluate timeliness is the data time-to-value. This is essential to measure and optimize this time, as it has many repercussions on the success of a business. The best moment to derive valuable information of data is always now, so the earliest you have access to that information, the better.

Whichever way you choose to improve the quality of your data, you will always need to measure the effectiveness of your efforts. All of these data quality metrics examples make a good assessment of your processes, and shouldn’t be left out of the picture. The more you assess, the better you can improve, so it is key to have it under control.

**Q5:** In about 350 words, describe the main challenges to effective data interpretation and analysis. (10 mrsk)

Data interpretation is the process of making sense of the information. It allows us to ask; what does this information tell me about the program. Therefore, certain data interpretation faces challenges or “pitfalls” that can occur when analyzing data as stated below:

**Correlation mistaken for causation:** our first misinterpretation of data refers to the tendency of data analyst to mix the cause of a phenomenon with correlation. It is the assumption that because two actions occurred together, one caused the other.

**Unreliable information**

A manufacturer thinks that they know the exact location of the truck transporting their finished products from the production site to the distribution center. They optimize routing, estimate delivery time, etc. And it turns out that the location data is wrong. The truck arrives later, which disrupts the normal workflow at the distribution center. Not to mention routing recommendations that turned out useless.

**Incomplete data**

Say, you are working to optimize your supply chain management. To assess suppliers and understand which ones are disciplined and trustworthy and which ones are not, you track the delivery time. But unlike scheduled delivery time, the actual delivery time field is not mandatory in your system. Naturally, your warehouse employees usually forget to key it in. Not knowing this critical information (having incomplete data), you fail to understand how your suppliers perform.

**Ambiguous data interpretation**

A machinery maintenance system may have a field called “Breakdown reason” intended to help identify what caused the failure. Usually, it takes the form of a drop-down menu and includes the “Other” option. As a result, a weekly report may say that in 80% of cases the machinery failure was caused by the “Other” reason. Thus, a manufacturer can experience low overall equipment efficiency without being able to learn how to improve it.

**Duplicated data**

At a first glance, duplicated data may not pose a challenge. But in fact, it can become a serious issue. For example, if a customer appears more than once in your CRM, it not only takes up additional storage but also leads to a wrong customer count. Additionally, duplicated data weakens marketing analysis: it disintegrates a customer’s purchasing history and, consequently, makes the company unable to understand customer needs and segment customers properly.

**Outdated information**

Imagine that a customer once completed a retailer’s questionnaire and stated that they did not have children. However, time passed – and now they have a newborn baby. The happy parents are ready to spend their budget on diapers, baby food and clothes, but is our retailer aware of that? Is this customer included in “Customers with babies” segment? No to both. This is how obsolete data may result in wrong customer segmentation, poor knowledge of the market and lost profit.

**Late data entry/update**

Late data entries and updates may negatively affect data analysis and reporting, as well as your business processes. An invoice sent to the wrong address is a typical example to illustrate the case. And to spice the story up even more, here’s another example on asset tracking. The system can state that the cement mixer is unavailable at the moment only because the responsible employee is several hours late with updating its status.

Irene Mikhailouskaya (Makaranka) 2018