USAID
FROM THE AMERICAN PEOPLE

fhi Family Health International

First Things First:

Guidelines on Management and Coding of Behavioural Surveillance Data

Acknowledgements

This guide was conceived and written by Elizabeth Pisani.

Thanks to Peter Ghys and Sara Hersey for useful comments, to Emma Slaymaker for the section on dealing with weighted data, and to Ties Boerma. Special thanks are due to Arizal Ahnaf, Happy Hardjo, Mohammad Noor Farid and other staff at Indonesia's Central Bureau of Statistics, who contributed significantly to developing much of the experience on which this work is based.

ISBN: 974-93953-1-X

This document has been funded by the United States Agency for International Development (USAID) through the Implementing AIDS Prevention and Care (IMPACT) Project (Cooperative Agreement HRN-A-00-97-00017-00), which is managed by FHI.

The views expressed by the named authors are solely the responsibility of those authors and do not necessarily reflect the views of FHI or (USAID). The mention of specific companies or products implies no endorsement and does not suggest that they are recommended by FHI or USAID over others of a similar nature that are not mentioned.

List of Acronyms

AIDS	Acquired Immuno-deficiency Syndrome
BSS	Behavioral Surveillance Survey
FSW	Female Sex Workers
HIV	Human Immuno-deficiency Virus
IDU	Injecting Drug User
MSM	Men who Have Sex with Men
RTI	Reproductive Tract Infection
STI	Sexually Transmitted Infection
TR	Time Reference

CONTENTS

CONTENTS

Model Codebook for BSS among Populations with High Risk Behaviours

Chapter 1: Introduction

Great strides have been made in recent years in improving the scope, quality and appropriateness of surveillance systems for HIV in general, and of behavioural surveillance in particular. A large number of countries now regularly collect information on risk behaviour and exposure to HIV prevention programmes in sub-populations most likely to be affected by the epidemic.

In almost every country, this wealth of data is massively underused. Why? Because those working to improve surveillance have, quite appropriately, taken a “first things first” approach. Without good quality data there can be no good analysis, and without good analysis there can be no appropriate use of data to improve programme performance. And so countries have concentrated first on improving population selection, sampling procedures and fieldwork. The result: lots of high quality data. It is time to turn our attention to how to use those data. But again, we need a “first things first” approach. Before they rush to strengthen data analysis, countries must deal with the most neglected area of behavioural surveillance: data management.

Data management includes data entry, consistency checks and cleaning. It should also include the combination of data from different populations, sites and years into a single data set. This will allow analysts easily to compare variables across time, between cities, between population groups, etc. But it is not as easy as it sounds. It involves careful renaming of variables to remove variations between questionnaires, recoding variables so that denominators are consistent even if questionnaire skip patterns differ, recoding again to produce variables which are most likely to be used in analysis, while being careful to document work and ensure that data labels remain correct.

The easier a data set is to use, the more it will get used

This is a nearly universal rule of data analysis. And yet we invest comparatively little time in data preparation, preferring to launch into analysis as quickly as possible. As a general rule, once data have been entered and consistency checks have been performed, another two weeks of full-time work should be invested in combining, coding and labelling data sets. While this may sound like a lot, it will slash the time needed for analysis, as well as creating a “user-friendly” dataset which can be used widely and repeatedly to meet any number of emerging data needs.

Investing time in data management is like investing time in preparatory work in the kitchen. It takes time to chop garlic, press coconut milk, brew fish-stock and wash basil, and white these tasks don’t take years of training, they need to be performed correctly and with care. Once these tasks are performed by the sous-chef, it takes the chef no time at all to whip up a whole range of dishes.

About this document

This document is part of a wider series of guidance on HIV, STI and behavioural surveillance (collectively known as “Second Generation Surveillance) and population size estimation, developed over the last decade by UNAIDS, the World Health

Organisation and their many partners in surveillance. One of the goals of this series is to ensure that surveillance data get used extensively in managing HIV prevention programmes at the local level and in planning, managing and evaluating HIV prevention and care programmes at the national level.

The work of data management must precede the work of analysis, but it need not necessarily be done by the same individuals. Data entry, cleaning, and coding are extremely technical tasks which do not require an extensive knowledge of HIV programming, just as chopping garlic and washing basil do not require a lot of knowledge about how to make the perfect sauce. Choice of appropriate recodes for use in analysis – the brewing of fish stock in our kitchen analogy — does require more specialised knowledge. But it is only at the data analysis stage that we really need a qualified chef to provide a full understanding of HIV prevention and care programmes.

This document is for the sous-chef.

It aims to give technical guidance on **creating data sets that are easy to use**, so that

1) it is easier **to train people to analyse and use data** effectively

2) people are more likely to **use the data on an ongoing basis to manage and evaluate** their HIV programmes

It is hoped that this guidance will be useful in many countries . The document contains examples and exercises which can be used in training, and which can easily be adapted for use directly on local datasets. It is meant for data editors and managers rather than for HIV prevention programme analysts, although the latter will be the end users of the data sets created using this guidance.

This guidance deals very briefly with issues of questionnaire design and data entry (on which more specific guidance is available elsewhere.) It devotes greater attention to issues of renaming, recoding and labelling data sets – the building blocks of combined, user-friendly datasets.

The examples and exercises given in this version of the document are based on Stata software – the software most commonly used for analysis of survey data by statistical bureaus.[1] It presupposes a basic working knowledge of Stata, though details on getting started are provided in Appendix 2. If the commands used in Example 1 on page 16 are not familiar, or if you do not know what a “do file” is, please read Appendix 2 before proceeding. Versions of this document are also being made available for users of other software packages.

The exercises in the text of this document can be done by simply writing code with a pen and paper, rather than by using Software. This encourages participants to think about the content of the exercise, rather than getting stuck on syntax issues.

This guidance was developed for countries which collect data from a number of different sub-populations with high risk behaviour, typically female, male and transvestite sex workers, men employed in occupations where buying sex is common, men who have sex with men, and drug injectors. The examples given focus on these groups. However the principles discussed are equally relevant to behavioural surveys in general, including data collected from groups representing the general population.

[1] This document is not intended as an endorsement by WHO, UNAIDS or any of their partners of any single commercial software package. The document is being prepared for use with a variety of commonly used commercial statistical software packages.

Users will notice that the guidelines contain a model codebook (Appendix 3), which suggests specific names and recodes for common variables. This is **NOT** intended to be an internationally standardised list. Countries will need to vary this list according to their own epidemic and programme needs, questionnaire formats and language preferences. The most essential points are that:

- consistency is maintained within a country between populations and over time.
- variable definitions, including variations from "international standards", are clearly documented

However, since many countries regularly provide data to international organisations in the course of reporting on UNGASS indicators and on other international agreements, it is worth considering using these formats unless there is a locally specific reason not to.

Chapter 2: Behavioural surveillance questionnaires

A good questionnaire is easy to understand (from the respondent's point of view) and easy to administer (from the interviewer's point of view). The answers are easy to read and code (from the data editor's point of view) and their meaning is clear (from the analyst's point of view). From the point of view of the programme managers and planners who are one of the most important end users of surveillance data, questionnaires need to provide information that will help explain the local epidemic and evaluate local HIV prevention and care programmes.

A note on standardising questionnaires internationally

There has been a great deal of debate in recent years about the extent to which behavioural surveillance questionnaires should be "standardised". International and bilateral agencies wishing to report similar indicators across countries often point to internationally standardised survey programmes such as the Demographic and Health Surveys (DHS) as a model for behavioural surveillance for HIV. DHS, which focuses on maternal and child health and contraception and which has included an AIDS module in some countries in recent years, uses very similar questionnaires to generate data sets which are then recoded to identical international standards. Reports and analyses are also largely standardised across countries.

There are a number of differences between DHS and similar surveys and behavioural surveillance for HIV, particularly in concentrated epidemics. DHS is paid for largely by a single bi-lateral donor, and technical assistance is provided largely by a single private company, which also has a large staff dedicated to data management. Increasingly, behavioural surveillance is an integral part of national HIV surveillance systems, designed, administered and sometimes paid for by national governments. Technical assistance is provided by a number of different agencies, and paid for by a number of different donors; none has dedicated data management staff. These factors combine to reduce the influence any agency can have in enforcing a "standard" questionnaire across different countries.

Where much behavioural surveillance takes place in the general population, as is the case in much of sub-Saharan Africa, questionnaires can more easily be standardised. However where HIV is driven by drug injection, sex between men, and commercial sex – almost all countries outside Africa — standard questionnaires are neither possible nor desirable. Sexual and drug-taking landscapes differ enormously between countries – many examples of the differences are given in the text below. These differences can determine what prevention and care programmes are appropriate. Since behavioural surveillance exists to meet the needs of national and local programme planners, questions must be appropriate to the local epidemic and response. Ironing out the differences in the interests of "standardising" questionnaires will iron out the utility of surveillance.

More information on questionnaire design is available in other documents in this series and elsewhere. The remainder of this section will focus only on those issues which have direct relevance for data management. Though many of the issues raised may seem to state the obvious, they appear here because errors in these areas continue to appear in national surveillance systems. Some of the questionnaire design errors are not fatal – indeed part of the reason we invest in recoding is to deal "after the fact" with inconsistencies in questionnaires. But to the extent that they can be avoided at the questionnaire design stage, it will make the task of data management much easier.

Standardising questionnaires nationally

The more similar questionnaires are, the easier it is for data management. The preceding box argues that it is not always possible or desirable to over-standardise questionnaires. And obviously, a questionnaire for drug injectors (which will have many questions about injecting behaviour) cannot be identical to a questionnaire for female sex workers (which will focus on sex with clients and condom use) or for male sex workers (which will contain additional questions about anal sex and lubricant use).

Having said that, however, there is much in the questionnaires that can easily be standardised, or rationalised to make the task of data coding and combining much easier. Rules of thumb include:

Organise questions into blocks, and keep block numbers the same

Questionnaires generally have various "blocks" of questions, for example socio-demographic background, HIV-related knowledge, sexual behaviour, drug-taking behaviour, STI experience and treatment-seeking, programme exposure, miscellaneous. Within each block, question numbering should begin again at 1, with the block number before it (e.g. Block 1: b1q1, b1q2…Block 2: b2q1, b2q2…; alternatively, more simply, Block 1: q101 q102… Block 2: q201 q202…)

It is very helpful to the data manager if these blocks have the same numbers across all questionnaires. **Unless there is a good reason** to change blocks around, do not switch the order or numbering of the blocks between questionnaires.

What is a recode file?

The phrase "recode file" appears frequently in this document. What is it?

In most data management and analysis software packages, users can write small computer programmes which tell the package what to do with the data. They lay out a series of tasks which are executed in a given order. In Excel these are called macros, and are written in a programme called Visual Basic. In Stata they are called do-files (see Appendix 2).

In this document, a recode file refers to a programme written by the data manager which performs all the tasks of combining, coding and naming BSS data sets so that they are ready for analysis and use. This includes all the tasks described in this document: adding new variables, changing the names of variables and the values of responses so that variables from different data sets can be combined, making sure that the denominators are clear and standardised between data sets, etc. Once written, recode files can be used repeatedly, and can be modified over time as necessary to cope with an expanding or evolving behavioural surveillance system.

Within blocks, maintain as much uniformity as possible

There are often whole blocks which can be kept identical or virtually identical between questionnaires within a country.

Examples include:

- Socio-demographic background
- HIV-related knowledge
- STI experience and treatment-seeking
- Other risks, e.g. alcohol and drug use for non-IDU populations

If the block number is the same across questionnaires, and all the questions and response codes in the block are identical, then the data manager only has to write one recode file for all questionnaires. This saves a huge amount of time and energy. **Unless there is a good reason** to change these blocks, try to keep them identical between questionnaires.

Within questionnaires, be consistent about response options

Very often in behavioural surveillance, similar response codes are possible for different questions. For example, a single questionnaire set may contain questions about frequency of condom negotiation, frequency of condom use, frequency of lubricant use and frequency of needle sharing. **Unless there is a good reason** to vary them, the response codes to these should be kept the same for all similar questions and across all questionnaires, for example 1 “never” 2 “occasionally/ sometimes” 3 “usually/often” 4 “every time”.

If the response codes are continuous categoricals (such as the above example, never/ sometimes/often/always, or age groups, for example) then it is a good idea to code them from lowest to highest (i.e. 1 = never …. 4 = always), rather than the other way around). This is because in analysis, we are often looking for a “dose-response relationship”. In other words, we expect people who have lower exposure to an intervention (.e.g. contact with outreach workers promoting condom use) will have a lower “response” (e.g. condom use with the most recent client). Stata gives output in the order of the numerical value of the categories, from low to high:

```
   frequency   |
    received   | used a condom at last
  condom from  |   commercial sex
   ngo last 3  |       (notes)
     months    |   no        yes    |    Total
---------------+--------------------+----------
   > 3 times   |   14        23     |    37
               |   37.84     62.16  |    100.00
---------------+--------------------+----------
   2-3 times   |   42        50     |    92
               |   45.65     54.35  |    100.00
---------------+--------------------+----------
   once        |   95        83     |    178
               |   53.37     46.63  |    100.00
---------------+--------------------+----------
   never       |   1,000     378    |    1,378
               |   72.57     27.43  |    100.00
---------------+--------------------+----------
   Total       |   1,151     534    |    1,685
               |   68.31     31.69  |    100.00
```

However if we hope to see rising rates of condom use with rising exposure to outreach workers, it is intuitively easier to read output that goes from low to high. Therefore, we should code our categories from low to high. It is easiest to do this at the questionnaire stage, although value labels can also be switched through recoding.

A yes/no question?

An increasingly common convention in data analysis for data that contain many yes/no responses, is to code 0 for “no” and 1 for “yes”.

In many existing questionnaires, data are currently coded 1 for “yes” and 2 for “no”, in part because we think more easily in terms of Yes/No than No/Yes. It is strongly recommended that new questionnaires should keep the same order of responses (yes, then no) but change the code for the “no” answer to 0. This means that the data will be entered with 0 and 1 values from the start.

When introducing variation, try to avoid disrupting uniformity

Sometimes, there are very good reasons to introduce variations even within blocks of questions which are largely similar, or in the response codes to a single question. If the variations are small and the rest of the block of questions is the same as in the other questionnaires, then consider numbering it in a way that does not disrupt the rest of the sequence.

For example, in questionnaires for female, male and transvestite sex workers, Block 3 Questions 1-6 are about numbers and frequency of clients. We want to ask the same questions for those IDU who sell sex, but not all IDU sell sex. One solution is to insert a new question q301 in the IDU questionnaire, asking if the respondent has sold sex. But then the numbers of 301-306 get disrupted (they would have to become q302-q307), and the rename and recode files for this one questionnaire will have to be adjusted for all six questions. An easier solution (at least for the data manager) is to number the filter q300, and keep the others the same. A common solution for non-uniform questions introduced in the middle of the block is to add "a" to the previous question number (so if a question unique to the IDU questionnaire were introduced after q302, then it would be called q302a, and the numbering of the identical questions that follow would not be disrupted)

Similarly, if the response codes for a question differ by population or region, it is easiest for the data manager if similar responses have the same response number in all questionnaires. For example: "Where did you meet your last client?"

Female sex worker: 1) street 2) bar 3) massage parlour 4) brothel

Male sex worker: 1) street 2) bar 3) massage parlour 4) bathhouse 5) gym

Minimising the risk of error with clear questionnaires

Clear questions, clear answers

Small changes in the wording of questions can make a huge difference to interpretation of responses. Every question should be as precise as possible. If the question refers to a particular TR period, it should say so. If it refers only to a specific type of partner, it should say so.

The surveillance manager should know, before the questionnaire is finalised, how each response will be used in analysis. The choice of question depends on how it will be used. Always aim for the simplest question that will give you the information you will use. For example, some questionnaires ask where the respondent is originally from. This may be answered in several ways – town, district, province etc. Province of origin may be useful in assessing mobility – more detailed information will probably be impossible to analyse sensibly. Ask only (and specifically) for the information you are likely to use. Don't ask "Where are you from?", ask "What province are you originally from?"

Among the most common sources of confusion are questions which deal with partner frequency. "How many times did you visit a sex worker last month?" is NOT the same questions as "How many sex workers did you visit last month?". The second may be used in models which require measures of exposure, while the first may be more useful for estimating the size of the client population using methods that balance total number of transactions. Equally, there is a huge difference between the questions "How many people did you have anal sex with in the last week?" and "How many times did you have anal sex in the last week?". Know why you are asking a question, and then formulate the question so that there is no room for confusion.

Some questions can be answered in different ways, not because the responses are different but because the format in which it is given is different. To find out how long a respondent has been working in a given location, the question might be: How long have you worked here? The response might be in weeks, months or years. Unless it is clear which is intended, it will not be possible to interpret the data. The questionnaire should be absolutely clear about the TR, giving separate spaces for months and years as necessary.

Questions with many possible responses

One of the most common sources of confusion in questionnaires is questions with a long list of potential responses. In some cases, respondents are only allowed to give one possible answer to the question. For example, there is generally only one possible response per person to the question "what brand of condom did you use at last sex?". In other cases, respondents can give more than one answer, as for example to the question "what brand of condom have you used?". From the point of view of both the data manager and the analyst, single-response questions are far easier to deal with than questions with more than one answer. And the easier a question is to analyse, the more likely is the data are to get used. **Unless there is a good reason** to allow several responses, it is better to stick to allowing respondents to give just one response per question. This usually gives you as much information as you need, and almost always gives you as much as you are likely to use productively. In the example above, for example, the first question is as likely to give as much information about which condom brand is the market leader as the second, and is far more likely to lead to more detailed analysis of brand preference by geographic area or population group because it is so much easier to analyse. In either case, the respondent should be absolutely clear about whether more than one response is allowed.

The format of questions to which more than one response can be accepted per respondent has important implications for data management. If only one answer per person is possible, the response codes should be numbered sequentially:

1 "Durex" 2 "Pleasure" 3 "Sublime Happiness" 4 "Toughnuts" etc.

If multiple answers are possible, the easiest approach is to number them exponentially, so that it is always clear which combination of answers was given (because there is no overlap between numbers):

1 "Durex" 2 "Pleasure" 4 "Sublime Happiness" 8 "Toughnuts" etc.

In this case, the editor will add up all the response codes and enter a single number. (For an alternative option, see Data entry, below)

Questionnaires should at all costs avoid giving interviewers the option of circling Yes or No for each one of a list of possible responses to a single question e.g.:

Durex 1 "Yes" 2 "No, Pleasure 1 "Yes" 2 "No, "Sublime Happiness" 1 "Yes" 2 "No"

This format creates extra work for the interviewer (they must go back and circle "No" for all the responses NOT mentioned). And it creates confusion for the data editor, since interviewers very often don't bother with the "No"s, and the editor often enters the response as missing. This leads to incomplete denominators, which in turn leads to headaches for the analyst. There is no good reason to use this question format.

Avoid open questions at (almost) all costs

Open questions (where interviewers write in the respondent's answer) are the data manager and the analysts nightmare. They are also usually unnecessary. Every round of surveillance should be preceded by an assessment, and every questionnaire should be pre-tested. By the time the questionnaire is finalised, surveillance managers should have a good enough idea of the likely responses to be able to pre-code almost every question. **Unless there is an absolutely compelling reason**, do not include open questions in a behavioural surveillance questionnaire. The data will not get analysed (and very often they will not even get entered).

Use standard response codes where available

Many questionnaires will contain information on city of interview, province of origin etc. Most countries have standard numerical codes for geographic information – used in the census and other statistical surveys. The coding is available from the national statistical office. It is a good idea to use these codes in the "Master" data so that datasets can easily be shared.

Chapter 3: Data entry

A good questionnaire can make a data manager's job a lot easier. But data entry is where "data management" in the traditional sense really begins.

Different users may have different needs

Choice of software is important but not critical. Commercial packages which "translate" software from one programme to another (such as DBMS Copy or StatTransfer) are widely available, and many software packages can read in datasets created using other software. It is very common practice to use one software for data entry and another for analysis.

If using "translated" data, keep an eye open for variables which may have been assigned to the wrong data type, for example variables that should be numeric, stored as strings. Most packages have tools for fixing this (such as the Stata command "destring").

Think, too, about ultimate users of the data. Some analysis packages restrict the names of variables to 8 characters for example. If these packages are to be used, "short names" must be used in data entry and recoding.

Some packages support multilingual labels; the same recode file can be used to create datasets with labels in more than one language (or use short and long labels, or other variants). For multilingual labelling in Stata, see **help label_language**

The basics: a comprehensive "Master List" and well designed software

Data can be entered in many types of software, ranging from simple spreadsheet programmes such as Microsoft Excel to sophisticated data management programmes such as FoxPro , or the openly accessible freeware CS Pro or EpiInfo. There is no "correct" programme, but it is worth noting that more sophisticated programmes adapted to national surveillance questionnaires may greatly cut down on data entry error. Dedicated programmes such as CS Pro are based on the questionnaires and use a "Master List" to ensure that response codes are correctly assigned (often translating the numerical code into the equivalent textual response for easy checking). These types of software can be configured to follow the questionnaire's skip patterns, minimising the risk that denominators (the key to good data analysis) are correctly entered. They can also build consistency checks into the data entry programme, so that any responses which are clearly inconsistent with earlier answers get flagged and can be rechecked against the questionnaire at the time of entry. For example, if the minimum age for respondents is 15 and a data entry clerk types "8" into the data entry form, the programme will refuse to accept the answer and will display a message "Below minimum age for respondents: re-check questionnaire". Similarly, if a respondent has previously reported never having had sex, and then reports three sex partners in the last month, the data entry form can be made to skip back to the question on first sex, with an appropriate message such as "Inconsistent response, re-check questionnaire".

It is important that the data entry software match the questionnaire exactly. If a response can be entered in years or in months, then the software should have separate variables for years or for months (or the data editor should translate the questionnaire data into a single agreed format before entering). The data entry software for a single questionnaire should be identical for every data entry clerk (if data are being entered at different surveillance sites nationally, there should be a standard entry format across all sites, for example). If data are to be entered at different sites, it is helpful to hold a single training for data clerks from all sites.

Data entry forms should be the same across all questionnaires, where questions are the same. This is especially an issue for questions for which respondents can give more than one answer, where there are several different ways of entering data (see below).

Data entry staff are only human – double enter to avoid errors

Even where specialised software is available, and especially if it is not, raw data should be entered from the questionnaires by two different people, each working on a parallel database. Because it is unlikely that two different data entry clerks will make the same errors, entering the data twice, comparing the two, and checking any discrepancies against the questionnaires will eliminate most of the human errors associated with simply typing in data incorrectly.

A code to identify the data entry clerk should be included in the data entry form and entered into the data set. This allows checking for error rates and retraining of staff as necessary.

Several data management packages, include the freely available CS Pro and Epi Info, have built-in functions which compare data sets and indicate discrepancies.

Questions with more than one response allowed

As mentioned earlier, there are two common ways of dealing with multiple response questions in questionnaires, and this leads to different ways of dealing with them in data entry. The first is that response codes are sequential:

1 “Durex” 2 “Pleasure” 3 “Sublime Happiness” 4 “Toughnuts” etc.

The interviewer will circle all those mentioned by the respondent. In this case, the data entry forms will have to create a separate variable for each possible response. If the question number is b3q5, the variables will probably be labelled b3q5_1, b3q5_2 etc.

The second method is to use exponential response codes, and to add together the responses.

1 “Durex” 2 “Pleasure” 4 “Sublime Happiness” 8 “Toughnuts” etc.

If a respondent says he uses Sublime Happiness and Durex, then the interviewer will circle both. The data editor will enter the response code 5 (4 + 1). Because of the exponential numbering, there can be no confusion as to the possible combination of responses.

From the point of view of the data manager, the second method is rather easier simply because it leads to a single variable (b3q5). It is easier to recode from a single variable with discrete values than from multiple variables. It also avoids the question of the denominator (since everyone who is asked the question will give some response and be entered into the variable). In the first method, the denominators have to be added during the recoding process in order to make a meaningful variable.

The exception is for questions for which there are a large number of possible precoded responses (for example "Can a person reduce the risk of getting infected with HIV in the following ways…") Where many answers are possible, exponential values quickly get very large, and the task of adding up multiple response codes becomes very tedious. In these cases, separate variable entries for each response option are preferable.

Skip patterns and denominators

Many errors in data analysis and interpretation are related to incorrect or poorly understood denominators. Some of these problems are introduced at the data entry stage; they are particularly likely to be related to skip patterns in the questionnaire.

As a rule of thumb, denominators in raw data should include **EVERYONE** who was asked the question, and **ONLY** those who were asked the question. People who were skipped on a variable because of a negative answer to an earlier question should not be coded as 0 or "no" (usually coded 2). If someone was not asked a question, then data for that person should be missing from the corresponding variable in the raw dataset. Note that missing data will be assigned a value by the data entry software (in Excel the cell remains blank, for example, while in State missing values appear as "."). This is NOT the same as a code for a respondent who was asked the question but did not reply to it (often coded as 99) or who said they could not remember or didn't know (often coded as 98). Respondents who should be included in the denominator because they were asked the question but for whom no response code is recorded should be entered using the code for the non-response value given in the master list (e.g. 98/99).

Open questions and write-in responses

Open questions should be avoided wherever possible. Sometimes, write-in responses are permitted for variables such as province or city of origin, because it would take up too much space to list all the possibilities on a pre-coded questionnaire. However the master list for the data editors will include a list of these responses with the standard codes, and the data editor should enter the numerical code into the data set, never the text.

Questionnaires often permit a response of "other" to a pre-coded question, and sometimes have a space for the "other" to be written in. Where this is an option, interviewers sometimes write in a response verbatim, rather than looking to see whether it fits into the precoded response options. The data editor should always check to see if write-in answers fit into any of the precoded response categories, and should assign them to those categories if they do. Take the example of the question: Who do you live with?

1 "Family" 2 "Friends" 3 "Boarding house" 4 "Street" 5 "Other" _GRANDFATHER_

In this case, the data editor should simply enter the response code 1. Careful editing of write-in responses will vastly increase the likelihood that these data will be used. If data entry is centralised and data editors see that high proportions of respondents are giving a response which was not precoded (perhaps because formative research or questionnaire testing were inadequate), they can assign a numerical code to that response. Very clear documentation is needed for any such decision.

Data types

Different software packages have different ways of recording data, but they all distinguish at a minimum between numerical values and text (usually referred to as "string variables").

From the analysts' point of view, numerical values are far, far preferable to string (text) values. Sex, for example, should be coded 1, 2 rather than M, F. String or text variables should only be used when absolutely unavoidable. Note that text variables will appear in the data base exactly as entered. Any variations or errors will increase the difficulty of analysis. For example "Where did you go when you last had an overdose?" might, if entered a string variable, produce the following output:

Value	Frequency
H CENTER	1
H. CENT.	2
HEALTH CENT.	5
HEALTH CENTRE	10
HEALTH CENTER	24
HELTH CENTRE	1

In fact, all these 43 observations should all have the same value, but because they are entered as strings and because of the variation in data entry, the show up as 6 different variables. Data such as these will never get analysed, so there is little point collecting them.

Data editors should ensure that data storage types are appropriate to the type of value entered. Sometimes, data types get mixed up by mistake in data entry or when databases are converted from one package to another, so that numbers are mistakenly read as string variables. In Stata, if you try to perform an operation on such data, you will probably get the error message "Type mismatch". This can usually be solved by using the "destring" or "encode" commands (see Stata Help files).

Protecting data and documenting changes in raw data

Recoding data is a dangerous business. It is easy to make mistakes, and vital that we use safeguards to protect original data and allow for the correction of any mistakes that arise. It is also absolutely critical that all the steps taken in renaming, recoding or otherwise manipulating raw data are carefully written down and preserved for future users (see Chapter 5 on documentation).

As soon as you receive a raw data set, create a directory with a name that clearly distinguishes this as backup data (e.g. c:\surveillance\behaviour\2005 data\do_not_touch. Copy all the raw data into this directory, and copy it also onto an external data storage device such as an external hard drive, CD Rom etc.

Chapter 4: Coding and combining data sets

A round of behavioural surveillance usually includes data from more than one sub-population and from more than one location. There may also be data for more than one year. These data are usually entered as separate data sets, one data set per population, per site and per year. Say a country has surveillance in four sub-populations in 10 cities for three years — that adds up to 120 separate data sets. Obviously, this makes analysis very difficult; statistically meaningful comparisons between populations, across sites and most importantly over time are virtually impossible unless the data sets are combined.

Several steps must be taken in combining data from different datasets into a user-friendly whole, and there is often more than one way to perform these steps. Current experience suggests that performing the steps in the order laid out below will minimise duplication of effort.

1. **Explore your data**
2. **By dataset: add any new variables if needed**
3. **By dataset, ONLY FOR QUESTIONS/QUESTION NUMBERS THAT ARE NOT IDENTICAL ACROSS DATASETS: Give variables names that do not relate to the question number, and make sure that these match across questionnaires (years and population groups)**
4. **By dataset: ONLY FOR VARIABLES THAT HAVE THE SAME NAME BUT DIFFERENT RESPONSE VALUES IN DIFFERENT QUESTIONNAIRES: Assign common values to variables that have the same name**
5. **Combine data sets**
6. **ON THE COMBINED DATA SET: Give variables names that do not relate to the question number, for all the variables that had identical question numbers and values across data sets**
7. **Keep only the variables you want to use in recoding or analysis**
8. **Recode data for analysis**
9. **Order the variables in the data set**
10. **Label and annotate variables and datasets**

You need to go through steps 1-7 BEFORE you can go on to the more complex recoding that will be used in analysis. When you are planning your work, please be aware that the data coding and preparation can take several days. It usually takes far more time than the analysis, especially the first time when you are making decisions about recodes and writing the first do files. A complete and well-documented set of recodes can easily be re-run as subsequent sites or rounds of data come in, so the process gets easier over time.

Step 1: Explore the data

Most of the data exploration will take place as we are recoding and labelling the data. However there are a few things that you have to do before you can start working with the data. You **must** have the questionnaires in front of you when you are coding data. You

should have the a code book as well (which states clearly what each numerical code stands for – this usually follows the questionnaires).

The **denominator** is the most important thing in data analysis. You need to know how many people are in each of your data-sets (a person, with all their associated data, occupies one horizontal line of the Stata database, and is known as an "observation").

describe

This will tell you how many people ("observations") are in the dataset, and how many variables there are. Write down the number of observations and stick it on the top of your screen. Describe will also list all of the variable names, with their labels if they have them, and tell you how the data are stored. Don't worry about this too much unless you see a lot of **"str".** This means "string variable" – letters instead of numbers — and should only appear if words rather than numbers have been typed in to the database. If you find that a numerical variable has been stored as a string variable by mistake, go to Help, type "**destring**" or **"encode"** and follow the instructions.

It is important at this stage to look at data sets in the same sub-population, but for different locations. If these data sets are identical in structure, then you should combine them before you take any further steps, using the "append" command given under Step 5 below. Make sure before you do this that there is a variable which identifies the location already in the data set. If there is not, generate a value for each location before you combine data sets for the various locations (see step 2). A quick way to see whether two data sets are largely the same is to compare them with the command "cf". First, read one data set in to memory. Then use the command

cf _all using "second dataset name.dta", verbose

This should tell you if the number of variables is the same, and give details of those that differ.

Step 2: Generate new variables

Because questionnaires are printed for a specific population, and often for a specific round of surveillance, they may not include variables which we will want in the user-friendly combined dataset. The questionnaire for female sex workers probably does not have a question about gender (sex), because all female sex workers are by definition female. But in analysis, we will be interested in analysing some variables by sex. So we have to add a variable for sex. We do that by generating a new variable. If you generate a variable and give it a single value, then it will carry that value for every individual in the dataset. For example: generate sex = 2 will assign a value of 2 to the variable sex for everyone in the data set.

Since we will certainly eventually want to compare changes over time, there should certainly be a year recorded in the data set. E.g. Generate year = 2005. Note that what we are really interested in is differences between rounds of surveillance. If one round of BSS spans December and January (i.e. two years), then add a variable for Round. If you add a variable such as 2004/5 it will be recorded as a string variable and will make the analysts life more difficult.

If you are not familiar with the file management commands in the first five lines of Example 1, please turn to the “Getting started with Stata” annexe (Appendix 2) and read it before continuing.

EXAMPLE 1: GENERATING VARIABLES FOR A WHOLE DATA SET

```
clear
set mem 250
cd: c:\surveillance\behaviour\data
use FSWdata
save FSWrecode, replace
generate sex = 2
generate year = 2005
save, replace
clear
use msmdata
save msmrecode, replace
generate sex = 1
generate year = 2005
save, replace
```

The number of variables generated in this way (assigning the same value to every person in the data set) is very small. More often, we generate new variables and then assign them different values according to whether an individual meets a certain set of conditions. These types of newly-generated variables are discussed under recoding, below.

If you need unique identifiers in your final data set, you need to think about this before combining any existing data sets. Sometimes, identifiers such as respondent numbers are unique *within* a data set but not *across* datasets. In this case, we need to create a unique identifier according to available data. A common way to do this is to create a variable out of the year, the target group identifier, the location identifier and the respondent identifier. We need these to be appended in sequence, so that they yield an identifier of the same number of digits. The easiest way is to look at the number of digits in each variable, figure out the final number of digits, then multiply to create the correct number of zeros after each value, and then add up. An example:

Target group: two digits

Year: four digits

City: two digits

Respondent number: three digits.

This gives us a total of 11 digits.

So we have to create nine zeros after the target group (to create space for another nine digits) i.e. we have to multiply the target group by 1,000,000,000. Then we need to create space for five digits after the year, and three digits after the city. Using mathematical commands as described in Appendix 2, our command would be:

```
gen idnum = target*1,000,000 + year*10,000 + city*1,000 +respno
```

EXERCISE 1: GENERATING NEW VARIABLES

You have five data sets. One is for female sex workers in Paris (file name c:\surveillance\behaviour\data\raw data round 1\FSW\paris, city code 1), collected in the first round of national surveillance in November 2003. The second, also first round data from female sex workers, was collected in Lyon (city code 5) in January 2004. The third female sex worker data set is second round data from Lyon, collected in January 2005. Finally, you have one data set for drug injectors in Paris, also for the second round (December 2004). Each respondent has a three-digit respondent number, beginning at 100 in each data set.

In preparation for combining these data sets, add any variables necessary to compare behaviour over time, between populations and by sex, and ensure that each respondent in your final data set has a unique identifying number.

Hint: You can refer to the suggested standard codes in Appendix 3 for some of the variable names and codes

Step 3: Rename variables so they match across questionnaires

In raw data sets variables tend to follow the questionnaire numbers. However questionnaires vary between populations, and may also change over time. Age might be Block 1 question 1 (b1q1) in the female sex worker questionnaire, followed by education (b1q2). But the IDU questionnaire first records of the sex of the respondent, so age becomes b1q2 and education b1q3 in the IDU dataset. To create data sets which combine data from different populations, we have to be sure that we are combining the same variables. The safest way to do this is to give the variable a name which is independent of the questionnaire, for example by calling both b1q1 in the sex worker questionnaire and b1q2 in the IDU questionnaire “age”. Similarly for “education”.

There are two ways to do this. The **safest** is to generate a new variable with the same name as the original variable, using the same command as above, “generate”. In this case, the original variable will stay in the data set (in case you need it later to use in a different context), until you tell Stata to drop it.

The **easiest** is to use the command “rename”. This simply changes the name of the variable in the data set. The original variable name will disappear from the data set. This is dangerous if you are going to change the variable in any way (i.e. recode it), because if it turns out later that you want to use the original variable, it will no longer be there.

EXAMPLE 2: RENAMING VARIABLES THE *SAFE* WAY

```
set mem 250m
cd: c:\surveillance\behaviour\data
use FSWdata
save FSWrecode, replace
generate education = b1q2
save, replace
clear
use idudata
save idurecode, replace
generate education = b1q3
save, replace
```

RENAMING VARIABLES THE *EASY* WAY

```
set mem 250m
cd: c:\surveillance\behaviour\data
use FSWdata
save FSWrecode, replace
rename b1q1 age
save FSWrecode, replace
clear
use idudata
rename b1q2 age
save idurecode, replace
```

Unless you are very sure of what you are doing, it is recommended that you always choose the safe rather than the easy way.

In choosing names for variables, look at Appendix 3. These are suggested standard names for variables commonly used in BSS. Using these names will make it easier to compare data across regions and countries if necessary. Training materials for data analysis and use are being developed using these names, so their use will also mean that international training materials can more easily be used to train on locally specific data. Clearly, this list does not include all the regional variability of BSS data. Where these names do not apply, develop names that make sense to you.

Make sure that the name makes sense for all the groups to whom it will apply. For example, you will have a variable for condom use at last sex with client for sex workers and condom use at last sex with a sex worker for male groups. This describes the same thing – the level of condom use in commercial sex – and one of the reasons you are combining the data sets is so that you can compare levels of condom use in commercial sex across different populations — both buyers and sellers. So you cannot use “condomlastclient” or “condomlastcsw”. You need to choose a name that works for both, such as “condomlastcs” (cs stands for commercial sex).

Step 4: Reassign response values so they match across questionnaires

Because sex and drug scenes differ, the same question may have different possible responses for different populations. Because most response codes are sequential (1,2,3) this means that the same response code value can have different meanings in different populations. Unless we deal with this before combining the data, we will no longer know what the code refers to once the data are combined. So we need to make sure that the

same value has the same meaning before combining the data. We use "recode" to change the values around so that they match.

EXAMPLE 3: REASSIGNING RESPONSE VALUES

All questionnaires include a question about the location at which the respondent was interviewed. The variable has been renamed "location". For sex workers, the response codes are:

1 "sauna" 2 "nightclub" 3 "karaoke" 4 "salon" 5 "street" 6 "other"

For MSM, the response codes are:

1 "bar/club" 2 "teahouse" 3 "bathhouse" 4 "park" 5 "public latrine" 6 "other"

We want to avoid too many responses, because it makes analysis very difficult, so we try to combine similar responses without creating confusion. In this case, the values might be rewritten like this:

(note to layout: this should just appear as ordinary table but can't seem to make that happen in shaded box — please fix)

FSW	MSM	Combined
1 Sauna	3 Bathhouse	1 Sauna/bathhouse
2 Nightclub	1 Bar/club	2 Bar/nightclub
3 Karaoke		3 Karaoke
4 Salon		4 Salon
5 Street	4 Park	5 Street/park
	2 Teahouse	6 Teahouse
	5 Public latrine	7 Public latrine
6 Other	6 Other	8 Other

In other words, the values for female sex workers would stay the same with the exception of 6, "other" which would become 8 so that it comes after the additional values added to ensure that MSM-specific codes can be included, while most of the MSM codes get switched around either to match FSW codes or to avoid clashing with them.

```
use FSWdata
save FSWrecode, replace
recode location 6=8
save, replace
clear
use msmdata
save msmrecode, replace
recode location 1=2 2=6 3=1 4=5 5=7 6=8
save, replace
```

EXERCISE 2: RENAMING AND REASSIGNING RESPONSE VALUES

The questionnaires for male, female and transvestite sex workers all ask about the most common occupation of clients. However, formative research has shown different client profiles, so response codes vary, as follows:

Female sex workers, Block 3 Question 4. Who are most of your clients?

1 "Civil servants" 2 "Military" 3 "Truck drivers" 4 "Sailors" 5 "Other" 6 "Don't know"

Male sex workers, Block 3 Question 6. Who are most of your clients?

1 "Businessmen" 2 "Civil servants" 3 "Other" 4 "Don't know"

Transvestite sex workers, Block 4 Question 12. Who are most of your clients?

1 "Students" 2 "Sailors/dock workers" 3 "Taxi drivers" 4 "Police" 5 "Other" 6 "Don't know"

Rename and recode the variables as needed in preparation for combining data.

Step 5: Combine data sets

Once necessary dataset-specific variables have been generated, all data sets are individually renamed, and adjustments have been made so that values match across data sets, you are ready to combine data sets.

In Stata, we simply use the command "append using" . Read one dataset into memory, then append as many other datasets as you need to.

EXAMPLE 4: APPENDING DATASETS

```
set mem 250m

use "c:\surveillance\behaviour\data\combined data\FSWdatarecode.dta"

append using "c:\surveillance\behaviour\data\combined data\msmrecode.dta"

append using "c:\surveillance\behaviour\data\combined data\idurecode.dta"

append using "c:\surveillance\behaviour\data\combined data\truckerrecode.dta"

save "c:\surveillance\behaviour\data\combined data\allcombined", replace

erase "c:\surveillance\behaviour\data\combined data\FSWdatarecode.dta"

erase "c:\surveillance\behaviour\data\combined data\msmrecode.dta"

erase "c:\surveillance\behaviour\data\combined data\idurecode.dta"

erase "c:\surveillance\behaviour\data\combined data\truckerrecode.dta"
```

Note that this cannot be used to combine datasets for the same individuals (for example if you have behavioural and biological data for the same people in two different datasets). For more information on combining datasets for individuals, see Appendix 2.

Note also that if you have data labels in your data sets, and if they differ between data sets, then in Stata the labels of the *last* data set to be appended will overwrite all the rest. It is safest to label data only *after* all data sets have been appended. If data are already labelled, "**lab drop**" will get rid of the labels.

Keeping things tidy

The process of combining data sets is likely to use data from several different locations, years and populations. If these are all kept in the same directory, the result can be great confusion. In addition, during the process, we often generate a number of "temporary" datasets which will never be used for analysis. These clutter up our directories and add greatly to the confusion. Because of this, it is a good idea to develop "tidy" data habits early on.

It is a good idea to keep raw data in folders of their own (these folders might be labelled "raw data round1" "raw data round2" etc.). Recoded data can be kept in another folder, together with the appropriate recode files. Analysis "do-files" and logs with the results of analysis can be kept in a separate folder again.

A directory structure for the data described in Exercise 1 might look something like this:

```
c:surveillance  - HIV

                - STI

                - behaviour   - admin

                              - questionnaires
```

- codebooks			
- data	- raw data round 1	- FSW	- paris.dta
			- lyon.dta
	- raw data round 2	- FSW	- paris.dta
			- lyon.dta
		- IDU	- paris.dta
	- combined data	- masterbss.dta	
		- combine all.do	
		- rename.do	
		- recode.do	
		- labelenglish.do	
		- labelfrench.do	
	- analysis files	- condom analysis.do	
		- condom analysis.log	
		- analysis for paper.do	
		- idu paper.log	

It is easy to switch back and forth between folders in a do-file. It simply involves typing the entire directory path when opening data files, running do files, or saving either data files or do files. (Remember that if you do not specify the whole path name, Stata will look for data and do-files and will save logs and do-files to the current working directory).

```
run “c:\surveillance\behaviour\data\combined data\combine all.do”

save “c:\surveillance\behaviour\data\combined data\masterbss.dta”, replace

log using “c:\surveillance\behaviour\data\analysis files\condom analysis.log”, replace

run “c:\surveillance\behaviour\data\analysis files\condom analysis.do”

log close
```

It is also easy to eliminate temporary files by using the command “erase” when the file is no longer needed. Ultimately, the only data files we really need are the raw data files, and a single master file of all the data combined.

EXERCISE 3: COMBINING DATA SETS

Go back to Exercise 1. Repeat the exercise by combining data sets at the appropriate time, to minimise the amount of repetition in the coding. Use the directory structure in the box “Keeping things tidy”, above. Get rid of any unnecessary files. You should end up with a single file with all the data sets in it, plus your raw data files.

Step 6: Rename variables in the combined dataset

The combined dataset will be a mixture of “friendly” and “unfriendly” variable names.

The “friendly” variable names — things like “year” and “clientswho” are the ones that you have generated from scratch to identify datasets, or the ones that you had to rename by data-set because there were variations in question numbers or response codes between datasets.

The “unfriendly” variable names are things like b3q23 — hard for analysts to remember, compared with a variable name such as “age” — and they belong to the questions that have the same question numbers across all of the data sets and have not yet been renames.

These now need to be given names to make it easier for analysts to cope with — it is easier to remember “condomlastcs” than b3q32. Some standard names are given in the example codebook in Appendix 3. It is recommended that you use these names if they match your data set and if there is no good reason not to use them. Note that while variable *labels* (and value labels) can be created in more than one language, variable *names* can only be in one language.

Remember that if you use the “rename” command the original variable name will disappear from the data set, while if you use the “generate” command the original variable name will stay in the data set until you drop it.

Step 7: Keep the variables you want to use in recoding or analysis

Large data sets become unwieldy and confusing. Before recoding data sets, it is a good idea to drop all the variables that you do not intend to use in recoding or analysis. This will include most of the variables that are based on the original questionnaire numbers.

There are two commands that can be used: **keep** (which will keep any variables specified on the keep line) and **drop** (which will drop any variables specified on the drop line). It is usually easier to drop than to keep, unless you intend to create a very limited data set for analysis of a few specific variables. It is possible to use the wild card “*” with keep and drop commands. For example

drop b1* b2* b3*

will eliminate from the data set any variable beginning with b1, b2, b3 etc. so b1q1 will go, so will b2q7, b3q2 etc. If you just use drop b* then you will also lose the variables “bisexual”, “brandcondom” etc.

Step 8: Recoding data sets for analysis

Combining data sets is the easy bit. The most important and also most complex part of data management is in recoding the data once the data sets have been combined. There are several ways to recode data, and also several different reasons to recode. These include:

- grouping data for ease of analysis
- standardising denominators to minimise errors in analysis
- ensuring that numerators and denominators are comparable across populations and across time
- creating variables which can be used as indicators of programme performance

It is possible to recode and label each of the data sets before you combine them, but since there is a lot of overlap it is more efficient to do it all at the end. It is also safer, because you can ensure that the codes and labels are correct for all of the data sets. Another advantage is that if you discover a mistake, you only need to correct it in one do-file, rather than in many.

The golden rule of recoding:
Look at your data before AND after you recode.

Once you get going on coding, it is easy to write line after line of code without ever checking the result. Often, it is not until someone tries to analyse the data and comes up with strange and unexpected results that errors get discovered. It is much, much easier to check your data before and after each code than to try to find an error in 2,800 lines of computer commands, which is what your final recode files may be.

Commands to use when looking at a variable:

tabulate, inspect, summarize, tabstat (see Stata Help for more details)

Things to look for before you touch a variable:

How many people are in the denominator? Is this what you expect?

How many missing and non-response values are there? Are there enough non-response values to worry about? What are you going to do with them?

How many different values are there? Is it what you expect? Does it match the questionnaire, and your earlier recodes?

What are the maximum and minimum values? Is it what you expect?

What is the data type? Is it what you expect?

Things to look for after you have recoded a variable:

How many people are in the denominator? Is this what you expect?

Are you sure that the denominators are correct for all the relevant sub-populations?

Are the means or proportions roughly what you would expect? Are they consistent with other variables in the dataset?

DOES IT LOOK LIKE IT MAKES SENSE?

Grouping continuous variables for ease of analysis

Some variables, such as age, income, price paid (for sex, drugs, condoms or treatment) have numerical values which rise in a continuous scale, with a lot of variation. It can be difficult to analyse continuous variables; to make things easier we often separate them into groups of four or five ranges. The easiest way to do this is simply by recoding (thinking of each numerical value as a response code above). First, generate a new variable that has the same contents as the variable you want to group. Then recode it into grouped values.

Beware: if you do not create a variable with a new name but recode a variable in the data directly, you will lose the original data, and only have the grouped data.

EXAMPLE 5: GROUPING CONTINUOUS VARIABLES

***Note that in this and subsequent examples, the code in brackets should not be included in your final do file. These lines of code should be typed directly into the command window, and you should look at the results before going on to the next step.**

```
clear

use “c:\surveillance\behaviour\data\combined data\allcombined.dta”, replace

save “c:\surveillance\behaviour\data\combined data\masterbss.dta”, replace

(bysort target: summarize age)

-> target = direct female sex workers

     Variable    |    Obs        Mean    Std. Dev.        Min        Max
-----------------+------------------------------------------------------------

     age         |    6168       27.28086.594533          13         53
------------------------------------------------------------------------------

-> target = indirect FSW
     Variable    |    Obs        Mean    Std. Dev.        Min        Max
-----------------+------------------------------------------------------------

      age        |    5483       26.57432    6.529596    15         75
------------------------------------------------------------------------------

-> target = high risk men
     Variable    |    Obs        Mean    Std. Dev.        Min        Max
-----------------+------------------------------------------------------------

     age         |    8801       30.92126    9.560373    14         99
```

We can see from the first glance that there is a “missing” code of 99 in the male data set. We can also see that the mean ages of men are rather different, though this may be affected by the “missing” code. First, check how likely that is by looking at the missings:

```
(tab target if age = = 99, m)

       target group     |       Freq.      Percent     Cum.
-------------------------+----------------------------------
       high risk men    |       1          100.00      100.00
-------------------------+----------------------------------
       Total            |       1          100.00
```

It is only one person and there are no other missing values. Out of 8801 in the sample, we can be pretty sure this one observation does not make that much difference to the mean.

```
(tabstat age, by (target) stats (min p25 p50 p75 max m))

     target group          |        min       p25      p50     p75       max
-----------------+-----------------------------------------------------------
     direct FSW            |        13        22       27      32        53
   indirect FSW            |        15        22       25      30        75
     high risk men         |        14        24       29      37        99
-----------------+-----------------------------------------------------------
     Total                 |        13        23       27      33        99
--------------------------------------------------------------------------
```

From this we can see that at the younger ages, the groups are not all that different. But a look at the third quartile shows that the male group is skewed upwards in comparison with the sex worker groups. Still, it would seem that the best grouping for age in this data set reflect those most often used in programme decision-making and in reporting indicators:

```
gen agegroup = age

recode age min/19 = 1 20/24=2 25/34=3 35/97 = 4

(tab target agegroup, m)
```

```
              |   target group
    age group |   direct fe  indirect   high risk |      Total
--------------+-----------------------------------+-----------
    <20       |   642          557        532     |      1,731
              |   10.41        10.16      6.04    |      8.46
--------------+-----------------------------------+-----------
    20-24     |   1,775        1,922      2,060   |      5,757
              |   28.78        35.05      23.41   |      28.15
--------------+-----------------------------------+-----------
    25-34     |   2,777        2,304      3,528   |      8,609
              |   45.02        42.02      40.09   |      42.09
--------------+-----------------------------------+-----------
   35+        |    974       700      2,680       |   4,354
              |  15.79      12.77      30.45      |   21.29
--------------+-----------------------------------+-----------
    .         |   0            0          1       |      1
              |   0.00         0.00       0.01    |      0.00
--------------+-----------------------------------+-----------
    Total     |   6,168        5,483      8,801   |      20,452
              |   100.00       100.00   100.00    |      100.00
```

Note that we have excluded the one person with missing data from this version of the code. That person will not be included in any analysis using the variable "agegroup". Further information on dealing with missing values can be found on page 40.

save, replace

If there are no non-response values, then the recode line could have been written as: **recode age min/19 = 1 20/24=2 25/34=3 35/max = 4**

This would have assigned all numeric values over 29 to the category "30 or over". It would include any missing values coded to 99, but would not include missing values coded to "."

The hardest part of any recode that involves categorising data is to decide on which categories make sense. In general, points to consider are:

- **Are the categories meaningful in terms of HIV programming?**

  If we are looking at injecting frequency, for example, there is a big difference in terms of risk as well as service needs between someone who injects only once or twice a month and someone who injects daily, but the difference between someone who injects three times a day and five times a day may have less relevance either for risk or for service provision.

- **Do the categories have political significance?**

  Sometimes we choose categories because they fit into political priorities or have important advocacy value. An example might be including an agegroup category for teenagers, even if they are only a small proportion of the population. Equally, we are sometimes influenced by nationally or internationally standardised indicators, even if they make little sense in the local epidemic context. If programme managers have to report such indicators, the least we data managers can do is make their job easier by recoding the data appropriately so they can spit the indicators out with minimum effort.

- **Is the distribution roughly even between categories?**

  In general, analysts would prefer that there be roughly the same number of people in each category. You can check the relative distributions by tabulating the data and looking at the cumulative totals — whatever the values at 25%, 50% and 75% will give you quartiles. This can prove difficult, because distributions can vary quite a bit between groups. For example the price charged by indirect sex workers in nightclubs and massage parlours may be much higher than that charged by sex workers in brothels, so it will be difficult to find categories that provide an even distribution for both groups.

The only absolute rules in creating categorical variables out of continuous variables are:

1. Look at the distribution in the raw data before making any decisions

2. Think about how you will use the categories in analysis

3. Deal appropriately with non-response values

EXERCISE 4: REGROUPING DATA

In your male, female and transvestite data sets and in your high risk men data sets, you have information on the price received or paid at last commercial sex. In your combined data set, the distributions are as follows:

```
               |   min    p25    p50    p75    max
---------------+-----------------------------------
FSW            |   0      75     90     200    2251
MSW            |   2      80     100    150    999
Transvestite   |   0      10     15     25     300
Client         |   0      30     50     60     2000
Total          |   0      50     70     150    7000
---------------------------------------------------
```

Write down the command you would have used to generate the information on distribution above, and other commands you would use to explore the data.

Create a variable which groups the price data appropriately for analysis.

Check the resulting variable.

Standardising denominators

It cannot be said too often: a very high proportion of errors in data analysis are related to ignoring or misunderstanding the denominator for any given variable. This problem arises very frequently in behavioural surveillance data, because BSS questionnaires typically have a lot of skip patterns.

If data entry is correct, then in the raw data everybody who was asked a question, and nobody who was not asked, should appear in the denominator for that question. However, in some cases these denominators are not particularly useful for analysis. They need to be manipulated before the data can be used sensibly for programme planning or management.

The safest way of ensuring that you have a correct recode with a clear denominator is to generate a new variable which is blank (missing) for all respondents, and then set criteria to “fill it up”, assigning values for different response codes.

Data management software tends to use “operators” that work just like in ordinary mathematical formulae, and they generally use brackets in the same way as maths, so that commands inside the brackets get resolved before commands outside the brackets.

In recoding just like in maths, there is very often more than one way to arrive at the “correct” answer. But in recoding, just like in maths, small differences or errors in the use of operators or brackets can lead to huge differences in the results. To avoid errors, it is wise to follow certain basic rules.

- Always generate new variables as blank
- If the new variable is a “yes/no” variable, always code the “No” – those who should be included in the denominator but will NOT be included in the numerator — first.
- Follow the logic of the questionnaire and include all the skip pattern possibilities to ensure completeness of the denominator.

High school maths: a reminder

These are the operators for Stata. Other packages will use very similar operators.

= = means “EQUAL TO”

& means AND, i.e. BOTH condition A AND condition B must be fulfilled

| means OR, i.e. EITHER condition A OR condition B must be fulfilled

~= means “NOT EQUAL TO”, ..i.e. the condition must not be fulfilled

< means “LESS THAN”

> means “GREATER THAN”

if is the syntax used for fulfilling conditions using the above operators. For example:

```
replace happy = 1 if monthsholiday = =3 & newpartner = = 1
```

BRACKETS: THINK CAREFULLY

```
replace potentialclient = 1 if (rich = = 1 | poor == 0) & blonde == 1
```

will code as potential clients only people who are rich OR are not poor, AND are blonde: in other words, all rich blondes. Rich brunettes are not potential clients with this code.

```
replace potentialclient = 1 if rich = = 1 | (poor == 0 & blonde == 1)
```

will code as potential clients people who are rich OR people who are not poor AND are blonde: in other words, all rich people, included but not limited to blondes. Rich people with dark hair would be considered potential clients with this code.

As a general rule, you do not need to use brackets if all the operators (&, | etc.) are the same in the command. The minute you start mixing operators (for example you have an & and an | in the same command line), then you need to use brackets.

Another general rule: variables whose positive value is “filled up” using “AND” operators will have a negative value assigned using “OR” operators, and vice versa.

```
replace safe = 0 if unprotectedsex = = 1 | shareneedle = = 1

replace safe = 1 if unprotectedsex = = 0 & shareneedle = = 0
```

EXAMPLE 6: STANDARDSING DENOMINATORS FOR ANALYSIS

In Block 5 on HIV/AIDS knowledge, the questionnaire asks:

q1) Have you ever heard of HIV or AIDS? 1) Yes 2) No (skip to block 6)

q2) Do you think AIDS can be prevented? 1) Yes 2) No (skip to block 6)

q3) Which of these ways can AIDS be prevented?

a) Always use condoms in sex

b) Avoid mosquito bites

c) Never share a syringe or injecting needle

d) Never have sex

e) Stick to one, faithful partner.

q4) Can you tell if someone is infected with HIV just by looking at them?

To protect herself against HIV, a female sex worker has to know, at a bare minimum, that she can avoid infection by using a condom. (Some of the other items of knowledge may be correct, but they are probably not directly relevant to this population – sex workers cannot choose to avoid HIV by having only one faithful partner unless they also choose to go hungry.)

A user-friendly dataset will have this measure of effective prevention knowledge already coded for the data analyst to pick up and use. The numerator (the 1 value code) is easy. It will be the same as the 1 for b5q3_a. But who should be in the denominator? For our HIV prevention programmes to be successful, we want everyone in high risk groups to know that condoms prevent HIV. So everyone should be in the denominator. The problem with b5q3_a is that we have already skipped the people who have never heard of AIDS or who don't think it can be prevented. They do not get asked the questions about prevention methods, so they are not in the denominator. When we make the recode, we have to put them back in.

First, make a new variable using "generate" and set all the values to missing (.).

```
generate condom = .
```

Now we need to "fill it up" according to logical criteria, using the command "replace". First, think of all the people who do NOT know that condoms prevent HIV. That is the people who answered "no" (2) to b5q3_a, PLUS the people who did not think HIV could be prevented, PLUS the people who have never even heard of HIV.

In this case, we want to code people to zero (do NOT know that condoms prevent HIV) if they fulfil the condition of not knowing condoms prevent HIV, or they don't know that anything prevents HIV, or they haven't even heard of HIV. We first rename variables so that we can use the same recode for a different population group even in the question numbers are not identical.

```
(tab b5q1, m)

b5q1          |   Freq.     Percent     Cum.
--------------+-----------------------------------
     1        |   5,836     88.30       88.30
     2        |     747     11.30       99.61
     9        |      26      0.39      100.00
--------------+-----------------------------------
     Total    |   6,609    100.00
```

```
gen knowaids = b5q1

recode knowaids 2/9 = 0
```

Note that in this case, we have made the decision to code those who did not respond to the question (less than half of one percent of respondents) to "no", with the assumption that in this cultural context, refusal to answer is an alternative to admitting ignorance.

(tab knowaids)

```
knowaids        |   Freq.     Percent      Cum.
----------------+--------------------------------
     0          |   773       11.69        11.69
     1          |   5,836     88.30        100.00
----------------+--------------------------------
     Total      |   6,609     100.00
```

(tab b5q2, m)

```
b5q2            |   Freq.     Percent      Cum.
----------------+--------------------------------
     1          |   4,815     72.86    72.86
     2          |   160       2.42     75.28
     8          |   887       13.42    88.70
     .          |   747       11.30    100.00
----------------+--------------------------------
     Total      |   6,609     100.00
```

Note that the 747 are missing on this code. These are the individuals who had never heard of AIDS, and were not asked the question. Those who were missing on that variable *(coded 9, i.e. 26 cases) were* asked the question. We have already coded them as 0, i.e. do not know AIDS. But we could do a further check here, to see if the assumption we made in doing that was fair:

tab b5q2 if b5q1 = = 9, m

```
b5q2            |   Freq.     Percent      Cum.
----------------+--------------------------------
     1          |   2         7.69         7.69
     2          |   1         3.85         11.54
     8          |   23        88.46        100.00
----------------+--------------------------------
    Total       |      26    100.00
```

We can see that only two of the people who did not answer the question about AIDS said that it could be prevented. For the others, our earlier assumption is probably correct. But we should probably correct the “knowaids” code in the do-file and run the file again before going any further.

```
drop knowaids

gen knowaids = b5 q1

recode knowaids 2 =0

replace knowaids = 0 if knowaids == 9 & b5q2 == 2 | b5q2 == 8

replace knowaids = 1 if knowaids == 9 & b5q2 == 1
```

```
(tab knowaids)

knowaids       |   Freq.      Percent      Cum.
---------------+-------------------------------------
     0         |   771        11.69        11.69
     1         |   5,838      88.30        100.00
---------------+-------------------------------------
     Total     |   6,609      100.00
```

Now we can get on with recoding b5q2. In this case, if people say they do not know if HIV can be prevented or not (b5q2= =8), then it seems fair to make the assumption that they do not know HIV can be prevented, so we code them to 0, do not know that HIV can be prevented.

```
gen canprevent = b5q2

recode canprevent 2/8 =0

(tab target canprevent, m)
```

```
canprevent     |   Freq.      Percent      Cum.
---------------+-------------------------------------
     0         |   1047       15.84        15.84
     1         |   4,815      72.86        88.70
     .         |     747      11.30        100.00
---------------+-------------------------------------
     Total     |   6,609      100.00
```

Now we can get on to coding the condom prevention variable.

```
gem condomp =  b5q3_a
```

To go back to our coding logic, people who do not know that condoms can prevent HIV are those who have never heard of AIDS, who don't know it can be prevented, or who don't know it can be prevented with condoms:

```
replace condom = 0 if knowaids == 0 |  canprevent == 0 |  condomp== 2
```

This gives us the people who will be in the denominator but NOT in the numerator. Now we have to "fill up" the numerator. This is much easier, because they only have to fulfil one condition, i.e. to know that condoms prevent HIV.

```
replace condom = 1 if condomp == 1

drop condomp

(tab condom)

condom          |   Freq.     Percent      Cum.
----------------+--------------------------------
0               |   1,989     30.10        30.10
1               |   4,620     69.90        100.00
----------------+--------------------------------
Total           |   6,609     100.00
```

In creating user-friendly data sets, we frequently code the same variable with different denominators, so that it can easily be used for different purposes in analysis, and to minimise the risk that people will misinterpret the data.

For example, a common indicator of risky sex among youth is the proportion of young people with more than one sex partner in the last year. One sometimes sees reports saying something like "58 % of young people have multiple sex partners". But what is the denominator? In part because people who have never had sex or who have not had sex in the last year are not asked about the number of partners, this indicator is often reported only for people who are currently sexually active. This gives us a profile of the level of high risk behaviour *among those who have any risk at all*. If this is only a small minority of the population, then it will give a very distorted picture of the level of high risk behaviour in the population as a whole. If only 19% of young people surveyed have ever had sex, and only two thirds of them have had sex in the last year, then the true population level of high risk behaviour among young people will be 0.19 x 0.66 x 0.58. So the more accurate headline would be: "7% of young people have multiple sex partners".

Name your denominators

To help analysts analyse the data quickly and interpret it correctly, the data manager can recode two separate and clearly labelled variables, "multiplesa" (multiple partners among those who are currently sexually active) and "multipleall" (multiple partners among all members of the population). This document suggests that by convention, the suffix "all" is used for any variable whose denominator includes all respondents, while "sa" is used for variables where only the sexually active are included. In populations where current sexual activity is part of the inclusion criteria for surveillance (sex workers or MSM, for example), these two variables will be identical. It is still worth coding them both ways, since you may want to use either one for comparison across populations.

A note on "lazy" recodes

With large and complex data sets, where recodes can run to several thousand lines, it is often tempting to take a "short cut". One common short cut for variables in which we would like everyone in the denominator is to generate the variable to 0 (so that everyone is in the denominator) and then simply "fill up" the numerator.

This is a **VERY BAD IDEA**. It leads to all manner of mistakes — sub-populations who should not be included in the variable at all get coded to zero, people for whom data was missing are assumed to have a negative value, people who should be in the numerator may be missed but we will not easily see that because they do not appear as missing values during the recodes, etc. etc. etc.

It takes far, far less time to write "thorough" and correct codes in the first place than it does to try and track down mistakes in recodes when you discover them during analysis, when you get results that simply don't make sense.

However tempting it may seem to take shortcuts: **DON'T DO IT.**

EXERCISE 5: STANDARDISING DENOMINATORS FOR ANALYSIS

Programme managers have said they intend to use BSS data:

1) to estimate the overall risk of HIV transmission in the MSM population in the last month

2) to asses the success of condom promotion programmes

From the following questionnaire sequence, create variables with appropriate denominators to meet both of these needs.

MSM questionnaire Block 3

q1) Have you had sex with a man in the last month?

1) Yes 2) No (go to Block 4)

q2) What type of sex with a man in the last month?

1) Oral only (go to b3q4) 2) Anal only 3)Oral and anal

q3) How often did you use condoms in anal sex last month?

1) Never 2) Occasionally 3) Often 4) Always

Ensuring that numerators and denominators are comparable across populations and time

Behavioural surveillance is a relatively new science; inevitably, the practice changes as countries become more experienced. New groups are added to surveillance, old groups may be dropped, and there are usually some changes to questionnaires. These changes often mean that datasets are not exactly comparable from one year to the next. In addition, questions may be asked differently for different populations, because programming needs are different or the risk landscape is different.

Recoding can help minimise these differences, so that data can still be compared (even imperfectly) between populations. This type of recode often involves some judgments or decisions on the part of the data manager. **These must be carefully documented (see Step 10 below)**.

EXAMPLE 7: RECODING NUMERATORS AND DENOMINATORS FOR CONSISTENCY BETWEEN POPULATIONS AND OVER TIME

In the first round of surveillance, female and transvestite sex workers were asked separately about condom use with last non-regular client and with last regular client. The data were renamed “condomlastreg” and “condomlastnonreg” for both populations. Data from all sites showed very little difference in risk in either population, and in the second round, the distinction was dropped, and the single “condom at last commercial sex” variable was called “condomlastcs”.

The results are no longer directly comparable. Like it or not, however, the data will still be used to compare trends in condom use over time. Two solutions are possible. The first is simply to pick one of the client types from the Round 1 questionnaire (probably non-regular clients only), and recode them to match (i.e. give the condom variables the same name) as that used for **all** clients in the Round 2.

THE LAZY APPROACH

replace condomlasts= condomlastnonreg if round = = 1

The problem is, a not-inconsiderable proportion of our sex workers may have had only regular clients in the last week. They would all drop out of the denominator for the critically important condom use variable if we used the lazy approach above, and so they would also be excluded from any analysis about levels of condom use — one of our most important programme indicators.

THE BETTER APPROACH

We can do better, by combining condom use at last sex with regular client and condom use at last sex with non-regular client. Where the sex worker only has one type of client, or where condom use with both types of client was the same (both yes or both no) we have no problem. The problem arises where a condom was used with one type of client and not the other, because we don’t know which was the most recent client.

The first step is to look at the data and see how big the problem is likely to be. The fastest way is probably to create a variable that captures the discordance.

```
gen discordant = .

keep if (target = = 10 | target = = 41)

replace discordant = 0 if (condomlastreg = = condomlastnonreg) & round = = 1

replace discordant = 1 if condomlastreg = = 1 & condomlastnonreg = = 0

replace discordant = 2 if condomlastreg = = 0 & condomlastnonreg = = 1
```

```
tab discordant

discordant      |  Freq.    Percent     Cum.
----------------+--------------------------------
     0          |  867      86.27       86.27
       1        |  44       4.38        90.65
       2        |  94       9.35        100.00
----------------+--------------------------------
     Total      |  1,005    100.00
```

So around 15 percent of the whole population reports different condom use with regular and non-regular clients; twice as many in the direction of using condoms with non-regular but not with regular clients.

The three options for coding these discordant data are

1) code as non-users those who did not use with regular but did use with non-regular. This would give us the lowest level of risk

2) code as non-users those who did not use with non-regular but did use regular. This would give us the middling level of risk

3) code as non-users those who did not use with either one of the two partner types. This would give the highest level of risk.

This is a matter for judgment. Generally, it is safest to code for the higher level of risk, to avoid overstating programme impact.

So the conditions for saying that a sex worker did NOT use a condom with her last client (of those with clients in the last week) are:

1) she had both regular and non-regular clients, and had unprotected sex with the last client of both types

OR

2) she had only regular clients, and did not use a condom with the last one. She had no non-regular clients

OR

3) she had only non-regular clients, and did not use a condom with the last one. She had no regular clients

OR (following the decision above)

4) she had both regular and non-regular clients, and had unprotected sex with EITHER her last regular client OR her last non-regular client.

In brief, those conditions boil down to just one: she must have had unprotected sex with either a regular or a non-regular client in the last week.

The conditions for saying that a sex worker DID use a condom with her last client (of those with clients in the last week) are as follows:

1) she had both regular and non-regular clients, and used a condom with the last client of both types

OR

2) she had only regular clients, and used a condom with the last one. She had no non-regular clients

OR

3) she had only non-regular clients, and used a condom with the last one. She had no regular clients

```
gen condomlastcs1 = .
```

Note that we cannot directly generate condomlastcs because it already exists for other populations in the dataset.

```
replace condomlastcs1 = 0 if condomlastreg == 0 | condomlastnonreg == 0

replace condomlastcs1 = 1 if (condomlastreg == 1 & condomlastnonreg == 1) |

(condomlastreg == 1 & condomlastnonreg == .) | (condomlastreg == . & condomlastnonreg == 1)

replace condomlastcs = condomlastcs1 if round == 1 & (target == 10 | target == 41)
```

It is vitally important that we note what we have done, not just in the do-file (which may not be available to all users), but in the dataset itself. Stata has a notes function, which is described in detail under Step 10.

EXERCISE 6: RECODING DENOMINATORS AND NUMERATORS FOR CONSISTENCY

In the first round of surveillance, female and transvestite sex workers were asked separately about the consistency of condom use with non-regular clients and with regular clients over the past week. The data were renamed "consiscondomreg" and "consiscondomnonreg" for both populations, and were coded 1 "Always" 2 "Often" 3 "Occasionally" 4 "Never". Denominators for different client types had already been standardised, as "nonregall" and "regall".

In the second round, the distinction was dropped, and the single "consistency of condom use last week" variable was called "consiscondomcs"

In addition, in both the first and the second rounds, male sex workers were asked about consistency of condom use with all clients last month. The variable, called "consiscondomsellsexman" was coded 1 "Never" 2 "Sometimes" 3 "Always"

You want to be able to compare levels of consistent condom use between sex working populations and over time. Write down the decisions you have to make before combining the data. In the master data set, make a variable that would serve this purpose.

How to cope with missing data

Missing data can be a real problems when it comes to recoding and analysis. Data can be missing for several reasons. If data are missing because of a skip pattern (as in several of the examples above) there is no problem. But if data are missing because the questionnaire was not completed correctly, there were errors in data entry, or the respondent did not answer a question, then the situation is more complex.

In most questionnaires, there are options including "don't know" (often coded to 88 or 98) and "no response" (often coded to 99). These codes make for very clumsy analysis outputs, and the tendency of some programmers is to code them to missing. This is usually not the most appropriate course of action.

In general, if you are using a number of variables in a recode or in analysis, any respondent who has missing data on **any one of the variables** will be dropped from the new variable or from the analysis. While we can usually tolerate a certain amount of missing data, composite indicators can lead to quite high proportions of missing data. So it is in our interests to try to minimise missing data. Judgement, common sense and good documentation are the three most important ingredients in minimising missing data. Steps to follow include:

- Check the number of missing observations

  tab numcsmonth, **m**

  will give the distribution of responses by variable value, including for value "." (missing). Look carefully for values assigned in the questionnaire to "don't know", "don't remember" and "no response"

- Think about what the raw variable means, what you will use the recoded variable for, and whether you can make any "common sense" assumptions in your recodes. For example there may be a question:

"How do you rate your own risk for HIV?" 1) High risk" 2 "Some risk" "3 Little or no risk" 8 "Don't know" 9 "No response"

You want to generate a variable "feelrisk" — a yes/no indication of whether people feel at risk for HIV, which you will use in analyses to see whether people who have high risk behaviour feel that are at risk for HIV. In this case, common sense suggests that the "Don't know" responses can be coded to "0" (no), because if someone does not know if they are at risk, then they do not feel at risk.

- Make "sensible" assumptions about responses. Imagine 18% of respondents reply "Don't remember" (code 97) to the question "How many sex workers have you had sex with in the last month. If we want to calculate the mean number of sex workers per client, you cannot leave this as coded to 97. Stata will treat this as though 18% of your sample have visited 97 sex workers in the last month, and the mean will be absurdly high. But if you simply drop these people (recode numcsmonth 97=.) then you will be calculating the mean for only 82% of your client population. What is the most likely reason that someone can't remember how many sex workers they visited in the last month? Probably, they can't remember because there were so many. If you drop all the people who can't remember, you may be underestimating the average number of sex workers per client. One solution is to look at the distribution for people who can remember, and assign those who can't remember to the higher end of the distribution (for example the 75th or 80th percentile). Stata can do this automatically if you tell it what centile to use

centile numcsmonth if numcsmonth<97, **centile(80)**

replace numcsmonth=**r(c_1)** if numcsmonth>=97 & numcsmonth~=.

- Document any assumptions you have made (see below)

Step 9: Order the variables in your data set

Variables will appear in the variable window in Stata in the same order as they are found in the data set. This depends on entry, appending and recoding. When you rename a variable, it stays in its original location in the dataset. If you append another dataset, those variables appear below the ones in the original dataset. If you generate a new variable, it goes to the end of the dataset (in the data editor) and to the bottom of the variable list. This is very inconvenient — we would like the variables to appear in a more orderly way, so that we can quickly check the variable list if we have forgotten the name of a particular variable or if we want to check the spelling.

Stata allows you to arrange the variables in a data set in any order you want. Once you have finished all the recoding work, simply type "order" and the names of the variables in the order you would like them to appear.

order round target sex age agegroup

Step 10: Labelling and annotating the dataset

Almost all variables will need to be labelled, and many will also need to be carefully annotated to explain the assumptions that have been made in the recoding process, any differences in TR periods etc. The process of labelling and annotating is especially important in complex, combined data sets, because they no longer relate directly to the questionnaires. Unless they are clearly labelled, users will not be able to conduct meaningful analysis.

Many people find it intuitively easier to label variables during the recoding process, at the time that they are creating the variables. However we recommend doing all the labelling in a separate do-file, *after* the recoding has taken place. This is largely for two practical reasons. Firstly, a single label file removes the risk of unintended duplication and the possibility that one set of data labels will unintentionally overwrite another. Secondly, and importantly, it is far easier to translate data labels if they are all in a single file, without other recoding information in between. If necessary, data managers can write the label file as a text file at the time they are doing the recoding in the recode do-file, and then convert the label text file to a do-file once it is complete.

With Stata, labelling is a three step process.

1. The variable itself has to be labelled, to tell the user what it is about.
2. The value labels have to be defined, so that we know what the categories within the variable (1, 2, 3 etc.) stand for.
3. We have to assign the value label to the variable.

In addition, data managers may wish to add labels in more than one language, and to add notes to the variable giving more detail than is possible in a variable label.

Labelling the variable

Use the “label variable” (or “lab var”) command, the name of the variable to be labelled, and then the label, in quotation marks. A variable label should contain as much information as possible, in as few characters as possible (the maximum in Stata is 80). Labels should usually contain information on:

- What the variable refers to
- What time period it refers to
- Who is in the denominator

“Used a condom” is NOT a good label. “Used condom at last commercial sex, of those with commercial sex last 12 months” is a good label. Note that when you are combining data from different data sets, there may be differences in TR. Write these in the label. If all the necessary details do not fit within 80 characters, use notes (see below).

lab var alwayscondomcs, “Used condom in all com. sex, of those w cs (FSW 1 week, msm 1 month, client 1 year)”

A very small number of variable names, such as age, sex, city or year may be self-explanatory and need no label in the language of origin, though most will still need to be labelled in other languages.

Define labels for values within a variable

Define labels for the values that go with the variables, using “label define” (or “lab def”). These will appear in the output and analysis instead of the underlying values which appear in the dataset.

Some value labels apply to many variables (e.g. 0 = no, 1 = yes, or 1 = never, 2 = occasionally, 3 = often, 4 = always). You can set these values just once in the dataset, and give them a name. You can then apply that set of label names to any number of variables. e.g., create a labels set called “yesno”, as follows:

lab def yesno 0 “No” 1 “Yes” 98 “Don’t know”

Many sets of labels will refer only to one variable. In this case we tend to call the label set by the same name as the variable to which it belongs. e.g.

lab def ethnic 1 “Asian” 2 “Afro-Caribbean” 3 “Caucasian” 4 “Other”

Assign value labels to a variable

Once you have defined a label set, you have to assign it to a variable before it will appear in the output. Type “label variable” or “lab var”, then the name of the variable you are assigning a label to, then the name of the label.

lab val ethnic ethnic

lab val everidu yesno

Note that numerical variables (e.g. age, income in the last month) do not need labels.

Add notes to give more details of recodes

Stata has a notes function that allows you to attach notes either to the data set in general, or to specific variables. This allows you to give information about the data set as well as about decisions made in recoding.

notes: Information about sampling; Female and transvestite sex worker samples were drawn following a complete mapping of commercial sex locations. Clusters were drawn at random from the sample frame with probability proportional to size, and a fixed number of sex workers were interviewed at each selected cluster. Respondent driven sampling was used for IDU populations. For more information on sampling, contact bss@statisticsbureau.gov”

notes numcsmonth: “for male respondents, don’t know reassigned to 80th percentile”

notes condomlastcs: “Round one has separate variables for regular and non-regular partners for FSW and transvestites. Condomlastcs is coded 0 if respondent did not use a condom with either one of these partner types”

To see your notes, type “notes” (to see all notes in the data set) or by variable:

notes condomlastcs

Notes do not appear automatically, so it is good to mention in the variable label that they are there.

lab var condomlastcs “used a condom at last commercial sex (notes), (FSW, transvestite 1 week, msw client 1 month,)”

These notes may be lost if the dataset is translated into a non-Stata format.

One dataset, many languages

Note that it is possible in Stata to create variable and value labels in more than one language. First tell Stata what its default language is. For example you may want to make the default language English but add labels in Spanish.

label language en, rename

will set the default language to English. Unless you tell it otherwise, Stata will assume the labels you are typing are in English (the programme itself has no language capacity, it accepts anything you enter, and would equally accept one set of labels for a “language” called female sex workers (FSW) and another for male (msw).

Once you have a full, labelled data set, create a new label language, for example Spanish (es). If you have done all your labelling in a single do file, this is relatively easy. Simply copy the label do file, and save it in another name.

Make sure that you have already set the default language to your original working language.

label language en, rename

Now simply start the new file by making another set of labels that is a copy of the first:

label language es, new copy

Your current working language for labels will now be Spanish. You now have to redefine all the variable and value labels in Spanish. You can do that by adding “_es” to the end of each of the label definitions in your existing label file, and then translating the labels. For value labels and notes, you don’t need to add the _es; since your working language is now Spanish, you can simply translate the label variables and notes.

lab def yesno_es 0 “No” 1 “Si” 98 “No se”

lab var condomlastcs “utilizado preservativo en el ultimo sexo comercial, (mujeres, transgeneros 1 semana, gay, clientes 1 mes)”

You do not need to repeat the task of assigning the defined labels to specific variables — once a label has been assigned in the default language, it will automatically be assigned in any other languages which are added.

If you want to switch between languages while using the data set, just specify the language you want to use.

lab lang en

will take you back to English labels, for example.

Labels in other languages may be lost if the dataset is translated into a non-Stata format, so you may want to save separate versions of the data set (one for each language) before converting. See Help label_language.

If you want to get a list of all the labels in your dataset, the values they carry and which variables they are affixed to, type

labelbook

You can also use this just to look at labels for specific variables.

labelbook province

This is a much easier way of checking the underlying numerical values of variables with a large number of values than “tab province, nolabel”. Unless you specify otherwise, “labelbook” will give labels in all the languages in the dataset.

Similarly, you can use

codebook

to give variable labels, either for the whole data set or for individual variables. If you want them in all specified languages, use the “languages” option

codebook condomlastcs, lang

EXERCISE 7: LABELLING DATA

Go back to Exercise 6.

Label the variable you have created, with appropriate value labels and notes.

Add labels in a second language.

Write out the commands you would use to check that the labels are correct in the dataset.

Chapter 5: Documentation

Throughout these pages and the exercises, the importance of documentation of data management and recoding has been stressed. Without careful documentation of the process of recoding, including assumptions made and the reasons for those assumptions, future users may not be able to recreate results. This may lead to inappropriate comparisons between populations and over time, and other errors of analysis and interpretation.

Documentation checklist

- **Original raw data set**

  It is absolutely critical that copies of the raw data set received from data entry be kept by data managers, in a separate and protected directory location.

- **Cleaning notes**

  Data managers should keep a log of changes made in the process of data cleaning, including inconsistencies between data that have been double-entered by different operators, how they were resolved, assumptions made in resolving internal inconsistencies within a questionnaire, additional codes added as a result of open questions, etc.

- **Code book - raw data**

  Once a clean set of raw data are available, it should be accompanied by a full code book.

- **Do-files for data management, including coding and combining data sets**

  Annotated do-files recording every action taken while coding and combining data sets should be carefully preserved.

- **Full data labels, in appropriate languages**

  Combined data sets should be carefully labelled (including full, appropriate variable labels and value labels) in every language in which analysis will be performed

- **Annotations WITHIN the data set**

  Important data transformations, especially those involving judgement on the part of the data managers or analysts, should be annotated within the data set itself. Notes should be attached to specific variables and to the data set as a whole, as appropriate.

- **Code book — recoded data**

  Once a combined, recoded data set is available, variables can no longer be traced back to the original questionnaires. These types of datasets MUST be accompanied by a full codebook. The code book should include each of the label languages, as well as any recoding notes.

Appendix 1 — Solutions to exercises

POSSIBLE SOLUTION TO EXERCISE 1

```
clear

set mem 250m

cd "c:\surveillance\behaviour\data\raw data round 1\FSW"

use paris

save parisrecode.dta, replace

gen sex =2

gen city = 1

gen round = 1
```

Note that in this case we define "Round" rather than "Year" for comparing trends over time, since both rounds span 2 years.

```
gen target = 10

gen idnum = target*100,000+city*10,000+round*1,000+respno

save, replace

clear

use lyon

save lyonrecode.dta, replace

gen sex =2

gen city = 5

gen round = 1

gen target = 10

gen idnum = target*100,000+city*10,000+round*1,000+respno

save, replace

clear

cd "c:\surveillance\behaviour\data\raw data round 2\FSW"
```

Note that we need to switch directories here, because Round 1 and Round 2 data are in separate directories.

```
use lyon

save lyonrecode.dta, replace

gen sex =2

gen city = 5

gen round = 2

gen target = 10

gen idnum = target*100,000+city*10,000+round*1,000+respno

save, replace

clear

cd "c:\surveillance\behaviour\data\raw data round 2\idu"

use lyon

save lyonrecode.dta, replace

ren sex = b1q1
```

Note that the IDU population is both male and female. Sex will already be included in the questionnaire and coded 1 for male and 2 for female, so we do not need to generate a new variable, merely rename the existing sex variable.

```
gen city = 5

gen round = 2

gen target = 41

gen idnum = target*100,000+city*10,000+round*1,000+respno

save, replace
```

Note that this process is quite repetitive. Later, we will learn to minimise repetition by combining datasets before adding variables (such as idnum in this example) which are common for all data sets, as well as to keep directories tidy by eliminating unwanted files. See Exercise 3.

POSSIBLE SOLUTION TO EXERCISE 2

```
clear

set mem 250m

cd "c:\surveillance\behaviour\data"
```

```
use FSWdata

save FSWrecode, replace

gen clientswho = b3q4
```

Note that we are generating a new variable rather than simply renaming the existing variable because we are going to make substantial changes to the values in some cases. If we rename, we will not be able to go back and look at the original if we want to.

```
recode clientswho 5=7 6=8

save, replace

clear

use mswdata

save mswrecode, replace

gen clientswho = b3q6

recode clientswho 1=5 2=1 3=7 4=8

save, replace

clear

use transvestitedata

save transvestiterecode, replace

gen clientswho = b4q12

recode clientswho 1=6 2=4 4=2 5=7 6=8

save, replace
```

*AFTER APPENDING DATA, YOU HAVE TO LABEL THE VARIABLE

```
use alldata.dta

lab var clientswho “occupation of majority of clients reported by sex workers”

lab def clientswho 1 “Civil servants” 2 “Military” 3 “Truck taxi/drivers” 4 “Sailors/
dockworkers” 5 “Businessmen” 6 “Students” 7 “Other” 8 “Don’t know”

lab val clientswho clientswho

save alldata.dta, replace
```

Note that in this example, data managers have decided to combine military and police, as well as sailors and dockworkers and truck drivers and taxi drivers. Such combinations are a matter of judgement, but should be based in large part with programme management priorities in mind. If potential interventions for truck drivers and taxi drivers are very different, then it may not be appropriate to combine them, because we may need to know which of the two is the majority in this case.

POSSIBLE SOLUTION TO EXERCISE 3

```
clear
set mem 250m
cd "c:\surveillance\behaviour\data\combined data"
use "c:\surveillance\behaviour\data\raw data round 1\FSW\paris.dta"
save "c:\surveillance\behaviour\data\raw data round 1\FSW\FSWparisrecode.dta", replace
gen city = 1
gen round = 1
save, replace
clear
use "c:\surveillance\behaviour\data\raw data round 1\FSW\lyon.dta"
save "c:\surveillance\behaviour\data\raw data round 1\FSW\FSWlyonrecode.dta",
gen city = 5
gen round = 1
append using "c:\surveillance\behaviour\data\raw data round 1\ FSW\FSWparisrecode.
dta", replace
```

Note that the Lyon dataset us still in memory, so we can immediately **replace** append the Paris dataset (which we saved earlier after labelling it), and save the joint dataset in a new name, then erase the unwanted files.

```
save "c:\surveillance\behaviour\data\combined data\FSWround1all.dta", replace
erase "c:\surveillance\behaviour\data\raw data round 1\FSW\FSWparisrecode.dta"
erase "c:\surveillance\behaviour\data\raw data round 1\FSW\FSWlyonrecode.dta"
```

Now we need to deal with the round 2 data for sex workers:

```
clear
use "c:\surveillance\behaviour\data\raw data round 2\FSW\lyon.dta"
save "c:\surveillance\behaviour\data\raw data round 2\FSW\FSWlyonrecode.dta"
```

```
gen city = 5

gen round = 2

append using "c:\surveillance\behaviour\data\combined data\FSWround1all.dta"
```

Again, the Lyon data are still in memory, so we can immediately append all of the Paris data and save in a different name.

```
save "c:\surveillance\behaviour\data\combined data\FSWall.dta", replace

erase "c:\surveillance\behaviour\data\raw data round 2\FSW\FSWlyonrecode.dta"

erase "c:\surveillance\behaviour\data\combined data\FSWround1all.dta"
```

Now that all the sex worker data are coded for round and city and combined, we can add codes for the things common to all female sex worker data sets.

```
gen sex = 2

gen target = 10

save, replace

clear

use "c:\surveillance\behaviour\data\raw data round 2\IDU\lyon.dta"

save "c:\surveillance\behaviour\data\combined data\IDUlyonrecode.dta", replace

ren sex = b1q1
```

The IDU dataset includes both men and women, so there will certainly already be a variable for sex, coded 1 for male and 2 for female. We don't need to create it; we can just rename it.

```
gen city = 5

gen round = 2

gen target = 41

append using "c:\surveillance\behaviour\data\combined data\FSWall.dta"

save "c:\surveillance\behaviour\data\combined data\masterbss.dta", replace

erase "c:\surveillance\behaviour\data\combined data\IDUlyonrecode.dta"

erase "c:\surveillance\behaviour\data\combined data\FSWall.dta"
```

Now all the data are in a single file, we can go ahead and generate any variables which are common to all the data sets.

```
gen idnum = target*100,000+city*10,000+round*1,000+respno

save, replace
```

POSSIBLE SOLUTION TO EXERCISE 4

```
clear

set mem 250m

use “c:\surveillance\behaviour\data\combined data\allcombined.dta”

save “c:\surveillance\behaviour\data\combined data\masterbss.dta”, replace

sort target

by target: summarize price

tabstat price, by (target) statistics (min p25 p50 p75 max)
```

This gives us the output printed above. It is strange that for nearly all groups some people say they had commercial sex for free. Before coding we would want to check and see that this is not a large proportion of the population, which would suggest errors in the data. Also, the 999 for male sex workers is suspicious, and probably indicates missing values. We want to check to see if other groups (whose maximum is higher than 999) also have suspicious values of 998 or 999.

```
tab target if price == 0

tab target price if price > 997

replace price = . if price == 999 and target == 42

gen pricegroup = price

recode pricegroup min/50 = 1 51/75 = 2 76/150 = 3 151/max = 4

tab target pricegroup, row

save, replace
```

POSSIBLE SOLUTION TO EXERCISE 5

```
use msmdata

rename b3q1 msmmonth

recode msmmonth 2=0
```

```
rename b3q2 sextype
```

```
rename b3q3 consiscondomanal
```

First recode: a variable which can be used to estimate the overall levels of high risk sex between men. Epidemiologists will be concerned with the proportion of the total population that is engaging in unprotected anal sex. So the numerator should be those who report unprotected anal sex, while the denominator should be all respondents. Right now, only men who have had anal sex in the last month are asked about use of condoms in anal sex . To get everyone into the denominator, we need to "put back" those who were removed by the skip patterns, i.e. those who have not had any sex with a man in the last month, and those who have not had anal sex with a man in the last month. We also need to think about assigning respondents to b3q3 correctly.

```
generate anyunprotectedanalall = .
```

Think about those who did NOT have unprotected anal sex. This includes

1) those who did not have sex

OR

2) those who had sex but did not have anal sex

OR

3) those who had anal sex but always used condoms

A person is not at risk if they fulfil any one of these criteria, so we use the operator OR

```
replace anyunprotectedanalall = 0 if msmmonth == 0 | sextype == 1 | consiscondomanal
== 4
```

Now think about those who DID have unprotected anal sex. This includes only those who

1) Had sex in the last month

AND

2) Had anal sex in the last month

AND

3) Did not always use a condom in anal sex

A person has to fulfil all three of these conditions to be at risk, so we use the operator AND. The most thorough way to write this code would be:

```
replace anyunprotectedanalall = 1 if msmmonth == 1 & (sextype == 2 | sextype ==3) & consiscondomanal < 4
```

However because people only get asked question b3q3 if they already fulfil the preceding criteria of having had sex in the last month and having had anal sex, a simpler way to achieve the same result would be

```
replace anyunprotectedanalall = 1 if consiscondomanal < 4
```

Second recode: A programme manager who is planning or evaluating condom promotion for MSM is less concerned with those members of the MSM community who don't have anal sex, since people who don't have anal sex are less in need of condoms than those who do have anal sex. For them, the more useful variable is the proportion *of those who have anal sex* who are using condoms. From the questionnaire, we get this directly in the variable b3q3.

If we start from scratch, we have to ensure that all the people who had anal sex get in to the variable. In this case the criteria for a negative code (they did NOT have unprotected anal sex) are:

1) Had anal sex

AND

2) Always used condoms.

```
generate anyunprotectedanalsa = .

replace anyunprotectedanalsa = 0 if (sextype == 2 | sextype == 3) & consiscondomanal == 4
```

The criteria for a positive code (they DID have unprotected anal sex) are:

1) Had anal sex

AND

2) Did not always use condoms.

```
replace anyunprotectedanalsa = 1 if (sextype == 2 | sextype == 3) & consiscondomanal < 4
```

In fact, when the denominator of a recoded variable is identical to the denominator of a variable in the raw data, there is another, simpler way of doing recodes. In this case, we can generate a new variable which will exactly replicate the existing variable, and then simply recode the response values, as we did in Exercise 2.

```
generate anyunprotectedanalsa = consiscondomanal
```

We now have two identical variables in the dataset, with different names. (This is different from renaming a variable, because when you rename, the original variable is lost. In this case we want to keep the original variable, because we still want to use it in other recoding and in analysis)

The denominator for anyunprotectedanalsa is now only the people who were asked about consistency of condom use, i.e. only those who had anal sex. All we have to do is turn this 4 response variable into a yes/no variable by recoding the response values.

```
recode anyunprotectedanalsa 1/3=0 4=1

save msmrecoded, replace
```

POSSIBLE SOLUTION TO EXERCISE 6

```
clear

set mem 250m

use “c:\surveillance\behaviour\data\combined data\allcombined.dta”

save “c:\surveillance\behaviour\data\combined data\masterbss.dta”, replace

tab target consiscondomreg

tab target consiscondomnonreg

tab target consiscondomcs

tab round consiscondomsellsexman
```

Decisions to be made: For some populations we have four values, and for others we have three values. We can either stick with 4, and choose to code the “sometimes” among male sex workers to either “occasionally” or “often”, or we can code both “occasionally” and “often” to “sometimes”, in accordance with the male sex worker data. If we choose the first course, we have to decide what to do when respondents use condoms “occasionally” with one type of partner and “often” with another type of partner.

The choice will depend in part on where you are in the epidemic. At earlier stages, when a lot of people are not using condoms at all, the shift between never and sometimes will be important, and we can go for the first choice, which is the easiest option. In later stages of the epidemic, when a higher proportion of people are using condoms at least occasionally, it may be more important to measure the shift from occasional condom use to frequent condom use. As a rule of thumb in recoding, it is always a good idea to make the slightly harder choice, because it saves having to go back and change it later on as the epidemic and the response develop. In other words, we should try to create a variable with four levels for all groups.

As we saw earlier, it is also usually a good idea to code “conservatively” for highest risk. In this case, we would decide to code to occasional use those who used condoms only occasionally with one client type, even if they used condoms more consistently with other partner types.

So for round one female and transvestite sex workers:

Conditions for always use:

1) Always used a condom with BOTH regular AND non-regular clients in the last week

OR

2) Always used a condom with regular clients in the last week, AND had no non-regular clients in the last week

OR

3) Always used a condom with non-regular clients in the last week, AND had no regular clients in the last week

Conditions for often use:

1) Often used a condom with BOTH regular AND non-regular clients in the last week

OR

2) Often used a condom with regular clients in the last week, AND had no non-regular clients in the last week

OR

3) Always used a condom with regular clients in the last week, AND often used a condom with non-regular clients in the last week

OR

4) Often used a condom with regular clients in the last week, AND always used a condom with non-regular clients in the last week

OR

5) Always used a condom with non-regular clients in the last week, AND had no regular clients in the last week

Conditions for occasional use:

1) Occasionally used a condom with EITHER regular OR non-regular client in the last week (regardless of level of condom use with any remaining client type)

Conditions for never use:

1) Never used a condom with BOTH regular AND non-regular clients in the last week

OR

2) Never used a condom with regular clients in the last week, AND had no non-regular clients in the last week

OR

3) Never used a condom with non-regular clients in the last week, AND had no regular clients in the last week

consiscondomcs already exists in the data for round 2, so we have to give the data we are recoding a different name.

```
gen consiscondomcs2 = .
```

Because it is easier to read low to high values in data sets, especially where we expect a dose-response relationship with interventions, we want to switch the values in this variable so that they go from 1 “Never” to 4 “Always”

```
replace consiscondomcs2 = 1 if (consiscondomreg == 4 & nonregall == 0) | (consiscondomnonreg == 4 & regall == 0) | (consiscondomreg == 4 & consiscondomnonreg == 4)
```

```
replace consiscondomcs2 = 2 if (consiscondomreg == 3 | consiscondomnonreg == 3)
```

```
replace consiscondomcs2 = 3 if (consiscondomreg == 2 & nonregall == 0) | (consiscondomnonreg == 2 & regall == 0) | (consiscondomreg == 2 & consiscondomnonreg == 2) | (consiscondomreg == 2 & consiscondomnonreg == 1) | (consiscondomreg == 1 & consiscondomnonreg == 2)
```

```
replace consiscondomcs2 = 4 if (consiscondomreg == 1 & numnonregweek == 0) | (consiscondomnonreg == 1 & numregweek == 0) | (consiscondomreg == 1 & consiscondomnonreg == 1)
```

Now check the variable

```
tab consiscondomcs2
```

Don’t forget that in the second round the distinction between client types has been dropped, but the value codes are still the same as they were in the first round — the opposite of what we want. So we need to switch them:

```
recode consiscondomcs 4=1 3=2 2=3 1=4
```

Now we can merge the two

```
replace consiscondomcs = consiscondomcs2 if round == 1 & (target == 10 | target == 41)
```

Now get rid of the interim variable, to avoid confusion later on.

```
drop consiscondomcs2
```

We still have to deal with the issue of male sex workers, who were only coded as “never, sometimes, always”. What criteria should we use to assign them to “occasionally” or “often”? As always, the first step is to look at the data.

```
tab consiscondomsellsexman condomlastsellsexman, row
```

```
                    last time use
consistent use   |   no        yes      |    Total
-----------------+----------------------+-----------
    never        |   894       3             897
                 |   99.67     0.33     |    100.00
-----------------+----------------------+-----------
sometimes        |   51        32       |    81
                 |   62.96     39.50    |    100.00
-----------------+----------------------+-----------
   always        |   1         43       |    44
                 |   2.27      97.73    |    100.00
-----------------+----------------------+-----------
    Total        |    949      73       |   1,022
                 |   92.86    7.14      |  100.00
```

We can see that in this population, condom use in commercial anal sex is very low indeed — overall, only 4% of respondents say they always using condoms in commercial sex with male clients, and 88% never do. In this case, there seem to be two choices. One would be simply to assign all of the “sometimes” to “rarely” (because it is likely in this context that even those who sometimes use condoms don’t do so very frequently). The other would be to assign those who DID use a condom last time to “often” and those who DID NOT use a condom last time to “occasionally”.

```
gen consiscondomcs3 = consiscondomsellsexman
```

First, get the “always” out of the way by shifting them to the value 4

```
recode consiscondomcs3 3=4

replace consiscondomcs3 = 2 if consiscondomsellsexman = = 2 & condomlastsellsexman = = 0

replace consiscondomcs3 = 3 if consiscondomsellsexman = = 2 & condomlastsellsexman = = 1

tab consiscondomcs3

replace consiscondomcs = consiscondomcs3 if target == 43
```

POSSIBLE SOLUTION TO EXERCISE 7

```
label language en

lab var consiscondomcs “freq of condom use with all recent clients, of sex workers (notes) FSW, transv 1 week, msw 1 month”
```

lab def consis 1 “never” 2 “occasionally” 3 “often” 4 “always”

lab val consiscondomcs consis

note consiscondomcs “in round 1, FSW and transv sw were asked separately about reg and non-reg clients. for round 1, this recode assigns those who used condoms ocassionally with any partner type to occasional. note questionnaire values inverted. for msw, only three values in quest. those who replied “sometimes” and did not use condoms with last partner assigned to occasional, those who did use condoms with last partner assigned to often.”

label language fr, new copy

lab var consiscondomcs “freq. d’utilisation de preservatif avec clients, tous travailleurs de sexe, (note) femmes, transv 1 semaine, msw 1 mois”

lab def consis_fr 1 “jamais” 2 “parfois” 3 “souvent” 4 “toujours”

note consiscondomcs “permiere annee, donnees pour vendeuses de sexe femmes et transvestis sont separees par type de client (regulier ou pas regulier). ici, ceux qui ont utilisees preservatif “parfois” avec client quelquonque sont codees comme parfois. valeurs l’inverse du questionnaire. pour vendeurs de sexe hommes, seulement trois valeurs. ceux qui disent “quelquefois” et n’ont pas utilise preservatif avec dernier client codes comme parfois, ceux qui disent “quelquefois” et ont utilise preservatif avec dernier client codes comme souvent.”

codebook consiscondomcs, languages

labelbook consiscondomcs

Appendix 2 — Getting started with Stata

Stata is a powerful statistical analysis package that works by typing in commands. It provides a quick and easy way to recode variables, to combine data sets, to analyse data, and to run statistical tests. You can do all of these things by learning a few simple commands. Some people who are used to using menu-driven packages like SPSS find Stata awkward at first. However command driven programmes tend to push us to think more carefully about our analysis, and that is always a good thing. The main weakness of Stata is that it is not very good at producing graphics.

Stata has very good built-in help, and even better on-line help. Please use it often.

What is on the screen when you open stata?

There are four windows shown when you run Stata. At the top right is the review window, which stores all the commands you have recently typed in. Below that is the variables window, which shows the variables in your data set, including the label if you have given the variable a label. The biggest window, at the top left, is where the output of your analysis appears. Below it is the command window, where you type in your commands.

Reading in a data set

Stata works in memory. The first step is always to allocate memory on your computer to Stata for the working session. How much memory you allocate depends on the size of your data files and the capacity of your computer. It is usually safe to start with around 250m. So every Stata session begins with

```
clear
```

```
set memory 250m
```

The next step is always to read a data set from your files into memory. You can then work with that data set. Stata will not make permanent changes to the data set until you tell it to save the dataset. Once an altered data set has been saved, **you will not be able to recover the original**. It is therefore absolutely vital that you keep a protected version of the original data set, clearly labelled, in a separate location. It is also vital that as soon as you open your data set for a day's work, you save it in a different name. This means that if you unintentionally save the file during the course of your work, your original working file will be unchanged.

The default directory for Stata is the directory where the programme is installed. This is almost never where you keep your data files. It is a good idea to change to your working directory before reading in the data (using the old DOS command cd for 'change directory'. Any data sets or command files you save will be saved in the working directory, unless you specify a full directory path to another directory in the "save" command. Stata will also look in the working directory for any dataset that you try to open using the "use" command, again unless you specify a full path to another directory.

```
cd “c:\surveillance\behaviour\data”
```

You only have to set the memory and the directory once at the beginning of each working session. It will remain unchanged (unless you deliberately change it) until you close Stata. If you have data in separate folders (e.g. you keep separate folders for population groups or for rounds of surveillance) you can switch back and forth between them in your command files (do files).

```
cd "c:\surveillance\behaviour\data\round2"
```

```
use FSW2005.dta
```

Now immediately save the file in a different name, to protect the original.

```
save FSW2005recode.dta
```

You are ready to work.

Viewing your data

You can view your dataset by clicking on the data browser or data editor tab. You will see all the variables, with their values. Most variables are stored numerically (e.g. marital status may be stored as 1 for single, 2 for married, 3 for divorced and 4 for widowed). In some data sets, they will appear in the window with the label you have given them, but if you click on that cell and look at the top of the screen, you will see the number that underlies the label. If a value is missing, it appears as a full-stop **(.)**.

You should NEVER make changes directly in the datasets using the data editor. If you change the data in the data editor and save the changes, you will not be able to recover the original data or correct any mistakes. Always use a "do file" to edit your data (see below).

If you want to view the contents of just one or two variables, use the command "browse", followed by the names of the variables you want to view.

The importance of "do-files"

Virtually all your work in data coding and analysis will be done using "do-files". These are mini computer programmes which tell Stata what to do. You can run them over and over again on the same data set or on different data sets that have the same variable names (e.g. BSS data collected using the same questionnaire from different locations). Do-files allow you to:

- automate tasks that are performed repeatedly
- perform identical tasks on large numbers of data sets
- make mistakes and correct them without damaging the data set
- document exactly how a variable was coded and why
- share coding and analysis tasks easily with other users

Use the "do-file editor" (under "Window" in the file menu) to open an existing do-file or to create a new one. Don't forget to save it when you have finished working. To save the do-file, **make sure the do-file window is active**. Otherwise, you may save changes to your dataset by mistake.

All do-files begin the same way: clear, set memory, change directory, read in data file, save data file in different name.

There are two ways to use a do-file. The first is to "Run" it. You can do this by typing **run** and the do-file name, or (more usually) by clicking on the right-hand icon at the top of the do-file window (a blank sheet with an arrow). This will perform all the tasks in your do-file (until it finishes or hits an error), without anything appearing on the screen. You will know when it is finished when a full stop appears in the results window.

The second way is to "Do" it, by typing **do** and the do-file name, or clicking on the "do" icon at the top of the do-file window (a written-on sheet with an arrow). This will perform all the tasks in your do-file**,** and you will see everything that is being done come up in the results window.

Do file-tips

- "Running" do-files is much quicker than "doing" them. When we are coding, we rarely use "do". One advantage of "do"ing a file is that if there is an error, it shows you where the error is. In a run, you get an error message but cannot tell which line it refers to. In a long file, it is often difficult to find the error without "do"ing the file.

- Stata is case-sensitive. All the commands are typed using small letters. Because of this, we use CAPITAL LETTERS to add notes to our do-files, reminding ourselves what we are doing in coding or in the analysis.

When Stata sees * at the beginning of a line, it ignores the line, so all notes lines start with * (so that the programme does not try to carry out a command it does not recognise).

* WE CAN PUT IN NOTES IN THE DO-FILES IN CAPITALS. THIS HELPS *US TO FIND THINGS IN THE DO-FILE WHEN IT GETS VERY LONG

- You can skip single lines of a do-file by just putting * at the beginning of the line. You can skip whole sections by putting in a line at the beginning and at the end of the section you want to skip, as follows:

***/** *(nothing after this point will be executed, unless Stata is told to resume)*

/* *(this is the resume command. Stata will execute any commands after this line)*

- Do files can get very long indeed. In part because Stata has a character limit on do files, it is a good idea to keep separate do-files for separate tasks — e.g. renaming and combining files, recoding files, analysis. Note that you can run files within files. If you want to run all your do-files in sequence, begin your final do-file with

run "rename and combine.do"

run "recode demog and injecting risk.do"

run "recode sexual risk.do"

run "append label order.do"

etc

- You should use do-files for all recoding and analysis work. However you do NOT have to type in to your do file the commands that you will use to check that your code has come out the way you wanted it. These don't need to be kept for the future, and so can be typed directly into the command window in Stata.

OPERATORS IN STATA

These are the operators for Stata. Other packages will use very similar operators.

= = means "EQUAL TO"

& means AND, i.e. BOTH condition A AND condition B must be fulfilled

| means OR, i.e. EITHER condition A OR condition B must be fulfilled

~= means "NOT EQUAL TO", ..i.e. the condition must not be fulfilled

< means "LESS THAN"

> means "GREATER THAN"

Brackets: think carefully

replace potentialclient = 1 if (rich = = 1 | poor == 0) & blonde == 1

will code as potential clients only people who are rich OR are not poor, AND are blonde: in other words, all rich blondes. Rich brunettes are not potential clients with this code.

replace potentialclient = 1 if rich = = 1 | (poor == 0 & blonde == 1)

will code as potential clients people who are rich OR people who are not poor AND are blonde: in other words, all rich people, included but not limited to blondes. Rich people with dark hair would be considered potential clients with this code.

As a general rule, variables whose positive value is "filled up" using "AND" operators will have a negative value assigned using "OR" operators, and vice versa.

replace safe = 0 if unprotectedsex = = 1 | shareneedle = = 1

replace safe = 1 if unprotectedsex = = 0 & shareneedle = = 0

MATHEMATICAL TRANSFORMATIONS

Note that mathematical operations can be undertaken within many Stata commands. The most commonly used in BSS is when we are generating new variables, e.g.

gen totalpartnersall = numFSWall + numsfall + spouseall

This will generate a variable whose value will be the total of the values of each of the three variables cited *for those people who have non-missing values for all three variables*. Note that if someone has a missing value on any of the three, they will be assigned to missing. This is why we take trouble to generate variables with everyone in the denominator.

Another simple example of a mathematical transformation would be translating between months and years, e.g.

```
gen yearsIDU = monthsIDU / 12
```

Stata tips and oddities

- Stata does not accept spaces in file names unless you enclose the entire file path in quotation marks.

- If you want a single command to span over more than one physical line on the screen, you have to tell Stata that it is still a single line. The easiest way to do this is to type “///” at the place you want to break the line, and continue typing on the next line. Stata will assume that the next line is part of the same command.

- Missing values appear in Stata as “.” Despite this, the programme assigns them an unspecified but very high value — one of the great mysteries of life. This means you have to be **very careful when using the command >** (greater than). If you use > in a recode, for example

  ```
  replace multiple = 1 if totalpartners >=2
  ```

  then everyone who has missing values for the total number of partners, including those who never had sex at all and were therefore not asked questions about partner numbers, will get coded as if they had multiple partners. This problem does not arise if you use “max”, which includes only assigned numeric values. Note, however, that the use of “max” would still assign as having multiple partners anyone who had a coded non-response (e.g. 98, don’t know, 99 no response).

  ```
  gen totalpartners = multiple
  ```

  ```
  recode multiple 0/1 = 0 2/max = 1
  ```

- Be aware that the “recode” command performs all actions simultaneously, while “replace” does them one after another. So a subsequent command will act on the earlier command.

  ```
  recode rich 1=0 0=1
  ```

  will turn all people who were coded as rich in the data set into people coded as poor (i.e. not rich) and will turn the poor people into rich people. But

  ```
  replace rich = 1 if rich = 0
  ```

  ```
  replace rich = 0 if rich = 1
  ```

  will make everyone poor. You have first made everyone rich (all the 0s turned in to 1s, so now the whole population is a 1, “rich” and THEN you have turned all the 1s into 0s, so now the whole population will be rich = 0 i.e. poor.

- The prefix **by (*varname*):** can be used with many commands in Stata; it will perform the command separately for each of the categories in the "by" variable, and is an easy way of performing stratified analysis. With some commands (principally summary statistics), the "by" option comes after the comma, followed by the variable in brackets. The former is more frequently used at the analysis stage, the latter at the data exploration and coding stage.

  by target: tab intervention condomlastcs, row chi2

  tabstat age, by (target) statistics (min p50 max var)

  If you use these commands, data must be **sorted** by the "by" variable before performing the command.

  sort target

  An alternative in the first case is to use the single command "bysort", which simultaneously sorts and produces results stratified by the specified variable.

  bysort target: tab intervention condomlastcs, row chi2

- Be careful with file names and directories. If you want to change a directory, you have to use the cd command. You can get a file from another directory by using the whole path name (including the file name) in quotation marks after the "**use**" command, and you can save a file to another directory in the same way. Note that **neither of these actions will change your working directory**. Do-files and any files saved with commands that do not use a directory path will continue to be saved in your working directory.

COMBINING DATASETS HOLDING DATA FOR THE SAME INDIVIDUALS

If you have two datasets that hold data for the same individuals, for example one with behavioural data and one with biological data, they can be combined as long as they share at least one variable which is unique to each participant. Usually, linked behavioural and biological surveys assign a unique ID to each respondent, and it is this that is used to link the two data sets. You can check that unique IDs really are unique by using the commands "isid" or "duplicates" (see Stata Help).

The datasets you want to link have to be sorted according to the unique identifier (for example "idnum"). Keeping one in memory, then add in the other, using the command "merge".

EXAMPLE: MERGING BIOLOGICAL AND BSS DATASETS

clear

use masterbss.dta

sort idum

save, replace

```
clear

use masterlab.dta

sort idnum

merge idnum using masterbss.dta

save mastersurveillance.dta, replace
```

A NOTE ON WEIGHTING DATA

If a survey did not use simple random sampling (SRS) (if, for example, it used time-location sampling or was a clustered household survey) the data cannot be analysed without some supporting information. Depending on the type of survey this information might include, for each respondent, some or all of the following:

- the stratum
- the cluster (also called primary sampling unit or PSU)
- sample weight

Data from surveys which have used stratified or clustered sampling techniques should contain the stratum and/or cluster from which each respondent was selected. If the survey was not designed to be self-weighting there should be a sample weight for each observation in the dataset.

If a survey is clustered or stratified there should be a place on the questionnaire to record the stratum and cluster for each respondent. These should be coded before data entry and this information should be entered with each questionnaire. If sample weights are needed these must be calculated for each respondent, based on information about the entire sample frame, and added to each record.

When combining surveys that have been stratified or clustered care must be taken to ensure that the clusters and strata are correctly identified in each of the surveys. If the same clusters and/or strata have been used in the different surveys it is important to make sure that, for each stratum and/or cluster, the same numerical code is used in each dataset. If the strata and/or clusters differ between the surveys then the codes used to identify these should be unique across the datasets.

The process for ensuring that the same codes are used in all datasets is the same as that set out in Step 4.

The process for ensuring that unique codes are used across datasets which do not have the same strata and/or clusters is similar to that given in Step 2. Unique codes can be built up from a combination of survey identifier (target population , year of survey etc), and the allocated cluster or stratum number. For the purposes of analysis it does not matter what values these codes take. The egen command with the group option can be very useful here (see help egen).

When combining data that is clustered and/or stratified with SRS data, it is essential to assign a unique code to the SRS data in the cluster and/or strata variables. The value of the cluster and/or strata variable should be the same for each observation from the SRS data.

If any of the survey data contains a sample weight this must be included when datasets are combined. If weighted data are combined with data that does not need weighting, assign a value of 1 to the weight variable in each of the records from the unweighted survey.

Sample weights may need to be recalculated for the analysis of data from different surveys. This depends heavily on the source of the data and the type of analysis and is beyond the scope of this document. See for example Korm and Graubard (1999) Analysis of Health Surveys. New York: Wiley. ISBN: 0471137731.

Example

You want to combine two household surveys of young people from 1999 and 2002. Both datasets are stratified by urban or rural residence and both samples are clustered. There are no sample weights in either dataset. The strata are the same in both surveys but the clusters are different.

In both datasets there are three relevant variables called: year strata cluster. The strata variable takes two values: 1 for urban and 2 for rural, and this is the same in both datasets. In both surveys ten clusters were chosen from each stratum. A summary of the data looks like this:

1999

	strata	
cluster	1	2
1	6	6
2	7	10
3	3	4
4	4	5
5	7	3
6	6	4
7	4	3
8	6	7
9	3	3
10	4	5

2002

	strata	
cluster	1	2
1	3	6
2	3	5
3	5	8
4	6	7
5	6	4
6	3	2
7	5	3
8	9	3
9	4	8
10	6	4

What must you do to these data to be able to work with a combined dataset?

Possible solution

You should notice two potential problems. The clusters have the same numbers in both surveys but you know that these codes do not refer to the same places. It is also a little confusing that the same codes have been used for the clusters in each stratum. You can solve this problem with one command:

```
use 1999_hh_recoded ,clear

egen global_cluster=group(year strata cluster)

gen cluster1999=cluster
```

```
use 2002_hh_recoded ,clear

egen global_cluster=group(year strata cluster)

gen cluster2002= cluster
```

The new variable called global_cluster now contains 40 different codes- one for each unique combination of year, strata and cluster.

The only drawback with this approach is that the codes for the new variable (global_cluster) do not have any geographical meaning- so to find out where a particular cluster is located you need the original information. Here, provision is made to retain this information by creating the new variables which contain the original codes. These are put into two separate variables so there is no chance of using the original codes by mistake.

Stata commands, Quick reference[2]

*FILE COMMANDS

clear

set memory 250m

cd c: \directorypath

use filename

save newfilename, **replace**

log using logname.log, **replace**

log close

append using secondfilename.dta

merge sortedcommonvariable **using** mergingdataset.dta

erase unneededfile.dta

/* ignore any commands in the do file from here until resume

***/** resume executing do files command

/// at end of line - resume this commandfd line on the next line (for long command lines)

*DATA MANAGEMENT COMMANDS

generate newvariable **= .**

rename oldname newname

[2] This list is provided as a convenient quick reference to the correct syntax of commands most commonly used in preparing behavioural surveillance data for analysis. It is by no means comprehensive, and represents only a tiny fraction of the programmes capabilities. Users are very strongly urged to use the Help functions in Sata to improve their skills and increase the efficiency of data processing.

```
recode variablename 1/5 = 0 6 7 = 1 8=.

replace variablename = 2 if (othervariablename = = 0 | thirdvariable ~ = 1)

lab def labelname 1 “good” 2 “bad” 3 “indifferent”

lab val variablename labelname

lab drop wronglabelname

lab var “description of variable, don’t forget to include denominator and TR”

notes variablename: “you can attach notes to variables in the data set to describe changes in the definition over time, give details for composite indicators and give other information that users of the data set may need”

descr variable, fullnames alpha

codebook, problems

labelbook variable

ds, has (varlabel *variablestring*) varwidth(25)

list variablename othervariablename othervariablename in 230/250

table rowvariable columvariable supercolumnvariable, row col

centile numcsmonth if numcsmonth<97, centile(80)

replace numcsmonth=r(c_1) if numcsmonth>=97 & numcsmonth~=.

keep usefulvar1 usefulvar2 otheruseful*

drop uselessvar1/uselessvar5 otheruseless*

destring stringvariable_that_ought_to_be_numerical

order variable6 variable10 variable23 variable1
```

*DATA ANALYSIS COMMANDS

```
*COMPARING PROPORTIONS: DO THEY DIFFER BY SOME EXPLANATORY VARIABLE? (FOR CATEGORICAL VARIABLES)

tab explanatoryvariable outcomevariable, row chi2

tab occupation incomegroup, row chi2

tab explanatoryvariable outcomevariable if conditionvariable = = 1, row chi2

tab occupation incomegroup if sex == 1, row chi2

bysort stratavariable: tab explanatoryvariable outcomevariable, row chi2

bysort city: tab occupation incomegroup, row chi2

*(IF EXPLANATORY VARIABLE HAS MANY VALUES)
```

tab outcomevariable explanatoryvariable, **col chi2**

tab condomlastcs religion, col chi2

*TEST FOR STATISTICALLY SIGNIFICANCE TREND OVER ORDERED CATEGORICALS

nptrend outcomevariable, by (orderedexplanatory)

nptrend anyinjectingrisk if target ==51, by (education)

*****COMPARING DISTRIBUTIONS AND OTHER DESCRIPTIVE STATISTICS: DO THEY DIFFER BY SOME EXPLANATORY VARIABLE? (FOR NUMERIC VARIABLES)

tabstat outcomevariable, **stats** (mean sem q n)

tabstat firstsell, stats (mean sem q n)

tabstat outcomevariable, **by** (explanatoryvariable) **stats** (min mean max n)

tabstat firstsell, by (target) stats (mean sem q n)

tabstat outcomevariable **if** (conditionvariable >1 & conditionvariable ~= .), by (explanatoryvariable) **stats** (mean sem q n)

tabstat age, by (IDUyear) if sellsex == 1 & round == 2, stats (mean sem q n)

*****COMPARING MEANS: IS THERE A SIGNIFICANT DIFFERENCE BETWEEN CATEGORIES BY SOME EXPLANATORY VARIABLE? (FOR NUMERIC VARIABLES)

ci outcomevariable

ci totalpartners

by explanatoryvariable: **ci** outcomevariable **if** conditionvariable <4

sort sex

by sex: ci freqyesterday frequsual if city ==2

by explanatoryvariable: **ci** outcomevariable**, level** (90) total

sort target1

by target1: ci totalclients, level (90) total

Appendix 3 —Example codebook of variables for use in combined data sets

A NOTE ON NAMING AND LABELLING CONVENTIONS

Most variables in this codebook follow a standard pattern and a few simple conventions. Once these have been grasped, the variables become much easier to use.

Naming conventions

Any variable that ends in "**all**" has all respondents in the denominator, except for respondents belonging to sub-populations for whom the variable is not relevant. Those beginning with "**any**" also include all respondents in the denominator.

Any variable that ends in "**sa**" has only sexually active respondents in the denominator

Variables that include "**ever**" refer to behaviours over the course of a lifetime

Variables that end in "**year**" refer to behaviours over the course of the 12 months preceding the interview date.

Variables that begin with "**num**" refer to the total number of reported partners of a specifed type

Variables that begin with "**freq**" refer to the frequency of a reported behaviour, including sex with a specific partner type, and exposure to interventions

Variables that begin with "**total**" refer to the total number of partners, computed from the reported number of partners of various types.

"**s**" at the end of a knowledge variable denotes that the response was given spontaneously, rather than in response to a prompted question

Partner types

"**spouse**" refers to sex with a marital or live-in parnter of the opposite sex

"**cs**" (for commercial sex) refers to sex with a commercial partner, where a woman is selling and a man is buying

"**sf**" (for special friend) refers to sex between non-marital, non-commercial partners of opposite sexes. In other words, for a woman it refers to a boyfriend or casual male partner, while for a man it refers to a girlfriend or casual female partner

Other partner types are self-explanatory, and include "**sellsexman**" (for transgenders and men selling sex to men), "**buysexman**" (for transgenders and men buying sex from men) "**sexmannopay"** (for men and transgenders having sex with other men where payment is not involved in either direction) etc

Condom variables

Condom use variables follow the same pattern for all partnertypes, generally ending with the partnertype itself. The denominator is always all those who reported that partner type. For condom variables in male-male sex, the denominator refers to anal sex only, unless oral sex is specified in the variable name.

"condomlastpartnertype" refers to the use of a condom in the most recent act of sex with the specified partner type

"consiscondompartnertype" refers to **consistency** of condom use in all acts of sex with all individuals of the specified partner type over the designated time reference period

"alwayscondompartnertype" refers to condom use in **all** acts of sex with all individuals of the specified partner type over the designated time reference period

"nevercondompartnertype" refers to condom use in **none** of the acts of sex with any individuals of the specified partner type over the designated period

Value coding conventions

For binary yes/no variables:

0 = No, 1 = Yes

For variables measuring consistency of a behaviour:

1 = Never, 2 = Occasionally, 3 = Often 4 = Always

8, 88, 888 = Don't know or don't remember

9, 99, 999 = No response

Note that "Don't know" responses are generally only included for variables derived directly from the questionnaires. Variables recoded from several data sources within the questionnaire would not generally include this code.

The example code book contains details for many variables of the denominator, the numerator, and the "zero-value". The denominator is the total number of respondents included in the variable. In yes/no variables, the type most commonly found in behavioural data sets, the numerator is made up of respondents coded as "1". The zero value is made up of respondents coded as "0". Added together, these make up the denominator for indicators based on these variables.

Labelling conventions

In this codebook, we use the letters "TR" to denote time reference. The greatest variation in behavioural surveillance data sets is probably found in the area of time references, and so these are not specified here unless they are genuinely agreed standards (such as sex within the last year to denote currently sexually active). The use TR in labels is meant to remind users that although there is flexibility about time references, they should always be specified in variable labels. Note that time references can be different for different populations **within the same variable**. It is important to note these differences in the variable labels or, at a minimum, in variable notes.

Organisation of codebook

The variables in the codebook should be in the same order as the data set. These in turn are most convenient if they are arranged in ways that are easy for programme analysis, for example with all the HIV knowledge variables together, all the STI/RTI variables together etc. In this example codebook, the variables most frequently used in analysis are at the beginning of each section, with less frequently used variables further down the section. The order of the variables does not necessarily reflect the order either of the questionnaires or of the recoding process.

SURVEY VARIABLES

year

Variable label: Survey year (notes)

Notes: One round of BSS can span calendar years. if you want to compare rounds, use the variable round

---

round

Variable label: Surveillance round (notes)

Notes: This variable exists for cases when surveillance rounds span more than one calendar year

---

trend

Variable label: Data for this population and site is comparable between surveillance rounds

Notes: This variable allows analysts to restrict analysis to sites or groups of sites that are comparable between rounds

Denominator: all respondents

Numerator: Respondents in sites and sub-populations that have been included in all rounds of surveillance

Zero value: Respondents in sites or sub-populations that have only been included in some rounds of surveillance, not in all rounds

---

DEMOGRAPHIC VARIABLES

sex

Variable label: Respondent’s biological sex

1 male

2 female

---

province

Variable label: Province of interview

(Values locally specific, if possible use Census standards)

---

district

Variable label: District of interview

(Values locally specific, if possible use Census standards)

---

target

Variable label: Target group (high risk men combined)(notes)

10	all FSW
11	direct FSW
12	indirect FSW
20	high risk men
41	transgender sex workers
42	male sex workers
43	MSM
51	IDU

Notes: Direct FSW sell sex in brothels or on the streets, indirect FSW have other occupations such as masseuses, bar hostess. high risk men are men in occupations where there is a culture that supports the buying of sex, including sailors, truckers and port workers

---

age

---

agegroup

Variable label: Age group

1	<20
2	20-24
3	25-34
4	35+

---

school

Variable label: Highest level of education, not necessarily completed

1	no school
2	primary
3	middle school
4	high school
5	tertiary
9	no response

---

educ

Variable label: Highest completed level of education

1	has not completed primary
2	completed primary
3	completed middle school
4	secondary graduate or higher

marital

Variable label: Marital status

1	single
2	married
3	divorced
4	widowed
9	no response

married

Variable label: Currently married

0	no
1	yes
8	don’t know
9	no response

single

Variable label: Currently unmarried

0	no
1	yes
8	don’t know
9	no response

provorigin

Variable label: Province of origin

(Values locally specific, if possible use Census standards)

districtorigin

Variable label: District of origin

(Values locally specific, if possible use Census standards)

migrate

Variable label: Working outside province of origin

0 no

1 yes

8 don't know

9 no response

SEXUAL BEHAVIOUR VARIABLES, NON COMMERCIAL

eversex

Variable label: Has ever had sex

0 no

1 yes

9 no response

---

firstsex

Variable label: Age at first sex

Denominator: Those who have ever had sex

---

sexyear

Variable label: Has had sex in the last year

0 no

1 yes

8 don't know

9 no response

Denominator: All respondents

---

numsexpartnerall

Variable label: Total number of sex partners in TR

Denominator: All respondents

---

multipleall

Variable label: Has had more than one sex partner in TR (of all respondents)

0 no

1 yes

9 no response

Denominator: All respondents

Numerator: Respondents who have had sex with more than one partner in the TR preceding the survey

Zero Value: Respondents who have never had sex, have not had sex in the last year, or have sex with only one partner in the survey TR (usually one year)

---

extramaritalall

Variable label: Has had sex with a non-martial/non-live-in partner in the last TR (of all)

0 no

1 yes

9 no response

Denominator: All respondents

Numerator: Respondents who have had sex with a person to whom they are not married or with whom they are not living in the TR period (usually 12 months) preceding the survey (includes premarital sex)

Zero Value: Respondents who have never had sex, have not had sex in the last year, or who have had sex in the TR only with a person to whom they are married or with whom they are living

---

A

Variable label: Has not had sex in the last year

0 no

1 yes

8 don't know

9 no response

Denominator: All respondents (often coded only for high risk men)

Numerator: Respondents who have never had sex or have not had sex in the 12 months preceding the survey

Zero Value: Respondents who have had sex with any partner type in the last 12 months

---

B

Variable label: Only has one marital/live-in sex partner

0 no

1 yes

9 no response

Denominator: All respondents (often coded only for high risk men)

Numerator: Respondents who have had sex with a person to whom they are married or with whom they are living in the last TR (usually 12 months), and who have had no other sex partners in that time

Zero value: Respondents who have not had sex in the last TR (usually 12 months), and also respondents who have had sex with any person to whom they are not married or with whom they are not living in the TR preceding the survey

---

C

Variable label: Uses condoms with all non-marital partners

0 no

1 yes

8 don't know

9 no response

Denominator: All respondents (often coded only for high risk men)

Numerator: Respondents who have had sex with a person to whom they are not married or with whom they are not living in the last TR (usually 12 months), and always used a condom with all of their non-marital partners in that time.

Zero value: Respondents who have not had sex in the last TR, respondents who have not had sex with any person to whom they are not married or with whom they are not living in the TR preceding the survey, and also respondents who have had sex with non-marital partner(s) but did not always use a condom in all sex with non-marital partner(s) in the preceding TR (usually 12 months)

ABC

1 no sex in past year

2 sex with only one non-commercial partner

3 condoms in all extramarital sex

4 risky behaviour

Denominator: All respondents (often coded only for high risk men)

---

condomlastspouse

Variable label: Use condom at last sex with spouse or live-in partner

0 no

1 yes

8 don't know

9 no response

Denominator: All respondents who are currently married or living with partner of the opposite sex

Numerator: Respondents who are currently married and used condom in most recent sex with spouse or live-in partner

Zero value: Respondents who are currently married and did not use condom in most recent sex with spouse or live-in partner

---

consiscondomspouse

Variable label: Frequency of condom use with spouse or live-in partner in the last TR

1 never

2 occasionally

3 often

4 always

Denominator: All respondents who are currently married or living with partner of the opposite sex

---

alwayscondomspouse

Variable label: Always used condom with spouse or live-in partner in the last TR

0 no

1 yes

9 no response

Denominator: All respondents who are currently married or living with partner of the opposite sex

Numerator: Respondents who are currently married and used condom in all sex with spouse or live-in partner in the last TR

Zero value: Respondents who are currently married and did not use condom in all sex with spouse or live-in partner in the last TR

---

nevercondomspouse

Variable label: Never used condom with spouse or live-in partner in the last TR

0 no

1 yes

9 no response

Denominator: All respondents who are currently married or living with partner of the opposite sex

Numerator: Respondents who are currently married and never used a condom in any sex with spouse or live-in partner in the last TR

Zero value: Respondents who are currently married and ever used a condom in any sex with spouse or live-in partner in the last TR

---

sfall

Variable label: Had male to female sex with non-marital, non-live-in, non commercial partner. in the last TR, of all

0 no

1 yes

9 no response

Denominator: All respondents

---

condomlastsf

Variable label: Used a condom in last male to female sex with a non-marital, non-live-in, non-commercial partner

0 no

1 yes

8 don't know

9 no response

Denominator: Respondents who have had sex with a non-marital, non-live-in, non-commercial partner of the opposite sex in the stated TR period

Numerator: Respondents who have had sex a non-marital, non-live-in, non-commercial partner of the opposite sex in the TR and used condom in most recent sex with this type of partner

Zero value: Respondents who have had sex with a non-marital, non-live-in, non-commercial partner of the opposite sex in the TR and did not use a condom in most recent sex with this type of partner

---

alwayscondomsf

Variable label: Always used condoms in male to female sex with a non-marital, non-live-in, non-commercial partner in the last TR

0 no

1 yes

9 no response

Denominator: Respondents who have had sex with a non-marital, non-live-in, non-commercial partner of the opposite sex in the stated TR period

Numerator: Respondents who have had sex a non-marital, non-live-in, non-commercial partner of the opposite sex in the TR and always used a condom in all sex with this type of partner over the TR

Zero value: Respondents who have had sex a non-marital, non-live-in, non-commercial partner of the opposite sex in the TR and did not always use a condom in all sex with this type of partner over the TR

---

nevercondomsf

Variable label: Never used condoms in male to female sex with a non-marital, non-live-in, non-commercial partner in the last TR

0 no

1 yes

9 no response

Denominator: Respondents who have had sex with a non-marital, non-live-in, non-commercial partner of the opposite sex in the stated TR period

Numerator: Respondents who have had sex a non-marital, non-live-in, non-commercial partner of the opposite sex in the TR and never used a condom in any sex with this type of partner over the TR

Zero value: Respondents who have had sex a non-marital, non-live-in, non-commercial partner of the opposite sex in the TR and ever use a condom in any sex with this type of partner over the TR

---

sfsa

Variable label: Had male to female sex with non-marital, non-live-in, non commercial partner. in the last TR of those with sex last year

0	no
1	yes
8	don't know
9	no response

Denominator: Respondents who have had sex in the previous 12 months

---

numsfall

Variable label: # Non-marital, non commercial male to female sex partners in TR

Denominator: All respondents (including those with no sex, no sex in the last year, or no sex with non-marital, non-commercial partners)

---

consiscondomsf

Variable label: Frequency of condom use in male to female sex with a non-marital, non-live-in, non-commercial partner in the last TR

1	never
2	occasionally
3	often
4	always

Denominator: Respondents who have had sex with a non-marital, non-live-in, non-commercial partner of the opposite sex in the stated TR period

---

whynocondomsf

Variable label: Why was no condom was used with boyfriend/girlfriend?

1	none available
2	man doesn't want to
3	partner is clean
4	in love
5	have already taken medicine
6	other

Denominator: Respondents who have had sex with a non-marital, non-live-in, non-commercial partner of the opposite sex in the stated TR period, and did not always use condoms with those partners in that time

---

multiplesa

Variable label: Has had more than one sex partner in TR (of sexually active)

0 no

1 yes

9 no response

Denominator: Respondents who have had sex in the last year

Numerator: Respondents who have had sex with more than one partner in the TR preceding the survey

Zero Value: Respondents who have had sex with only one partner in the survey TR (usually one year)

---

extramaritalsa

Variable label: Has had sex with a non-martial/non-live-in partner in TR (of sexually active)

0 no

1 yes

9 no response

Denominator: Respondents who have had sex in the last year

Numerator: Respondents who have had sex with a person to whom they are not married or with whom they are not living in the TR period (usually 12 months) preceding the survey (includes premarital sex)

Zero Value: Respondents who have had sex in the TR only with a person to whom they are married or with whom they are living

---

MALE TO FEMALE COMMERCIAL SEX VARIABLES

csall

Variable label: Commercial sex between male and female in last TR

0 no

1 yes

8 don’t know

9 no response

Denominator: All respondents

Numerator: Male respondents who have bought sex from a woman in the questionnaire TR, and all women who have sold sex to a man in the TR

Zero value: Respondents who have never had sex or have not had sex in the last year, men who have not bought sex from a woman in the male TR period, and women who have not sold sex (not had any male clients) in the femaleTR.

---

anypayingsexall

Variable label: Had any commercial sex in last month/in the last year (all respondents) (notes)

0 no

1 yes

9 no response

Notes: This includes people who paid girlfriends in cash after having sex with them last time, if this question is asked.

Denominator: All respondents

Numerator: Male respondents who have bought sex from a woman in the questionnaire TR, or who have sex with a "girlfriend" or casual partner and then paid her in cash, and all women who have sold sex to a man in the TR, including those who had sex with a "boyfriend" who paid them cash after sex

Zero value: Respondents who have never had sex or have not had sex in the last year, men who have not paid cash to a woman after sex in the male TR period, and women who have not sold sex (not had any male clients or been paid cash after sex) in the female TR.

---

anyunprotectedcs

Variable label : Any unprotected commercial sex between male and female last TR, of all

0 no

1 yes

9 no response

Denominator: All respondents

Numerator: Male respondents who have bought sex from a woman in the questionnaire TR, or who have sex with a "girlfriend" or casual partner and then paid her in cash, and who did not **always** use condoms with those partners, and all women who have sold sex to a man in the TR, including those who had sex with a "boyfriend" who paid them cash after sex, and who did not **always** use condoms with those partners.

Zero value: Respondents who have never had sex or have not had sex in the last year, men who have not paid cash to a woman after sex in the male TR period, or who have paid cash but always used condoms with all paid partners, and women who have not sold sex (not had any male clients or been paid cash after sex) in the female TR, or have sold sex but have always used condoms with all paying clients in the TR

---

condomlastcs

Variable label: Used a condom at last commercial sex

0	no
1	yes
8	don't know
9	no response

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR

Numerator: Men who bought sex from a woman in male TR and used a condom last time they bought sex or women who sold sex to a man in female TR and used a condom last time they sold sex

Zero value: Men who bought sex from a woman in male TR and did not use a condom last time they bought sex or women who have sold sex to a man in female TR and did not use a condom last time they sold sex

alwayscondomcs

Variable label: Always used condoms in all commerical sex in TR

0	no
1	yes
9	no response

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR

Numerator: Men who bought sex from a woman in male TR and always used a condom when they bought sex or women who sold sex to a man in female TR and used a condom every time they sold sex

Zero value: Men who bought sex from a woman in male TR and did not always use a condom when they bought sex or women who have sold sex to a man in female TR and did not always use a condom every time they sold sex

nevercondomcs

Variable label: Never used condoms in commercial sex in TR

0	no
1	yes
9	no response

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR

Numerator: Men who bought sex from a woman in male TR and never used a condom when they bought sex, or women who sold sex to a man in female TR and never used a condom any time they sold sex

Zero value: Men who bought sex from a woman in male TR and ever used a condom when they bought sex or women who sold sex to a man in female TR and ever used a condom any time they sold sex

consiscondomcs

Variable label: Frequency of condom use in commercial sex btwn men and women

1 never

2 occasionally

3 often

4 always

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR

---

whynocondomcs

Variable label: Why was no condom was used in commercial sex (notes)

1 none available

2 man doesn't want to

3 partner is clean

4 in love

5 have already taken medicine

6 other

Example of notes: Round 1 this was asked only of people who didn't use condoms at last sex, round 2 of anyone who didn't use in all commercial sex in last week (FSW), last month (men)

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR, and did not always use condoms with commercial partners in that time (see notes)

---

price

Variable label: Money recieved (by sex workers)/paid (by clients) at last comm sex, (currency) (notes)

Notes: "Can't remember coded as missing"

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR, and who give a price for the most recent transaction

---

pricegroup

Variable label: Money recieved (by sex workers)/paid (by clients) at last comm sex, (currency) (notes)

Values vary locally according to currency and price structure

Notes: "Can't remember coded as missing"

Denominator: Men who have bought sex from a woman in male TR or women who have sold sex to a man in female TR, and who give a price for the most recent transaction

---

foreignclient

Variable label: Last client was foreigner

0	no
1	yes
8	don’t know
9	no response

Denominator: Women who have sold sex to a man in TR

Numerator: Women who have sold sex to a man in TR and say their last client was a foreigner

Zero value: Women who have sold sex to a man in TR and say their last client was not a foreigner

---

localclient

Variable label: Last client was a local resident, not migrant or foreigner

0	no
1	yes
8	don’t know
9	no response

Denominator: Women who have sold sex to a man in TR

Numerator: Women who have sold sex to a man in TR and say their last client was a local resident, not a migrant e.g. from another province, or a foreigner

Zero value: Women who have sold sex to a man in TR and say their last client was a foreigner or a migrant from another part of the country

---

SEX INDUSTRY VARIABLES: FEMALE, TRANSGENDER AND MALE SEX WORKERS

firstsell

Variable label: Age at first selling sex

Denominator: Sex working populations

---

timesell

Variable label: Length of time selling sex (months)

Denominator: Sex working populations

---

timesellgroup

Variable label: Length of time selling sex (years)

1	<=1
2	1-2
3	2-4
4	4+

Denominator: Sex working populations

---

sellsexelsewhere

Variable label: Have sold sex in another city

0	no
1	yes
9	no response

Denominator: Sex working populations

---

timehere

Variable label: Length of time selling sex here (months)

Denominator: Sex working populations

---

timeheregroup

Variable label: Length of time selling sex here

1	< 6 months
2	6 - 11 months
3	12 - 23 months
4	24 + months

Denominator: Sex working populations

---

sexworker

Variable label: Is the respondent a sex worker?

0	no
1	yes

Denominator: All respondents

Numerator: Respondents who sell sex for a living

Zero value: Respondents who do not regularly receive cash in exchange for sex

---

location

Variable label: Works (was interviewed) at what type of location

1	brothel
2	street/park
3	hotel

4	massage parlour/salon
5	karaoke/disco/bar
6	port
7	truck/taxi stop
8	factory
9	barracks
10	other

Denominator: All respondents

---

establishment

Variable label: Establishment based sex worker

0	no
1	yes

Denominator: All sex worker respondents

Numerator: Sex workers who sell sex out of an establishment such as a brothel or bar

Zero value: Sex workers who sell sex on the streets or in other public spaces

---

street

Variable label: Street based sex worker

0	no
1	yes

Denominator: All sex worker respondents

Numerator: Sex workers who sell sex on the streets or in other public spaces such as parks

Zero value: Sex workers who sell sex out of an establishment such as a brothel or bar

---

clientwho

Variable label: Most common ocuupation of clients in last week

1	student
2	police/military
3	civil servant
4	private sector
5	labourer
6	unemployed
7	trader
8	other
98	don’t know
99	no response

Denominator: All sex worker respondents

---

violence

Variable label: Has been raped/forced to have sex without payment in TR

0 no

1 yes

9 no response

Denominator: All sex worker respondents

Numerator: Sex workers who say they have been raped or forced to have sex without payment in the TR (usually preceding 12 months)

Zero value: Sex workers who say they have not been raped or forced to have sex without payment in the TR (usually preceding 12 months)

---

sexotherprovince

Variable label: Has sold (sex worker)/bought (client) sex in another province

0 no

1 yes

9 no response

Denominator: Men who have bought sex from a woman in male TR or sex workers

Numerator: Men who bought sex from a woman in male TR in province other than province of interview, and sex workers and who have ever sold sex in province other than province of interview

Zero value: Men who bought sex from a woman in male TR but did not buy sex in province other than province of interview, and sex workers who have never sold sex in any province other than province of interview

---

COMMERCIAL SEX BETWEEN MALES

MEN AND TRANSGENDER SEX WORKERS SELLING SEX TO MEN

sellsexmanall

Variable label: Sold sex to a man last TR, of all male/transgender respondents

0 no

1 yes

9 no response

Denominator: All male respondents (note this is usually not asked of “high risk heterosexual male” groups, but is often included in IDU questionnaires)

Numerator: Male/transgender respondents who have bought sex from male sex worker in TR

Zero value: Male/transgender respondents who have never had sex, have not had sex in the last year, or have not bought sex from male sex worker in TR

---

anyunprotectedmaleclient

Variable label: Had unprotected anal sex with male client in TR, of all male/transgender sex working respondents

0 no

1 yes

Denominator: Male/transgender sex working populations (note in some countries this variable can also be constructed for male IDU and female sex workers)

Numerator: Male/transgender who have sold anal sex to a man in TR and did not always use condom in all anal sex with male clients during TR

Zero value: Male/transgender who did not have anal sex clients in TR or who sold anal sex to a man in TR but always used condom in all anal sex with male clients during TR

---

numanalsellsexmanall

Variable label: Number of anal male clients in TR, of male sex working groups (notes)

Notes: "Don't remember" coded to 80th percentile"

Denominator: All male/transgender sex worker respondents

---

numsellsexmanall

Variable label: Number of male clients in TR, of male sex working groups (notes)

Notes: "Don't remember" coded to 80th percentile, includes anal, oral and non-penetra tive clients"

Denominator: All male/transgender sex worker respondents

---

condomlastsellsexman

Variable label: Used a condom when last selling anal sex to male client in TR

0 no

1 yes

8 don't know

9 no response

Denominator: Male/transgender who have sold anal sex to a man in TR

Numerator: Male/transgender who have sold anal sex to a man in TR and used a condom last time they sold anal sex

Zero value: Male/transgender who have sold anal sex to a man in TR and did not use a condom last time they sold anal sex

---

alwayscondomsellsexman

Variable label: Always used a condom when selling anal sex to all male clients in TR

0 no

1 yes

9 no response

Denominator: Male/transgender who have sold anal sex to a man in TR

Numerator: Male/transgender who have sold anal sex to a man in TR and always used a condom every time they sold anal sex in TR

Zero value: Male/transgender who have sold anal sex to a man in TR and did not always use a condom every time they sold anal sex in TR

---

nevercondomsellsexman

Variable label: Never used condoms when selling anal sex to any male client in TR

0 no

1 yes

9 no response

Denominator: Male/transgender who have sold anal sex to a man in TR

Numerator: Male/transgender who have sold anal sex to a man in TR and never used a condom any time they sold anal sex in TR

Zero value: Male/transgender who have sold anal sex to a man in TR and ever used a condom any time they sold anal sex in TR

---

consiscondomsellsexman

Variable label: Frequency of condom use among male/transgender sex workers in anal sex with male clients in TR

1 never

2 occasionally

3 often

4 always

Denominator: Male/transgender who have sold anal sex to a man in TR

MEN/TRANSGENDERS BUYING SEX FROM MEN

buysexmanall

Variable label: Bought sex from a man last TR, of all male/transgender respondents

0 no

1 yes

9 no response

Denominator: All male respondents

Numerator: Male/transgender respondents who have bought sex from male sex worker in TR

Zero value: Male/transgender respondents who have never had sex, have not had sex in the last year, or have not bought sex from male sex worker in TR

---

anyunprotectebuysexman

Variable label: Paid for unprotected anal sex with male sex worker (not transgender) in TR, of all male/transgender respondents (notes)

0 no

1 yes

Notes: Refers to male sex wokers only. For all anal commercial sex from buying side (including buying from transgender), use anyunprotectedbuysexmale

Denominator: Male/transgender populations (note in some countries this variable can only be constructed for MSM populations)

Numerator: Male/transgender who have bought anal sex from a man in TR and did not always use condom in all anal sex with male sex workers during TR

Zero value: Male/transgender who did not have anal sex in TR, or did not pay a man for anal sex in TR, or bought anal sex from a man but always used condom in all anal sex with male sex workers during TR

---

numanalbuysexmanall

Variable label: Number of male sex workers from whom bought anal sex in TR (notes)

Notes: “Don’t remember” coded to 80th percentile. Includes only male-identified part ners, not transgender sex workers”

Denominator: All male/transgender respondents (note this is usually not asked of “high risk heterosexual male” groups, but is usually included in MSM questionnaires)

---

numbuysexmanall

Variable label: Number of male sex workers from whom bought sex in TR (notes)

Notes: “Don’t remember” coded to 80th percentile. Includes only male-identified part ners, not transgender sex workers. Includes oral and non-penetrative sex”

Denominator: All male/transgender respondents (note this is usually not asked of “high risk heterosexual male” groups, but is usually included in MSM questionnaires)

---

condomlastbuysexman

Variable label: Used a condom when last buying anal sex from male sexworker in TR

0 no

1 yes

8 don’t know

9 no response

Denominator: Male/transgender who have bought anal sex from a man (not transgender) in TR

Numerator: Male/transgender who have bought anal sex from a man (not transgender) in TR and used a condom last time they bought anal sex from this type of partner

Zero value: Male/transgender who have bought anal sex from a man (not transgender) in TR and did not use a condom last time they bought anal sex from this type of partner

---

alwayscondombuysexman

Variable label: Always used a condom when buying anal sex from all male sexworkers in TR

0 no

1 yes

9 no response

Denominator: Male/transgender who have bought anal sex from a man (not transgender) in TR

Numerator: Male/transgender who have bought anal sex from a man (not transgender) in TR and always used a condom every time they bought anal sex in TR from this type of partner

Zero value: Male/transgender who have bought anal sex from a man (not transgender) in TR and did not always use a condom every time they bought anal sex in TR from this type of partner

---

nevercondombuysexman

Variable label: Never used condoms when buying anal sex from any male sexworker in TR

0 no

1 yes

9 no response

Denominator: Male/transgender who have bought anal sex from a man (not transgender) in TR

Numerator: Male/transgender who have bought anal sex from a man (not transgender) in TR and never used a condom any time they bought anal sex in TR from this type of partner

Zero value: Male/transgender who have bought anal sex from a man (not transgender) in TR and ever used a condom any time they bought anal sex in TR from this type of partner

---

consiscondombuysexman

Variable label: Frequency of condom use among male/transgender sex workers in anal sex with male sexworkers in TR

1 never

2 occasionally

3 often

4 always

Denominator: Male/transgender who have bought anal sex from a man (not transgender) in TR

---

anyunprotectebuysexmale

Variable label: Paid for unprotected anal sex with male/transgender sex worker in TR, of all male/transgender respondents (notes)

0 no

1 yes

Notes: Refers to male/transgender sex workers. For only male sex workers, see anyunprotectedbuysexman, for only transgender sex workers, see anyunprotectedbuysextransgender

Denominator: Male and transgenders populations (note in some countries this variable can only be constructed for MSM populations)

Numerator: Male/transgender who have bought anal sex from a man or a transgender sw in TR and did not always use condom in all anal sex with male/transgender sex workers during TR

Zero value: Male/transgender who did not have anal sex in TR, or who did not pay a man or transgender for anal sex in TR, or who bought anal sex from a man but always used condom in all anal sex with male/transgender sex workers during TR

MEN BUYING SEX FROM TRANSGENDER SEX WORKERS

buysextgall

Variable label: Bought sex from transgender sex worker last TR, of all male respondents

0 no

1 yes

9 no response

Denominator: All male respondents

Numerator: Men who have bought sex from transgender sex worker in TR

Zero value: Men who have never had sex, who have not had sex in the last year, or who have not bought sex from transgender sex worker in TR

---

anyunprotectebuysextg

Variable label: Paid for unprotected anal sex with transgender sex worker in TR, of all male respondents

0 no

1 yes

Denominator: All male respondents (note in some countries this variable can only be constructed for MSM populations)

Numerator: Men who have bought anal sex from a transgender sex worker in TR and did not always use condom in all anal sex with transgender sex workers during TR

Zero value: Men who did not have anal sex in TR, or who did not pay a man for anal sex in TR, or who bought anal sex from a transgender sex worker but always used condom in all anal sex with transgender sex workers during TR

---

numanalbuysextgall

Variable label: Number of transgender sex workers from whom bought anal sex in TR (notes)

Notes: "Don't remember" coded to 80th percentile."

Denominator: All male respondents (note this is usually not asked of "high risk heterosexual male" groups, but is usually included in MSM questionnaires)

---

numbuysextgall

Variable label: Number of tansgender sex workers from whom bought sex in TR (notes)

Notes: "Don't remember" coded to 80th percentile. Includes oral and non-penetrative sex"

Denominator: All male respondents (note this is usually not asked of "high risk heterosexual male" groups, but is usually included in MSM questionnaires)

---

condomlastbuysextg

Variable label: Used a condom when last buying anal sex from transgender sex worker in TR

0 no

1 yes

8 don't know

9 no response

Denominator: Men who have bought anal sex from transgender sex worker in TR

Numerator: Men who have bought anal sex from transgender sex worker in TR and used a condom last time they bought anal sex from this type of partner

Zero value: Menwho have bought anal sex from transgender sex worker in TR and did not use a condom last time they bought anal sex from this type of partner

---

alwayscondombuysextg

Variable label: Always used a condom when buying anal sex from all transgender sex workers in TR

0	no
1	yes
9	no response

Denominator: Men who have bought anal sex from transgender sw in TR

Numerator: Men who have bought anal sex from transgender sw in TR and always used a condom every time they bought anal sex in TR from this type of partner

Zero value: Men who have bought anal sex from transgender sw in TR and did not always use a condom every time they bought anal sex in TR from this type of partner

---

nevercondombuysextg

Variable label: Never used condoms when buying anal sex from any transgender sex worker in TR

0	no
1	yes
9	no response

Denominator: Men who have bought anal sex from transgender sex worker in TR

Numerator: Men who have bought anal sex from transgender sex worker in TR and never used a condom any time they bought anal sex in TR from this type of partner

Zero value: Men who have bought anal sex from transgender sex worker in TR and ever used a condom any time they bought anal sex in TR from this type of partner

---

consiscondombuysextg

Variable label: Frequency of condom use among male clients when buying sex from transgender sex workers in TR

1	never
2	occasionally
3	often
4	always

Denominator: Men who have bought anal sex from transgender sex worker in TR

SUMMARY OF ALL MALE TO MALE COMMERCIAL SEX RISK

anymalecs

Variable label: Male/transgender bought or sold sex to or from man or transgender sw in TR, of all male/transgender (notes)

0 no

1 yes

Notes: This refers to oral and non-penetrative sex as well as anal sex. For anal sex, use anyanalmalecs

Denominator: All male/transgender respondents

Numerator: Male/transgender respondents who paid a male/transgender sex worker for sex in TR, or was paid by male client for sex in TR

Zero value: Male/transgender respondents who did not have any sex with a male partner, or had sex with a male partner but did not give or receive money in exchange for sex in TR

---

anyanalmalecs

Variable label: Male/transgender bought or sold anal sex to or from man or transgender sex worker, of all male/transgender (notes)

0 no

1 yes

Notes: This refers only to anal sex. For commercial sex including oral and non-penetra tive sex, use anymalecs

Denominator: All male/transgender respondents

Numerator: Male/transgender respondents who paid a male/transgender sex worker for anal sex in TR, or was paid by male client for anal sex in TR

Zero value: Male/transgender respondents who did not have any anal sex with a male partner, or had anal sex with a male partner but did not give or receive moneyin exchange for anal sex in TR

---

anyunprotectedmalecs

Variable label: Paid or was paid for unprotected anal sex between males or males & transgenders, of all males/transgender

0 no

1 yes

Denominator: All male respondents (note in some countries this variable can only be constructed for MSM populations)

Numerator: Male respondens who have bought anal sex from a male/transgender sex worker in TR or sold anal sex to a male in TR and did not always use condom in all commercial anal sex during TR

Zero value: Males respondents who did not have anal sex in TR, or who did not exchange money anal sex in TR, or who exchanged money for anal sex with a male/transgender sex worker but always used condom in all commercial anal sex during TR

---

whynocondomcsman

Variable label: Why was no condom used in commercial anal sex between males or males and transgender sex worker?

1 none available

2 sex worker doesn't like condoms

3 client doesn't like condoms

4 expensive

5 other

Denominator: Male/transgender respondents who have sold anal sex to a man in TR and did not always use condoms with clients in that time, men/transgender who have bought anal sex from a man/transgender in TR and did not always use condoms with sex worker in TR.

MALE TO MALE SEX WITH NON-COMMERCIAL PARTNERS

sexmannopayall

Variable label: Had sex with man without payment in TR, of all male/transgender respondents (notes)

0 no

1 yes

9 no response

Denominator: All male respondents

Numerator: Male/transgender respondents who have had sex in TR with male with no payment

Zero value: Male/transgender respondents who have never had sex, have not had sex with a male/transgender partner in TR, or have not had any male/transgender sex partners in TR with whom no payment was involved

---

anyunprotectedsexmannopay

Variable label: Had unprotected anal sex with male/transgender with no payment in volved in TR, of all male respondents

0 no

1 yes

Denominator: All male respondents (note in many countries this variable can only be constructed for MSM groups)

Numerator: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR and did not always use condom in all anal sex with unpaid, nonpaying man during TR

Zero value: All male/transgender who did not have anal sex with a man, did not have non-paying, unpaid anal sex with a man in TR, or had non-paying, unpaid sex with a man and always used condoms in all anal sex with non-paying, unpaid man during TR

---

numanalsexmannopayall

Variable label: Number of men with whom respondent had non-paying, unpaid anal sex in TR, of all males/transgender (notes)

Notes: "Don't remember" coded to 80th percentile"

Denominator: All male/transgender respondents (in some countries this can only be calcuated for MSM populations)

---

numsexmannopayall

Variable label: Number of men with whom respondent had non-paying, unpaid sex in TR, of all males/transgender (notes)

Notes: "Don't remember" coded to 80th percentile, includes oral and non-penetrative sex partners"

Denominator: All male/transgender respondents (in some countries this can only be calcuated for MSM populations)

---

condomlastsexmannopay

Variable label: Used a condom at last anal sex between males or males & transgender without payment in TR

0	no
1	yes
8	don't know
9	no response

Denominator: Male/transgender respondents who have had sex with a non-paying, unpaid man in TR

Numerator: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR and used a condom at last anal sex with that partner type

Zero value: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR and did not use a condom at last anal sex with that partner type

---

alwayscondomsexmannopay

Variable label: Always used a condom in anal sex between males or males & transgender without payment in TR

0 no

1 yes

9 no response

Denominator: Male/transgender respondents who have had sex with a non-paying, unpaid man in TR

Numerator: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR and always used a condom in all anal sex with that partner type in TR

Zero value: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR but did not always used a condom in all anal sex with that partner type in TR

---

nevercondomsexmannopay

Variable label: Never used a condom in anal sex between males/ males & transgender without payment in TR

0 no

1 yes

9 no response

Denominator: Male/transgender respondents who have had sex with a non-paying, unpaid man in TR

Numerator: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR and never used a condom in any anal sex with that partner type in TR

Zero value: Male/transgender respondents who have had anal sex with a non-paying, unpaid man in TR and ever used a condom in any anal sex with that partner type in TR

---

consiscondomsexmannopay

Variable label: Frequency of condom use in anal sex between males/ males & transgender without payment in TR

1 never

2 occasionally

3 often

4 always

Denominator: Male/transgender respondents who have had sex with a non-paying, unpaid man in TR

SUMMARY OF ALL MALE TO MALE SEX RISK

msmall

Variable label: Had anal sex with other man (not transgender) in TR (of all non-transgender males)

0 no

1 yes

9 no response

Denominator: All male respondents

Numerator: Male respondents who have had anal sex with a man (not transgender) in TR (usually 12 months preceding survey)

Zero value: Male respondents who have never had sex, have not had sex in the last year, or have not had anal sex with a man in the last year.

---

tgall

Variable label: Had anal sex with transgender in TR (of all males)

0 no

1 yes

9 no response

Denominator: All male respondents

Numerator: Male respondents who have had anal sex with a transgender in TR (usually 12 months preceding survey)

Zero value: Male respondents who have never had sex, have not had sex in the last year, or have not had anal sex with a transgender partner in the last year.

---

anyanal

Variable label: Had anal sex with any male/transgender partner in TR, of all males/ transgender

0 no

1 yes

9 no response

Denominator: All male/transgender respondents

Numerator: Male/transgender respondents who have had anal sex with a man or transgender in TR (usually 12 months preceding survey)

Zero value: Male/transgender respondents who have never had sex, have not had sex in the last year, or have not had anal sex with a man or transgender partner in the last year.

---

anyunprotectedanal

Variable label: Had unprotected anal sex with any type of male/transgender partner, of all male/transgender respondents

0 no

1 yes

Denominator: All male respondents (note in many countries this variable can only be constructed for msm groups)

Numerator: Male/transgender respondents who have had anal sex with any male/transgender partner type in TR and did not always use a condom in all anal sex with every partner type in TR

Zero value: Male/transgender respondents who did not have anal sex with any male/transgender partner type in TR, or who had anal sex with any male or transgender partner type in TR but always used a condom in all anal sex with every partner type in TR

---

totalanalpartnersall

Variable label: Number of anal sex partners of any type in TR

Denominator: All male/transgender respondents

---

totalanalgroup

Variable label: Number of anal sex partners of any type in TR

0 none

1 one a week or fewer

2 up to three a week

3 more than three a week

Denominator: All male/transgender respondents

LUBRICANT USE AND OTHER RISK FACTORS IN ANAL SEX

waterbasedlubelast

Variable label: Used water-based lubricant at last anal sex, of those reporting anal sex in TR

0 no

1 yes

9 no response

Denominator: All male respondents reporting anal sex (usually only MSM groups)

Numerator: Male respondents who have had anal sex in TR and used water-based lubricant in most recent anal sex

Zero value: Male respondents who have had anal sex in TR but never used any lubricant, or never heard of water-based lubricant, or did not use water-based lubricant in last anal sex

---

waterbasedlubeever

Variable label: Has ever used water-based lubricant in anal sex, of those reporting anal sex in TR

0	no
1	yes
9	no response

Denominator: All male respondents reporting anal sex (usually only MSM groups)

Numerator: Male respondents who have had anal sex in TR and ever used water-based lubricant in anal sex

Zero value: Male respondents who have had anal sex in TR but never used water-based lubricant in anal sex

---

knowwaterbasedlube

Variable label: Has heard of water-based lubricant for use with condoms and in anal sex, of all MSM groups

0	no
1	yes
9	no response

Denominator: All respondents in MSM groups

Numerator: Male respondents who have heard of water-based lubricant for use with condoms and in anal sex

Zero value: Male respondents who have never heard of water-based lubricant for use with condoms and in anal sex

---

consislube

Variable label: Frequency of use of water-based lubricant in anal sex in TR, of those with anal sex

1	never
2	rarely
3	often
4	always

Denominator: All respondents in MSM groups who report anal sex in TR

---

whynolube

Variable label: Reason for not always using water-based lubricant lube in anal sex, of those not always using it in TR

1	expensive
2	shy to buy
3	don’t know where to get it
4	no need
5	use other non-water-based lubricant
6	other

---

whatlubeall

Variable label: Lubricant used at last anal sex, of all respondents with anal sex

0	none
1	saliva
2	oil/lotion
3	water-based lubricant
4	other

Denominator: All respondents in MSM groups who report anal sex in TR

---

anytop

Variable label: Reports often being penetrative partner in anal sex, of all with anal part ners

0	no
1	yes
9	no response

Denominator: All male respondents reporting anal sex (usually only MSM groups)

Numerator: Male respondents who have had anal sex in TR and reports often or lways being penetrative partner in anal sex

Zero value: Male respondents who have had anal sex in TR and reports rarely or never being penetrative partner in anal sex

---

anybottom

Variable label: Reports often being receptive partner in anal sex, of all with anal partners

0	no
1	yes
9	no response

Denominator: All male respondents reporting anal sex (usually only MSM groups)

Numerator: Male respondents who have had anal sex in TR and often or always is the receptive partner in anal sex

Zero value: Male respondents who have had anal sex in TR and rarely or never is the receptive partner in anal sex

---

analhow

Variable label: Of those with anal sex, position in anal sex is usually:

1	receptive partner
2	insertive partner
3	insertive and receptive
8	don’t remember

Denominator: All male respondents reporting anal sex (usually only MSM groups)

---

oralhow

Variable label : Of those with oral sex, position in oral sex is usually:

1 receptive partner

2 insertive partner

3 insertive and receptive

8 don’t remember

Denominator: All male respondents reporting oral sex (usually only MSM groups)

CONDOM PROGRAMMING VARIABLES

condomaccess

Variable label: Ease of access to condoms, as reported by survey staff (notes)

1 no condoms

2 condoms only outside establishment

3 condoms inside establishment

Notes: Before filling the questionnaire, survey staff check to see if condoms are avail able in or around the establishment and record the brands. For street based sex workers “inside” means at the drinks stall where they wait for clients

Denominator: All sex working respondents

---

knowcondom

Variable label: Recognises a condom when shown one by the interviewer

0 no

1 yes

Denominator: All sex working respondents (note, this is sometimes asked of MSM and high risk men also)

Numerator: Female, male/transgender sex workers who correctly identify a condom when shown one by the interviewer

Zero value: Female, male/transgender sex workers who are shown a condom but do not know what it is (do not identify it as a condom or slang equivalent)

---

showcondom

Denominator: All sex working respondents (note, this is sometimes asked of MSM and high risk men also)

Variable label: Is carrying a condom and can show it to the interviewer

0 no

1 yes

---

askcondom

Variable label: Frequency with which sex worker proposed condoms to clients in TR

1	never
2	occasionally
3	often
4	always

Denominator: All sex working respondents

---

condomad

Variable label: Has heard/seen an ad promoting condoms (notes)

0	no
1	yes
9	no response

Notes: "Don't know" coded to zero

Denominator: All respondents (note denominator may be restricted according to target audience for condom campaign, e.g. coded for high risk men only)

HIV KNOWLEDGE VARIABLES

knowaids

Variable label: Has heard of HIV/AIDS, of all (notes)

0	no
1	yes

Notes: "Don't know" and "no response" coded to zero

---

knowenough

Variable label: Knows enough information relevant to themselves to prevent HIV, of all (notes)

0	no
1	yes

Notes: This varies by group. Sex workers must know that condoms prevent HIV and that you cannot tell partners are infected by looking at them (male/transgender sex workers must specify condom use in anal sex). High risk men must know at least that A or B or C are protective (abstinence, monogamy or condom use). IDU must know that sharing needles spreads HIV. MSM must know they can prevent HIV by avoiding anal sex or using condoms in anal sex)

Denominator: All respondents

---

knowabc

Variable label: Knows that abstinence Monogamy and Condoms prevent HIV

0 no

1 yes

Denominator: All respondents in “high risk male” groups e.g. truckers, military etc, and IDU

---

knowabcd

Variable label: Knows that abstinence, Monogamy, Condoms and not sharing needles prevent HIV, of all

0 no

1 yes

Denominator: All respondents in “high risk male” groups e.g. truckers, military etc, and IDU

---

healthy

Variable label: Knows you cannot tell who is infected from their appearance, of all (notes)

0 no

1 yes

Notes: Don’t know, no response, never heard of AIDS, coded to zero

---

bloodtest

Variable label: Knows the only way to tell HIV is through blood test, of all (notes)

0 no

1 yes

Notes: Don’t know, no response, never heard of AIDS, coded to zero

---

canprevent

Variable label: Knows HIV can be prevented, of all (notes)

0 no

1 yes

Notes: Don’t know, no response coded to zero

---

*NOTE THAT FOR THE FOLLOWING QUESTIONS, THERE IS SOMETIMES AN OPTION FOR RESPONDENTS TO ANSWER SPONTANEOUSLY, THEN BE PROMPTED. BY CONVENTION, WE USE THE SUFFIX “S” FOR A SPONTANOUS RESPONSE, TO DISTIGUISH IT FROM A PROMPTED (INCLUDING SPONTANEOUS RESPONSE). TO SAVE SPACE, THIS IS ILLUSTRATED HERE ONLY IN THE FIRST EXAMPLE, ABSTAINS/ABSTAIN

abstains

Variable label: Knows abstinence can prevent HIV, spontaneous, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of AIDS, does not know AIDS can be prevented coded to zero

Denominator: All respondents

Numerator: Respondents asked about methods to prevent HIV/AIDS who spontaneously volunteer that it can be prevented by abstaining from sex

Zero value: Respondents who have never heard of HIV/AIDS, do not believe that it can be prevented, or do not spontaneously mention abstaining from sex when asked how HIV/AIDS can be prevented.

---

abstain

Variable label: Knows abstinence can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of AIDS, does not know AIDS can be prevented coded to zero

Denominator: All respondents

Numerator: Respondents (who spontaneously volunteer that HIV/AIDS can be prevented by abstaining from sex, or) who agree on prompting that HIV/AIDS can be prevented by abstaining from sex

Zero value: Respondents who have never heard of HIV/AIDS, do not believe that it can be prevented, or say no when asked if HIV/AIDS can be prevented by abstaining from sex

---

medicine

Variable label: Thinks taking medicine can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

condom

Variable label: Knows using condoms can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

inject

Variable label: Knows not sharing needles can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

mosquito

Variable label: Thinks avoiding mosquito bites can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

shareplate

Variable label: Thinks avoiding sharing eating utensils can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

monog

Variable label: Knows sticking to one monogamous partner can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

eatwell

Variable label: Thinks healthy diet can prevent HIV, spontaneous and prompted, of all (notes)

0	no
1	yes

Notes: Don't know, no response, never heard of HIV/AIDS, does not know HIV/AIDS can be prevented coded to zero

---

feelrisk

Variable label: Feels at medium or high risk of being infected with HIV, of all (notes)

0 no

1 yes

9 no response

Notes: Never heard of HIV and don't know if at risk coded to zero

---

friendAIDS

Variable label: Has a personal friend infected with HIV or died of AIDS, of all (notes)

0 no

1 yes

8 don't know

9 no response

Notes: Never heard of HIV coded to zero

VCT AND TESTING VARIABLES

evervct

Variable label: Has ever had an HIV test at own request, of all respondents

0 no

1 yes

9 no response

Denominator: All respondents

Numerator: Respondents who have ever had an HIV test, and who say the test was at their own request, regarless of whether they got results

Zero value: Respondents who have never heard of HIV, have never taken an HIV test, or have taken an HIV test at the request of someone else (not voluntary)

---

vctgoodall

Variable label: Has ever had an HIV test, got counselling and received test results, of all respondents

0 no

1 yes

Denominator: All respondents

Numerator: Respondents who have ever had an HIV test, who had counselling when they had the test, and who got the test results

Zero value: Respondents who have never heard of HIV, have never taken an HIV test, or have taken an HIV test but did not get counselling, and/or did not get results

---

evertest

Variable label: Has ever had an HIV test, of all respondents

0	no
1	yes
8	don’t know
9	no response

Denominator: All respondents

Numerator: Respondents who believe they have ever been tested for HIV, whether or not they received counselling or test results

Zero value: Respondents who have never heard of HIV, or have never knowingly been tested for HIV

---

testresults

Variable label: Of those who have had HIV test, got results

0	no
1	yes
9	no response

Denominator: All respondents who believe they have ever been tested for HIV

Numerator: Respondents who believe they have ever been tested for HIV, and were given their test results

Zero value: Respondents who believe they have ever been tested for HIV, but never received their test results

---

testresultsall

Variable label: Tested for HIV and got results, of all respondents

0	no
1	yes
9	no response

Denominator: All respondents

Numerator: Respondents who believe they have ever been tested for HIV, and were given their test results

Zero value: Respondents who have never been tested for HIV, or believe they have ever been tested for HIV, but never received their test results

---

counselling

Variable label: Of those who have had an HIV test, received counselling

0 no

1 yes

9 no response

Denominator: All respondents who believe they have ever been tested for HIV

Numerator: Respondents who believe they have ever been tested for HIV, and received pre/post test counselling

Zero value: Respondents who believe they have ever been tested for HIV, but did not receive pre/post test counselling

---

vctyearall

Variable label: Been tested for HIV with counselling and results in the last year, of all respondents

0 no

1 yes

Denominator: All respondents

Numerator: Respondents who had an HIV test in the 12 months prior to survey, had counselling when they had the test, and got the test results

Zero value: Respondents who have never heard of HIV, have never taken an HIV test, did not have an HIV test in the 12 months prior to survey or have taken an HIV test but did not get counselling, and/or did not get results in the 12 months prior to survey

---

refervct

Variable label: Has been refered to VCT services by outreach worker in TR (notes)

0 no

1 yes

9 no response

Notes: “Don’t know” coded to no

Denominator: All respondents

Numerator: Respondents who report being refered for specific HIV counselling and testing services in TR

Zero value: Respondents who have not had contact with outreach workers, or have had contact with outreach workers but say they were not refered for specific HIV counselling and testing services in TR

STI/RTI VARIABLES

anysti

Variable label: Reports any symptom of STI/RTI in TR, of all sexually active respondents (notes)

0 no

1 yes

Notes: This variable is created from answers to prompted questions about particular physical symptoms (not related in the questions to STIs). For women: abnormal vaginal discharge, genital ulcers, genital swelling. For men: urethral discharge, genital ulcers, genital swelling. Male-male sex groups also includes anal discharge

Denominator: All sexually active respondents

Numerator: Sexually active respondents reporting at least one STI/RTI symptom, when prompted (as per notes) in TR (usually 12 months prior to survey)

Zero value: Sexually active respondents saying they did not experience any STI/RTI symptoms in TR (usually 12 months prior to survey) in response to specific prompted questions (see notes)

---

treat

Variable label: What respondent did about most recent episode of STI/RTI symptoms, of those reporting symptoms in TR

1 treated at medical facility

2 treated by traditional healer

3 did nothing

4 self-treated with antibiotic

5 self-treated with herbs

Denominator: All respondents reporting specific symptoms of STI/RTI s in TR

---

treatwhere

Variable label: Treated at which health facility, of those treated at health facilities

1 hospital

2 primary clinic

3 private doctor

4 nurse

5 ngo clinic

6 brothel clinic

7 other

Denominator: All respondents reporting specific symptoms of STI/RTI in TR, and saying they sought treatment at a medical facility for the most recent episode of symptoms

---

selftreatfirst

Variable label: Self-treated before going to health facility, of those who went to health facility for STI/RTI symptoms

0	no
1	yes
9	no response

Denominator: All respondents reporting specific symptoms of STI/RTI in TR, and saying they sought treatment at a medical facility for the most recent episode of symptoms

Numerator: Respondents who reported going to medical facilities for most recent STI/RTI symptoms who reported, on further probing, that they had self-medicated before seeking medical help that time.

Zero value: Respondents who reported going to medical facilities for most recent STI/RTI symptoms who reported, on further probing, that they did not self-medicate before seeking medical help that time.

---

selftreatany

Variable label: Any self-treatment for STI/RTI, of those reporting STI/RTI symptoms in TR

0	no
1	yes

Denominator: All respondents reporting specific symptoms of STI/RTI in TR

Numerator: Respondents reporting specific symptoms of STI/RTI in TR who reported any self-medication, whether or not they also sought other treatment services.

Zero value: Respondents reporting specific symptoms of STI/RTI in TR who did not report any self-medication at all.

---

ulcer

Variable label: Has had ulcer on genitals in TR, of all sexually active respondents

0	no
1	yes
8	don’t know
9	no response

Denominator: All sexually active respondents

Numerator: Sexually active respondents who said in response to prompted question that they had an ulcer or sore on their genitals in TR

Zero value: Sexually active respondents who said in response to prompted question that they never had an ulcer or sore on their genitals in TR

---

swelling

Variable label: Has had swelling around genitals in TR, of all sexually active respondents

0	no
1	yes
8	don’t know
9	no response

Denominator: All sexually active respondents

Numerator: Sexually active respondents who said in response to prompted question that they had a swelling around their genitals in TR

Zero value: Sexually active respondents who said in response to prompted question that they never had a swelling around their genitals in TR

---

discharge

Variable label: Has had smelly discharge (FSW), urethral discharge (men) in TR, of all sexually active respondents

0	no
1	yes
8	don’t know
9	no response

Denominator: All sexually active respondents

Numerator: Sexually active respondents who said in response to prompted question that they had discharge in TR

Zero value: Sexually active respondents who said in response to prompted question that they never had discharge in TR

---

analdischarge

Variable label: Had anal discharge in TR, of all sexually active respondents

0	no
1	yes
8	don’t know
9	no response

Denominator: All sexually active respondents (usually only asked for MSM groups)

Numerator: Sexually active respondents who said in response to prompted question that they had anal discharge in TR

Zero value: Sexually active respondents who said in response to prompted question that they never had an anal discharge in TR

---

stdclinic

Variable label: Attended clinic for routine STI/RTI checkup in TR, of sex worker populations

0	no
1	yes

Denominator: All sex worker respondents

Numerator: Sex worker respondents who report attending a clinic for routine STI/RTI screening and treatment in TR

Zero value: Sex worker respondents who report not having attended clinic for routine STI/RTI screening and treatment in TR

---

freqstdclinic

Variable label: Frequency with which attended clinic for routine STI/RTI checkup in TR (often 3 months)

1	never
2	once
3	2-3 times
4	> 3 times
8	don’t remember
9	no response

---

referstd

Variable label: Has been refered to STI/RTI clinic by outreach worker in TR (notes)

0	no
1	yes
9	no response

Notes: “Don’t know” coded to no

Denominator: All sex worker respondents

Numerator: Sex worker respondents who report being refered to STI/RTI clinic for routine screening and treatment in TR (usually three months)

Zero value: Sex worker respondents who have not had contact with outreach workers, or who have had contact with outreach workers but say they were not refered to STI/RTI clinic for routine screening and treatment in TR (usually three months)

---

partnerreferal

Variable label: Encouraged partner to seek std treatment at medical facility, of non sex worker men with STI/RTI symptoms in TR

0 no

1 yes

9 no response

Denominator: All non-sex worker male respondents reporting specific symptoms of STI/RTI in TR

Numerator: Non-sex worker male respondents reporting specific symptoms of STI/RTI in TR who said they encouraged their partner to seek treatment at a medical facility

Zero value: Non-sex worker male respondents reporting specific symptoms of STI/RTI in TR who did not encourage their partner to seek treatment at a medical facility, regardless of whether they themselves sought treatment.

ALCOHOL AND DRUGS (all populations)

alcohol

Variable label: Has drunk alcohol in TR, of all

0 no

1 yes

9 no response

Denominator: All respondents

---

drunk

Variable label: Has been drunk in TR, of all

0 no

1 yes

9 no response

Denominator: All respondents

---

drugs

Variable label: Has ever used illegal drugs, of all

0 no

1 yes

9 no response

Denominator: All respondents

---

iduever

Variable label: Has ever injected drugs for recreation/to get high, of all

0 no

1 yes

9 no response

Denominator: All respondents

Numerator: Respondents who say they have ever injected drugs for recreation/to get high

Zero value: Respondents who have never used drugs, or who have used drugs but never injected drugs

---

iduyear

Variable label: Has injected drugs for recreation/to get high in the last year, of all

0 no

1 yes

9 no response

Denominator: All respondents

Numerator: Respondents who say they have ever injected drugs for recreation/to get high within the 12 months preceding the survey

Zero value: Respondents who have never used drugs, or who have used drugs but never injected drugs, or who have injected drugs for recreation/to get high, but not in the last 12 months

---

partneridu

Variable label: Sex partner injects drugs for recreation/to get high

0 no

1 yes

8 don’t know

9 no response

Denominator: All respondents (sometimes asked only of IDU respondents and sex worker respondents)

Numerator: Respondents who say they have a sex partner who currently injects drugs for for recreation/to get high

Zero value: Respondents who are not sexually active, who say they do not have any sex partners who currently inject drugs for recreation/to get high

---

PREVENTION INTERVENTION VARIABLES

anyintervention

Variable label: Reports participating in any HIV prevention intervention in TR, of all respondents (notes)

0 no

1 yes

9 no response

Notes: A respondent is coded as participating in an intervention if they report contact with an outreach worker in TR, having received IEC material in TR, having received free condoms (and lubricant for MSM groups) in TR, and having attended a clinic for routine STI/RTI screening and treatment in TR. In addition, IDU are coded to 1 if they attend a needle exchange or methadone maintenance programme.

Denominator: all respondents

---

anyoutreach

Variable label: Has been contacted by any prevention outreach worker with information about HIV in TR, of all (notes)

0 no

1 yes

9 no response

Notes: “Don’t know” coded to zero

---

anycondom

Variable label: Has received condom from outreach worker or other free source in TR, of all (notes)

0 no

1 yes

9 no response

Notes: “Don’t know” coded to zero

---

anyiec

Variable label: Has received brochure or comic about HIV in TR, of all (notes)

0 no

1 yes

9 no response

Notes: “Don’t know” coded to zero

---

needleexchange

Variable label: Has participated in needlee exchange in TR, of all IDU

0	no
1	yes
9	no response

Denominator: All IDU respondents

Notes: “Don’t know” coded to zero

---

freqneedleexchange

Variable label: Frequency of use of needle exchange in TR (often 1 month), of all

1	never
2	less than once a week
3	once a week
4	more than once a week
8	don’t remember
9	no response

Denominator: All IDU respondents

---

whointerventionall

Variable label: Reached by intervention run by whom, of all respondents

1	no intervention
2	only NGO
3	only govt
4	NGO and govt
8	don’t know

---

anyngo

Variable label: Has participated in HIV prevention with NGO in TR, of all

0	no
1	yes
8	don’t know
9	no response

Denominator: All respondents

Numerator: Respondents who say they participated in intervention(s) run by an NGO in TR

Zero value: Respondents who did not participate in any intervention in TR, do not know if they participated in any intervention in TR, or participated in intervention(s) in TR, but not one run by an NGO

Don’t know value: Respondents who say they participated in intervention(s) in TR, but do not know who the intervention is run by.

---

anygovernment

Variable label: Has participated in HIV prevention run by government agency in TR, of all

0	no
1	yes
8	don't know
9	no response

Denominator: All respondents

Numerator: Respondents who say they participated in intervention(s) run by any government agency in TR

Zero value: Respondents who did not participate in any intervention in TR, do not know if they participated in any intervention in TR, or participated in intervention(s) in TR, but not one run by a government agency.

Don't know value: Respondents who say they participated in intervention(s) in TR, but do not know who the intervention is run by.

---

freqoutreach

Variable label: Frequency of contact with prevention outreach worker in TR (often 3 months), of all

1	never
2	once
3	2-3 times
4	> 3 times
8	don't remember
9	no response

Denominator: All respondents

IDU VARIABLES

Note that drug injecting cultures very widely between countries and locations. IDU questionnaires are therefore necessarily far more varied than questionnaires for other high risk group. This list of variables is illustrative — may of the recodes may be impossible or inappropriate in different situations.

DRUG USING BACKGROUND

timedrugs

Variable label: Years using any kind of drugs

Denominator: All IDU respondents

---

timedrugsgroup

Variable label: Years using any drugs (grouped)

1 a year or less

2 2-3 years

3 4-10 years

4 over 10 years

Denominator: All IDU respondents

---

timeidu

Variable label: Years injecting drugs

0 less than one year

Denominator: All IDU respondents

---

timeidugroup

Variable label: Years injecting drugs (grouped)

1 a year or less

2 2-3 years

3 4-10 years

4 over 10 years

Denominator: All IDU respondents

---

transition

Variable label: Years on drugs before starting to inject (notes)

Note: This is coded from the length of time taking any drugs, minus the length of time injecting drugs

Denominator: All IDU respondents

---

injectelsewhere

Variable label: Injected drugs in another city/district in TR

0 no

1 yes

9 no response

Denominator: All IDU respondents

INJECTING FREQUENCY

freqinjectday

Variable label: Times injected yesterday (day preceding survey)

Denominator: All IDU respondents

---

freqinjectweek

Variable label : Times injected in the seven days preceding survey (notes)

Notes: “Don’t remember” coded to 80th percentile, excluding zeros

Denominator: All IDU respondents

---

freqinjectgroup

Variable label: Usual frequency of injection

1	once a day or less
2	2-3 times a day
3	4 or more

Denominator: All IDU respondents

NEEDLE SHARING AND CLEANING

sharelast

Variable label: Took used needle/syringe from other user or gave to other user at last injection

0	no
1	yes
8	don’t know
9	no response

Denominator: All IDU respondents

Numerator: Respondents who injected in TR and used a needle/syringe that had previously been used by another injector at last injection, or who passed on the needle/syringe to another user after using it themselves at last injection

Zero value: Respondents who did not inject in TR or who injected in TR but did not use a needle/syringe that had previously been used by another injector at last injection, and did not pass on the needle/syringe to another user after using it themselves at last injection

---

anyinjectingrisk

Variable label: Shared a needle, used a public needle or had professional injection in TR

0	no
1	yes
8	don’t know
9	no response

Denominator: All IDU respondents

Numerator: Respondents who used a previously used needle/syringe in TR (usually one week). This includes use of a needle stored in a public place, and injection by a shooting gallery/professional injector in situations where shooting galleries do not use a new needle for every injection.

Zero value: Respondents who did not inject in TR or who only ever used a new needle or a needle which no-one but themselves had previously used, and did not pass on any needles to other users.

---

numsharelastall

Variable label: Used the same needle/syringe before and after respondent at last injection in TR (note)

Note: People who did not inject in TR are coded to zero

Denominator: All IDU respondents

---

sharetakelast

Variable label: Took used needle/syringe from other user at last injection

0 no

1 yes

9 no response

Denominator: All IDU respondents

Numerator: Respondents who injected in TR and used a needle/syringe that had previously been used by another injector at last injection

Zero value: Respondents who did not inject in TR or who injected in TR but did not use a needle/syringe that had previously been used by another injector at last injection

---

sharegivelast

Variable label: Gave needle/syringe to other user after using it at last injection

0 no

1 yes

9 no response

Denominator: All IDU respondents

Numerator: Respondents who injected in TR and passed on the needle/syringe to another user after using it themselves at last injection

Zero value: Respondents who did not inject in TR or who injected in TR but did not pass on the needle/syringe to another user after using it themselves at last injection

---

shareworkslast

Variable label: Shared any injecting equipment at last injection, including needle, syringe, water, setting

0 no

1 yes

9 no response

Denominator: All IDU respondents

Numerator: Respondents who injected in TR and shared any injecting equipment at last injection, incl needle, syringe, water, setting at last injection

Zero value: Respondents who did not inject in TR or who injected in TR but did not share any injecting equipment at last injection, including needle, syringe, water, setting at last injection

---

professional

Variable label: Respondent had injection at shooting gallery/by professional injector in TR

0 no

1 yes

9 no response

Denominator: All IDU respondents

Numerator: Respondents who injected in TR and had injection at a shooting gallery or by a professional injector in TR

Zero value: Respondents who did not inject in TR or who injected in TR but did not have injection at a shooting gallery or by a professional injector in TR

---

cleaningpractices

Variable label: How needle was usually cleaned, of those sharing in the last week (notes)

1 dangerous

2 ineffective but not dangerous

3 most effective

Notes: In this code, "dangerous" practices are no cleaning at all, or cleaning with shared water, wiping with rag, "ineffective but not dangerous" is cleaning with clean, hot or soapy water, "most effective" is cleaning with bleach or alcohol

Denominator: IDU respondents who ever shared a needle in the last week

---

cleanhow

Variable label: How needle was usually cleaned, of those sharing in the last week

1 needle not usually cleaned

2 cleaned with used water

3 cleaned with new/ hot/ soapy water

4 cleaned with bleach

5 cleaned with alcohol

6 wiped with rag

Denominator: IDU respondents who ever shared a needle in the last week

---

bleach

Variable label: Usually cleans needle with bleach of those sharing in the last week

0 no

1 yes

9 no response

Denominator: IDU respondents who ever shared a needle in the last week

Numerator: Respondents who shared a needle in the last week and usually cleans needles with bleach or clinical alcohol

Zero value: Respondents who shared a needle in the last week, but does not clean needles or does not usually cleans needles with bleach or clinical alcohol between users

IDU VULNERABILITY VARIABLES

prisonyear

Variable label: Respondent has been to prison in 12 months preceding survey

0 no

1 yes

9 no response

Denominator: All IDU respondents. Note in some countries this variable can also be coded for sex worker groups

Numerator: Respondents who have been to prison in 12 months preceding survey

Zero value: Respondents who have never been to prison, or have been to prison but not in 12 months preceding survey

prisonever

Variable label: Respondent has ever been inprisoned

0 no

1 yes

9 no response

Denominator: All IDU respondents. Note in some countries this variable can also be coded for sex worker groups

Numerator: Respondent who have ever been inprisoned

Zero value: Respondent who have never been inprisoned

injectprison

Variable label: Respondent had drugs injected in prison, of those who have ever been inprisoned

0 no

1 yes

8 don’t know

9 no response

rehabyear

Variable label: Respondent has been in a drug rehabilitation in 12 months preceding survey

0 no

1 yes

9 no response

Denominator: All IDU respondents.

Numerator: Respondent who have been in a drug rehabilitation programme in 12 months preceding survey

Zero value: Respondent who have never been in a drug rehabilitation programme, or have been in a drug rehabilitation programme but not in 12 months preceding survey

---

rehabever

Variable label: Respondent has ever been in a drug rehabilitation programme

0 no

1 yes

9 no response

Denominator: All IDU respondents.

Numerator: Respondent who have ever been in a drug rehabilitation programme

Zero value: Respondent who have never been in a drug rehabilitation programme

---

odever

Variable label: Ever overdosed on drugs

0 no

1 yes

9 no response

Denominator: All IDU respondents.

---

frienddiedod

Variable label: Person in respondent's injecting circle has ever died of drug overdosed

0 no

1 yes

9 no response

Denominator: All IDU respondents.

Family Health International, Asia and Pacific Department
19th Floor, Tower 3, Sindhorn Building
130-132 Wireless Road, Lumpini, Phatumwan
Bangkok 10330, Thailand
Tel: +662.263.2300
Fax:+662.263.2114