The DATA COLLECTION chapter covers several topics:

- Overview
- Collection Techniques -- Data Entry, Formats
- Data Sources
- MS Access Database

### Overview

This section provides an overview of how to collect PANCEA data. It is intended to help you plan site visits with your primary contact(s), and to guide your day-to-day data collection schedule.

### Contents

- What type of data collection will we conduct? Abbreviated vs. full data collection sites.
- Where will we go to collect data? The sampling plan.
- How will the data collection proceed? Scheduling on-site visits.
- How should data collection begin and end? A few notes on starting and finishing the data collection.

### Abbreviated vs. Full Data Collection

We will collect cost and output information from all PANCEA sites. At most sites, we will collect data only for a short period of time—specifically, for the last month and for the most recently completed fiscal year. This is an “abbreviated” data collection.

We will visit a few sites for a “full” (or “intensive”) data collection. During these visits, will want to understand fluctuations in costs and outputs for the entire life of the intervention(s), and to understand the factors understanding variation in efficiency. These sites will represent the full range of intervention types, and a range in efficiency based on data from abbreviated visits. See the “Timing” section below for additional information on the timing of collecting data from abbreviated and full data collection sites.

The table below provides a description of abbreviated and full sites for each country; it also lists the PANCEA instruments that are used for each type of data collection.
**Sampling Plan**

The PANCEA sampling plan has evolved out of discussions regarding sample representativeness, statistical power, and practical limitations. Creating a sample plan has been arduous because our data collection goals are ambitious: We want to gather extensive expenditure and output data from 8 different types of HIV prevention interventions spread across 5 countries, suitable for two different kinds of analyses (see the manual section “Conceptual Approach”). At the same time, we have limited resources in the form of modest sample size (~200 facilities), considering the number of variables we want to study. It goes without saying that we also have limited time and money, which brings up a host of practical considerations.

Our current sampling plan balances the ideal of a randomized or representative sample with the realities of collecting detailed information that provides at least some opportunity for cross-national
comparison. We start with a general description of the sampling approach (section 1), and then discuss the specific composition of the sample (section 2).

1. General Description: high generalizability plus some variation across countries
The PANCEA sampling plan is a compromise solution that allows our econometric sample to reflect, as much as possible, the universe of existing HIV intervention programs in 5 countries. Our plan is to select one “standard” intervention to sample consistently across countries, and one “country” intervention within each country that is most suited to econometrics.

The primary advantage of this approach is that we combine one effort at comparisons across countries with several independent efforts at within-country analyses. For the “standard” intervention, we have a cross-national test of econometric methods. For the “country” interventions, we obtain important information and demonstrate econometric feasibility in one setting. The within-country analyses may do well at combining methodological suitability with generalizability for that type of intervention.

There are a few disadvantages. First, we are unable to obtain cross-national, econometric samples for each of our 8 interventions of interest. However, with our second analytic method (individual-program CEA) we can document cost per output and generate hypotheses on the correlates and causes thereof across all 8 interventions and multiple countries. Second, it is desirable to gather substantial amounts of information in choosing “country” intervention (e.g., number of programs, level of integration), though reasonable choices may be possible with relatively limited information. Third, we are putting all our eggs into one basket for the “standard” intervention; it is imperative that we select an intervention that is considered as essential and effective in curbing the HIV/AIDS epidemic.

See “Sample Composition” section below for a detailed breakdown of abbreviated/full data collections within each country.

VCT: the “standard” intervention
We have selected VCT as our “standard” intervention. We have made this choice for the following reasons:

- Previous research has shown that VCT programs have had a significant impact on the HIV/AIDS epidemic. Evidence for STI programs, another option, is more mixed.
- VCT exists in adequate numbers across all PANCEA countries.
- VCT links to treatment, as well as to prevention.
- VCT programs are less likely than some of our other interventions to be highly integrated into much larger programs or facilities. For example, counseling is more likely to be done by only a few staff than is STI management. Highly integrated programs are harder to effectively assess using econometrics, due to the prevention intervention outputs being dwarfed by other, varied outputs. (While truly stand-alone programs are easiest for econometrics, they are generally not the predominant mode of programs with adequate sample size across countries, e.g., SW programs.)
“Country” Interventions

A “country” intervention will be selected within each country. This intervention should ideally have a natural distribution amenable to econometric analysis. Criteria in selecting this intervention include the following:

- The intervention approach has had, or is likely to have, a significant impact on the HIV/AIDS epidemic within the country.
- There is an adequate number of existing intervention programs from which to select a sample.
- Programs tend to be stand-alone, or housed in units that are operationally separate from the larger facilities where they are located.
- The prospective intervention overlaps with the selection made in another country. While not necessary, econometric analyses will be helped by having some overlap among the country interventions since this will create a larger sample size. All else being equal, redundancy is preferable.

Drawing upon existing information, possible “country” interventions are as follows: (To date, these have not final been finalized.)

- Mexico – SW, with a heavy STI component.
- Uganda - micro IEC or MTCT
- Russia – risk reduction
- South Africa – SW or STI
- India – SW or STI
2. Sample Composition
There will typically be ~37 sites per country (Russia may have more, Uganda fewer), split along two dimensions. These are as follows.

The First Dimension: type of intervention
As described above, our focus is on two interventions within each country – the Standard Intervention (VCT) and the Country Intervention. However, we will collect data on 7 different intervention types within each country. (Note that we have included 8 different intervention types in the PANCEA study. These are: CSM, IEC, MTCT, RR (IDU), Schools, STI, SW, VCT. Each country will typically use 7 of the 8.) The numerical breakdown on this dimension is typically as follows:

<table>
<thead>
<tr>
<th>Intervention Type</th>
<th>Number in Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCT</td>
<td>16</td>
</tr>
<tr>
<td>Country Intervention</td>
<td>16</td>
</tr>
<tr>
<td>Additional Type 1</td>
<td>1</td>
</tr>
<tr>
<td>Additional Type 2</td>
<td>1</td>
</tr>
<tr>
<td>Additional Type 3</td>
<td>1</td>
</tr>
<tr>
<td>Additional Type 4</td>
<td>1</td>
</tr>
<tr>
<td>Additional Type 5</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>37</strong></td>
</tr>
</tbody>
</table>

The Second Dimension: type of data collection
In addition to collecting information on multiple types of HIV prevention interventions, the PANCEA project also uses multiple data collection and analysis methods. After collecting data, we will use three techniques: econometrics, simple cost-effectiveness analyses (CEA), and intensive program analyses combining CEA and limited qualitative data. However, it is unreasonable to collect information for all analyses at every site. For this reason, we will collect the abbreviated information we need for econometric analyses and simple CEAs at EVERY SITE. In a SUBSET of the sample – the “full” sites – we will collect additional information for use in the intensive program analyses.

We will also conduct several more formal case studies, like the one we did in Uganda, but the sample is not simply prevention programs, and will be developed and discussed separately. Specifically, 11 of the 37 sites in each country will be full sites.
Timing

Clearly, more time and effort will be expended at full sites. Additionally, we have few full sites – only 11 in each country and 55 across all countries. Consequently, we want to select full sites carefully so that we have data on the most interesting programs. For example, we want to select sites that have variable cost-effectiveness. We also hope to select programs that have interesting histories.

We know we can make the best selections when we have substantial amounts of information about prospective full sites. For this reason, we plan to collect data in two tiers. This plan is as follows:

Tier - I

- Collect econometric data (HIPPI-I, ARQ, Exit Interviews) for all Standard Intervention and Country Intervention sites. These are also called “abbreviated sites.” (typically ~32 sites).

Tier - II

- Select 3 Standard Intervention sites and 3 Country Intervention Sites for full data collection. Return to these sites to administer the ADC and HIPPI-II. (6 sites) The information to make the judgment about which is likely to justify an full effort will be derived from the ARQ and from HIPPI-I, particularly the last three questions of HIPPI-I which ask about the major accomplishments, challenges and external circumstances that determine the intervention’s ability to get clients and deliver services. We will select three Standard Intervention sites and three Country Intervention sites for full data collection. These will include one intervention that appears, based on ARQ data, to be relatively efficient; one that appears to be inefficient; and one that is of about average efficiency.

- Select 5 additional sites for examination of the 5 additional intervention types. Administer all instruments (HIPPI-I, ARQ, ADC, HIPPI-II, exit interviews) at each of these sites. (5 sites). These can occur during Tier I activities if most practical.
The graphic below represents PANCEA’s two sampling dimensions and the timing of data collection, with a typical distribution of number of sites.

Note that the numbers above create a sampling template; there are several likely variations. The sample might be split unevenly between interventions, depending on available sites. In some countries (e.g., Uganda), the sample might be smaller, and in others (e.g., Russia) larger.

Make a Plan, Stan¹: scheduling on-site visits
Communicating with your primary site contact prior to your first day of data collection will help you decide whether it is feasible to collect PANCEA data from a prospective site. It will also set the stage for you to conduct an efficient and considerate data collection.

Please refer to the manual section “Working with a Facility”, which contains several support documents for this stage of data collection. Documents include an interview protocol for your initial conversation with each facility’s primary contact, guidelines for determining whether the facility should be included in your sample, and documents that you can use as you establish day-to-day data collection schedules.

Data Collection Timeline
The data collection timeline depicted in the following figure is an ideal representation; we created it based on the design of our instruments, and not on practical considerations such as the day-to-day schedules of your respondents. Therefore, please use this graphic as a guide to:

- Determine which instruments should be administered at abbreviated/full data collection sites.

¹ With thanks to Paul Simon … “50 Ways to Leave Your Lover” (Still Crazy After All These Years, 1976).
- Identify the order in which instruments should be administered.
- Let your respondents know how much time you need to complete each instrument.
- Develop a day-by-day schedule with appropriate respondents. (See the instrument descriptions below to help you determine who appropriate respondents may be.)

A Few Notes
- You may want to share the “Data Collection Timeline” (DCT) figure with your primary site contact so that s/he can get a sense of how the data collection will proceed.
- Even though the DCT figure indicates that a full data collection will take 5 days, you will typically not visit a site on 5 continuous days. Instead, the PANCEA preference is to do one visit for the abbreviated collection, and a return visit some weeks later for some sites for the full visit. This spreads the data collection burden over time, facilitates digesting the data from the abbreviated visit to inform further discussion, and permits us to select an interesting sample for the intensive/full data collection.
- However, depending on logistics and your respondents’ preferences, you might want to complete all instruments for a full data collection during one week. This will be done in consultation with UCSF, to assure that it doesn’t compromise the sampling plan.
- Most aspects of the day-to-day data collection schedule are flexible, and should be adjusted to reflect respondent availability. At the same time, the time required to administer each instrument should approximate the schedule illustrated in the DCT figure.

Data Collection Timeline

Note: For most full sites, the “abbreviated” visit will stop after 2-3 days, and the remaining parts will occur weeks later if the site is chosen for full data collection. See text above.
Instrument descriptions

HIPPI, part I

**Time required:** 1-1.5 hours.
**Focus:** Interventions.
**Records:** Generally unnecessary.
**Appropriate Respondents:** Should have broad experience of the facility and intervention(s) over the entire course of intervention(s), if possible. This will typically be the director or a senior manager.
**Note:** The HIPPI part I has questions that should be administered once for each relevant intervention, up to two, at a facility. You may have a different respondent for each of these administrations.

ARQ non-intervention (ni)

**Time required:** ~4-6 hours.
**Focus:** Facility and interventions.
**Records:** Desired for expenditure information for the last month and the most recently completed fiscal year (by quarter or month if possible).
**Appropriate Respondents:** Has current facility-level and intervention-level knowledge of inputs (e.g., personnel, buildings, capital goods, and recurrent goods/services expenditures) and facility-level knowledge of outputs.
**Note:** The ARQni expenditure forms should be done after ARQint if there is any question of respondent fatigue (i.e., the same respondent), because the detailed outputs data in ARQint are critical “Forest” issues whereas the fine details of expenditures are not. Thus, a good order is: non-expenditure sheets in ARQni (e.g., governance), then ARQ intervention, then ARQni expenditures. It would also be ok to do ARQ intervention first, then ARQni.

ARQ intervention(s)

**Time required:** ~2-3 hours.
**Focus:** Intervention.
**Records:** Desired for outputs information for the last month and the most recently completed fiscal year (by quarter/month)
**Appropriate Respondents:** will have current knowledge of intervention(s) outputs (e.g., # of counseling sessions conducted, # of condoms distributed, etc.).

**Notes:**
-- One ARQ intervention sheet should be administered for each relevant intervention at a facility. You may have a different respondent for each of these administrations.
-- Whether a second additional ARQ intervention sheet is warranted depends on the extent of activity. For example, if an SW program conducts a limited amount of VCT, that is captured in the SW sheet with little detail. The VCT sheet is warranted only if the VCT activity comprises a significant portion of effort (e.g., more than 10%). Likewise, a VCT program with SW clients who get a bit of extra counseling would probably not need an SW sheet. *If in doubt, please consult with the UCSF PANCEA team.*

---

*For discussion of what constitutes the facility, see manual chapters “Definitions,” “Working with a Facility,” or “ARQni” (“Facility Expenditures” section).*
ADC non-intervention (ni)

- **Time required:** ~6-8 hours.
- **Focus:** Intervention.
- **Records:** Highly preferred for expenditure information across the entire lifespan of the intervention(s). Personnel records and records documenting recurrent goods/services expenditures are especially important.
- **Appropriate Respondents:** have accounting knowledge and access to records of the intervention(s)' expenditures, across the entire lifespan of the intervention(s), if possible.

ADC intervention(s)

- **Time required:** ~2-4 hours.
- **Focus:** Intervention.
- **Records:** Essential for output information across the entire lifespan of the intervention(s) (e.g., # of counseling sessions conducted, # of condoms distributed, etc.).
- **Appropriate Respondents:** will have direct knowledge of intervention outputs and access to relevant outputs records, across the entire lifespan of the intervention(s), if possible.

**Note:** One ADC intervention sheet should be administered for each relevant intervention at a facility. You may have a different respondent for each of these administrations.

HIPPI, part II

- **Time required:** ~2-3 hours.
- **Focus:** Intervention.
- **Records:** Unnecessary.
- **Appropriate Respondents:** have broad experience of the facility and intervention(s) over the entire course of intervention(s), if possible. This will typically be the director or a senior manager.

Exit Interview

- **Time required:** ~15 minutes each including consent, aim for 20 respondents.
- **Focus:** Intervention.
- **Records:** Not applicable.
- **Appropriate Respondents:** are clients of the intervention(s) being examined.

Starting and Finishing the Data Collection

There are several things you can do to keep your data collection on track.

- Discuss data collection goals at the beginning of each day you are on-site. This will help focus both your and your respondents’ efforts.

- Spend a few minutes at the end of each data day to summarize the work you have completed, and to discuss goals and confirm the schedule for the next day of data collection.
Prioritize developing rapport with respondents. You will be asking for a substantial amount of your respondents’ time. In addition, you will want to examine records with the help of site staff. This is a big favor to us, and can be a hassle to our respondents. Therefore, be considerate of your respondents’ obligations outside helping us with our data collection. Your schedule should work around your respondents’ schedule.

After you have said good-bye to your respondent(s), there are a few more loose ends you need to take care of before the end of the day.

- Review all collected data, making notes of any items that are unclear so that you can follow-up on them at some point on the following day. Ideally, data collectors can work together to achieve consensus on particular points, or to highlight areas that still need clarification. See the “Quality Assurance” manual section, which is a guide to this process.

- Ensure that all data are saved in a master file as well as in a back-up file.

- Record person-hours spent for each instrument and any additional comments on the facility report.

After you have finished the entire data collection at a site, the following tasks also need to be done.

- Complete facility report.

- Make a back-up copy of all data.

- Email finalized data set to UCSF.

- Complete summary report for site.

- Send the thank-you letter, summary report if available, and gift to your primary site contact.
Collection Techniques – data entry

This section describes appropriate procedures for entering data into the PANCEA database.

Contents:
- Two general techniques for successful data collection.
- What are the common complications of data collection?
- “Reasonable” and “acceptable” answers.
- How to differentiate between blanks, zeros, and “not applicables”.
- When to enter data.
- How to clarify complex, ambiguous, or conflicting responses.
- Crucial: the day-by-day data review.

Two general techniques for successful data collection.

- It is vital that respondent(s) feel comfortable throughout the data collection process, especially since the PANCEA data collection requires substantial amounts of time and attention to detail. Take the time to develop rapport and maintain an awareness of how your respondent is feeling – Is s/he tired? Does s/he have other obligations to take care of? A collaborative data collection effort will help keep your respondent(s) engaged in the data collection process.

- As the data collection occurs over several days, maintain an awareness of how information builds on itself. Some questions cover related issues, and therefore indicate reasonable answers to subsequent questions. Make sure you see this big picture as you collect a lot of detailed information, and maintain a colloquial style of questioning by referring back to previous questions. This will facilitate effective and efficient data collection.

What are the common complications of data collection?

The PANCEA data collection requires that you collect extensive qualitative and quantitative data. As you work through the HIPPI, ARQ, ADC, and exit interviews, there will be instances in which the following occur:

- The respondent doesn’t know the answer to the question.
- The question doesn’t apply to the intervention under investigation, and so must be skipped.
- The respondent gives long answers that are difficult to enter into forms completely.
- A comment needs to be made about an answer to clarify what the answer means, or what the answer does and does not refer to.
- Data are identical over time and need to be recorded for multiple time periods.
- The respondent gives a response that is inconsistent with another response.

In each of these situations, you will have to decide how to enter data so that your respondents’ answers are recorded clearly, correctly and completely. The following paragraphs outline appropriate procedures for doing this.
“Reasonable” and “acceptable” responses

In many cases, your respondent(s) will give you an answer to your question. However, this answer may not be in a form that is “reasonable” or “acceptable” for our database. There are four ways in which an answer may fail a quality assurance review.

- Wrong data type.
- Inadequate level of detail
- Possible but unlikely response.
- Inconsistent with another response.

Wrong data type
We may want an answer in number form, but a text response is given. The MS Access database prevents this from happening in many cases; however, refer to the relevant question-by-question guide in the manual to determine the data type that is expected.

Inadequate level of detail
We usually want specifics, but a vague response is given. For example, “we faced problems with regulations” is vague; “we couldn’t get a building permit because of zoning regulations” is specific. You will need to decide when to ask additional questions to obtain a more specific answer, and when more detail is unnecessary. In general, we want enough specifics to provide real insight into intervention operation. The best tools you have for making these judgments successfully is to understand the purpose of the PANCEA data collection, and having familiarity with each question.

Possible but unlikely response
For example, it is very unlikely that the number of people receiving intervention services is the same as the number of people in the area’s client risk group. It is more probable that the number receiving services is substantially smaller than the number who could receive services.

As another example, extremely high (90-100%) or low (5-10%) return rates after testing are unlikely. Whenever answers are possible but unlikely, it is important to verify their accuracy and to note in a comment that they have been verified. For example, “Verified; 95% return rate achieved due to very aggressive follow-up effort.”

Inconsistent with another response
For example, the number of people in the area’s client risk group is smaller than the number of people receiving intervention services. As another example, ARQcsm data show that 1,000 condoms were distributed in the previous fiscal year, but ADCcsm or HIPPI data indicate that 2,000 condoms were distributed in the same period. Another inconsistency is recording outputs (e.g., condoms, STI tests, etc.) that have no associated costs.

- As you collect data, consider whether your respondent’s answers make sense. Confirm and clarify answers as you go. This will help both you and your respondent(s) see the big picture.
- When you have determined that an unlikely or inconsistent response is correct, provide a comment to explain the entry.
- In addition, review the Quality Assurance Guidelines to find and troubleshoot unlikely and inconsistent responses.
Ideally, you will identify unlikely and inconsistent responses during your interviews and/or during your data review at the end of the data collection day. That way, you can clarify such answers with your respondents as soon as possible.

How to differentiate between blanks, zeros, and “not-applicables”

A blank cell is not a good idea … it can mean several things:

- **Not asked**—the question wasn’t asked at all (yet).
- **Not applicable**—the question was asked, but it didn’t apply to the intervention.
- **Don’t know**—the question was asked, but the respondent didn’t know the answer.
- **Zero**—the question was asked, and the answer was “zero” or “none”.
- **Repeated value**—the question was asked, and the answer was the same as the previous one, so the cell was left blank.

As you can imagine, it is VERY problematic to try to decipher what is meant by blank cells after the data collection has been completed. For this reason, *use the following rules in all cases!*

**Not (yet) asked**—the cell should remain blank. (All cells start out blank.) *By the end of your data collection, there should be no blanks – all questions should have been asked and all response fields filled – even if the answer is “not-applicable”.*

**Not applicable**—enter “n/a”.

**Don’t know**—enter a “dk” to indicate that the question was asked, but the respondent did not know the answer. In the case of date fields, enter all “1”s to indicate that the date is unknown (e.g., 11/1111). Note that if the respondent has a recall estimate or guess, that information can be recorded, with an appropriate note, under Data Source.

**Zero or none**—enter a zero (0) (not an “o”).

**Repeated value**—if possible, enter the correct values for these in real time. However, if this process is unreasonable—in other words, your respondent is idle as you copy and paste values—make careful notes for yourself so that you can fill in these data at the end of your data collection day.

*NOTE: These fields should be completed at the end of the relevant data collection day, at the very latest. This process becomes lengthy and error-prone if deferred until the end of the data collection week.*

**When to enter data**

As much as possible, *ENTER ALL DATA AS YOU GO*. Sometimes, this can be difficult because you will be talking with your respondent and asking for large amounts of information. In addition, some answers will repeat across multiple time periods, and these responses may have to be copied and pasted. However, for most, if not all of the PANCEA instruments, you will work in pairs. Therefore, one data collector can be the primary interviewer while the other data collector can focus on recording answers.

In the worst case scenario, it will not be reasonable to record all data and associated comments in real time. If this happens, make careful notes for yourself so that you can complete data entry
at the end of each data collection day. It is vital to wrap up data entry at the end of each data
collection day so that this process stays brief, and so that you don’t forget what your notes
mean.

HIPPI: Most of the data you enter during the HIPPI interviews is open-ended. This type of data
can be more difficult to record than the quantitative data you will be recording for the ARQ and
ADC because as respondents give you HIPPI information, you will need to create accurate
summaries as opposed recording answers verbatim. For this reason, it is possible that the
information you record is not what your respondents actually intended to say. At the end of each
relevant data collection day, you and the other data collector who worked on the HIPPI interview
should spend a few moments reviewing your data to achieve consensus on each summary. If
you can’t achieve consensus, and/or you find some answers to be ambiguous, you should plan
to ask your respondent(s) to clarify these issues as soon as possible. Although we will give
respondents an opportunity to review key findings, we can’t rely on them to do a thorough
review.

How to clarify complex, ambiguous, or conflicting responses

- If you are in doubt regarding what a particular question means, refer to the relevant
  section of the manual. Often, these instructions may guide your decision.

- Remember that the purpose of PANCEA is to examine the cost-effectiveness of HIV
  prevention programs. For example, PANCEA wants to obtain a list of all required staff for
  an intervention; so, an answer like “I know we had volunteers, but I don’t remember
  exactly how many,” is not the kind of information we want to obtain. Rather, an answer
  like this should prompt you to ask a couple more probing questions to perhaps obtain a
  clearer response. For example, “We had between 6 and 8 volunteers, I’m uncertain of
  the exact number.” Fine detail is not critical, but omitting a major expense is (see
  “Forest and Trees” in the first part of the manual).

- If the relevant manual section does not answer your question, do your best to be as
clear as possible with each answer. Provide enough comments and/or added detail so
  that you can discuss and resolve the issue with the PANCEA team later. This part of the
  process is also valuable because the feedback you give us helps us create protocols for
  the issues you come across.

- It is a good idea to repeat answers back to the respondent for confirmation whenever
  possible.
  - You can use this technique to ensure that your summaries of longer, qualitative
    responses are accurate in the main points: “So, to summarize, you said that your
    ability to distribute more condoms was due primarily to increased funding at that
time, but also to successful publicity campaigns. Is that correct?”
  - You can also use this technique for short, quantitative answers: “OK, so you said
    you purchased a car in the second quarter of 1995 for US $2,000?”

Crucial: the day-by-day data review

- At the end of each day of data collection, both interviewers should expect to spend some
time reviewing all data collected for the day to ensure that data are complete and
unambiguous. Some problems you may find are as follows:
You may find sections where the data are incomplete. Did you miss entering any data?

You may find cases in which you and the other data collector had different ideas about what the respondent meant. If you discuss these issues, you may be able to resolve the issue yourself by adding more notes to clarify the correct response.

In some cases, similar questions are asked across multiple PANCEA instruments, and these answers may conflict. This type of problem is difficult to find while you are on-site collecting data, but may be a prominent contradiction as you examine the big picture.

During this process, highlight answers that need to be clarified and/or completed by your respondent on the following day, or as soon as your respondent(s) have more time for you.

**NOTE:** Use the “QA Guidelines” to guide your review.
Collection Techniques - Formats

Most of the data collection should be administered in teams of two data collectors. One person will have primary responsibilities for asking the questions, while the second person will record the answers. This facilitates a smooth flow to the interview.

There are advantages and disadvantages to using a laptop computer, paper and tape recorder to record answers during the data collection. This section reviews these pros and cons. Ultimately, you will have to determine the method of data collection that works best for your team. However, the decision to use laptops, paper and/or tape recorders to record data should be made fairly early in the data collection process and thus remain as consistent as possible across data collection sites.

Contents:
- Laptop computers
- Paper questionnaires
- Cassette tapes

Laptop Computers
The advantages to using laptop computers are as follows:

- Laptops eliminate the need to transcribe answers following the interview.
- Laptops allow interviewers to scroll through and refer to data that have already been entered, facilitating a colloquial interviewing style. This is especially true as interviewers become more adept at navigating the database.
- Unlike paper questionnaires, laptops store the data in a secure (private) environment.

The disadvantages are as follows:

- Laptops may create a distraction or separation between interviewers and respondents. However, this distraction is minimized when two interviewers are participating—one person to focus on interviewing while the second person enters data.
- Laptops can not maintain a hard copy of data. Therefore, unlike hand-written information, it is possible to mis-enter data into the database.
- Unless care is taken to regularly save files, it is possible to lose data.

Using laptops is a suggested format for data collection. The final decision, however, should be made by each study team.

Paper Questionnaires
The advantages to using paper questionnaires are as follows:

- Paper questionnaires do not rely on technology, and allow interviewers to write additional notes directly on the questionnaire.
Respondents may feel more comfortable with paper questionnaires.

Paper questionnaires provide a quality-control measure of the electronic database since all information entered into the database can be double-checked against paper copies.

The disadvantages are as follows:

- Additional time is required after interviewing to enter data into the database, and to check that all data have been correctly entered into the database.
- It may be more difficult to “see the big picture” with paper as it may involve flipping through many sheets of paper to determine what respondents said previously to a related question.

Exit interviews must be administered with paper questionnaires, and while not required, this format may be well-suited to the HIPPI interview too.

**Cassette Tapes**

The advantages to using cassette tapes are as follows:

- Cassette tapes provide a recording of interviews that can used to resolve differences of opinion among data collectors of what was actually said during a key part of the interview.
- Cassette tapes can be used to collect verbatim, open-ended answers as opposed to summaries. This ensures that that respondent(s)’ meaning is not combined with interviewers’ interpretation as information is collected.
- Cassette tapes can serve as a back-up in case computer or paper files get lost or destroyed. However, the tape recorder should never substitute for paper or computer data entry.

The disadvantages are as follows:

- Respondents may not feel comfortable being recorded. It is essential that respondents feel at ease so that we can collect candid and truthful responses.
- Data entry requires considerably more time since it can not be done during the interview process.
- Cassette tapes and recording equipment are an added expense. In addition, they require considerable storage space.
Data Sources

The PANCEA project collects detailed expenditure and outputs information, sometimes across multiple years. Consequently, we have to rely on multiple data sources to obtain complete data. Sometimes the data we want will be recorded in program records. Other times, information from records provide a partial answer, and these data have to be adjusted in some way to find the data we need. In yet other instances, the information we want does not appear in records at all; in these situations, we rely on respondent(s)' memories.

It is important to keep track of data sources for a couple reasons. First, we will need to write up a methods section in each of our papers. A clear record of data sources will allow us to develop a map of how data were collected across many sites – pure records data, a combination of data sources, recall data, etc. Second, while all data contains some error, it may be the case that some data sources contain more error than others. For example, records may be a more accurate representation of intervention outputs than respondent(s)' recollections. Although we can never be sure of just how much data error exists, we want to keep track of the sources used for data collection so that we can examine this issue later on.

This section outlines the procedure for recording data sources.

One data source field exists for groups of related questions in every PANCEA instrument. This is an open-ended field in which you can describe the data sources used for the relevant questions. We have created a classification system to describe record sources; these are as follows:

- **SR**: Written summaries or reports whose numbers are used directly. Examples: Recurrent spending from audited report; annual payroll reports used for salaries; reports to funder on number of HIV tests.
- **SR-A**: Written summaries or reports adapted to our data needs: Example: Funding history translated to %s by type of funder.
- **RR**: Raw written records whose numbers are used directly. Examples: Complete registers of salary or per diem payments; simple sum across 3 months to generate quarterly data; invoices for condom purchases, including number purchased; register of HIV tests completed.
- **RR-R**: Raw written records informed by recall. Example: Incomplete registers of salary or per diem payments; invoices for condom purchases, with recall for number purchased and thus price.
- **WP**: Written policies/protocols. Example: Counseling protocol.
- **RO**: Recall only. Example: Mix of types of clients; percent effort on 2 interventions; supply outages; STI follow-up return rates.
- **EE**: Estimation extrapolated from similar data. Example: Salary to hire someone to do volunteer’s work, based on a similar employee’s salary.
- **Guess**: Really rough estimate. No basis in data.
Try to classify all data sources into these categories. However if you encounter sources of data that do not fit, do not try to make them fit. Instead, describe the data source you used with explanations describing its possible inaccuracies. We will improve this data classification system as we gain experience. The information you provide on the data sources you find is important to this process.

Recording data sources should not take more than a minute per section. To make this process as quick and easy as possible use the classification “codes” above as a shorthand. For example, instead of writing out “estimation extrapolated from similar data”, you can write “EE”.

Sometimes, one section may have been informed by several data sources; record all data sources for each section.

Data source information is useful when it is specific. For example, “mix of records and recall” is too vague. This answer does not indicate which data points came from records and which came from recall. A better description is something like, “Titles, salaries, and hire periods are SR. Market pricing is RO. Allocations are RR-R.”

Complete data source fields immediately after collecting data for each section, otherwise you might forget how data were collected.
MS Access Database

PANCEA Data Collection: MS Access (status as of May 2003)

- The MS Access database is designed and partially implemented.
- As of this writing, it houses the ARQni, ARQvct and ARQsti instruments. ARQsw will be done soon. All other instruments should be administered in either MS Excel or MS Word.

Contents
- Introduction and basic terms
- Navigation

Introduction and basic terms

Primary Terms

The database is composed of two basic layers.

- The top layer allows users to enter data, using “forms”. The bottom layer stores all the data in “tables”.
- The bottom layer is the control center for the top layer. You will not need to know much about the bottom layer; you may actually never see this layer. However, some bottom layer terms are included below so that you can have a conceptual understanding of what happens to the data that you enter into the database.

Top Layer (what you see)

Forms are what you work with when you enter information into the database. The first form you see in our database (below) is the form to select a PANCEA instrument to examine or edit.

Forms come in many different sizes. For example, the form below has multiple pages, each accessed by a button on the navigation bar to the right of the form.
Forms can be formatted in several ways. For example, the “Output Details” form below “pops-up” when you press the “Last FY” buttons on many of the “Services” pages. This form pops-up so that you can enter data for an entire year, in addition to the “last month”.

---

<table>
<thead>
<tr>
<th>Time and Place</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What “last month” is covered by the information provided on this sheet?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month (mm/yyyy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start Date (mm/yyyy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End date (mm/yyyy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. What “previous fiscal year” is covered by the information provided on this sheet?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCT Start Date (mm/yyyy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. When did you begin providing VCT services?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Where were your VCT activities conducted in the last month?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In a health facility with other services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In a stand-alone facility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In a mobile unit (e.g., van)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In another type of location (Specify)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last month</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (by the same for All of the last 12 months)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If (yes) When did you make the change?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

PANCEA Manual

Data Collection Guidelines
Forms allow you to enter data in a variety of ways. In some cases, you have the freedom to enter any information you like; in other cases, the form controls what you can and can’t enter.

All information is entered in one of the following ways:

**Option buttons**
These allow you to select from a list of choices that appear on the form. A primary example of option buttons is on the instrument menu (see above).

**Combo boxes**
These are white rectangles with a down arrow on the right-hand side. If you click on the arrow, you will see a list of options. You can select one of these options, but you can’t enter additional information.

**Text boxes**
These are white rectangles that allow you to enter nearly anything you like, although data entry may be restricted to text or numerical responses, based on the expected answer to particular questions.
Command buttons help you navigate forms. The grey “go” rectangle is an example of a command button (See the instrument menu above.). When you click on this control, it will take you to another form.

Bottom Layer (what you don’t see)

So, what happens to all the information you enter? These data are stored in tables in the bottom layer of the database. You will not see tables as you enter data; however, if you are involved with analyses after our data collection is complete, you will work with these tables. Tables contain records and fields. A record is a row in a table. Often, a record is equal to all the information collected for a certain facility. A record is composed of numerous fields. A field is a column in a table. Each piece of information collected about a facility (record) is recorded in a field.

Navigation

Security—

1. Enter the password when you open the database. (If you have forgotten the password, ask your team leader.)
   a. This password is the same for everyone on the PANCEA team, and therefore only restricts non-PANCEA personnel from entering the DB.

Select an instrument—

1. Select the instrument for which you would like to enter or edit data by clicking on the radio button to the left of the instrument name.
2. Click “go” to navigate to the next form.

Select an intervention facility —

1. Click on the facility/intervention for which you would like to enter/edit data. Notice the name of the facility/intervention does not appear. This is to protect confidentiality. You must know the UNQID and RECID(s). These are the two ways in which each facility/intervention is identified.
   a. The UNQID refers to a facility; the RECID refers to an intervention in a facility. Each facility can have only one UNQID, but multiple RECIDs.

2. Remember the UNQID because you will need to re-enter this ID on the next screen.
3. Click on either “new data” to start data entry for an instrument or “edit data” to continue data entry for an instrument.

*Enter/edit data—*

1. Enter/edit all data you would like. Each section of the instrument is listed on the navigation bar to the right of the form; you can click on these to navigate to each page of the intervention form.
2. Note that there is a “comments” field for each question. Use these as much as you like.
3. Note that there is a “data source” field for each sub-section of the instrument. See the “Data Sources” section of the manual.
4. Save your data frequently by clicking on the “save” button in the upper left-hand corner, or hitting “control-S”. Your data are also saved when you click on the button “Exit Forms” to return to the instrument menu.