Time and Motion Studies for Demining: Snapshots of Operations

Geneva International Centre for Humanitarian Demining

Centre International de Déminage Humanitaire - Genève

Time and Motion Studies for Demining: Snapshots of Operations

The **Geneva International Centre for Humanitarian Demining** (GICHD) supports the efforts of the international community in reducing the impact of mines and unexploded ordnance (UXO). The Centre provides operational assistance, is active in research and supports the implementation of the Anti-Personnel Mine Ban Convention.

Geneva International Centre for Humanitarian Demining
7bis, avenue de la Paix
P.O. Box 1300
CH-1211 Geneva 1
Switzerland
Tel. (41 22) 906 16 60
Fax (41 22) 906 16 90
www.gichd.ch
info@gichd.ch

Time and Motion Studies for Demining: Snapshots of Operations, GICHD, Geneva, November 2005.
This project was managed and written by Ian McLean and Rebecca J. Sargisson, GICHD (info@gichd.ch).

ISBN 2-88487-038-5

Contents

Acknowledgements

The Geneva International Centre for Humanitarian Demining (GICHD) would like to thank Norwegian People's Aid and the Cambodian Mine Action Centre for supporting studies that contributed to this report. Tony Fish ran the study on manual demining. The GICHD would also like to thank the Governments of Norway and Sweden for funding the study.

Photo credits: Ian McLean.

1 Introduction

One way or another, anybody involved in the management of demining programmes will do cost-benefit analyses. In most cases, the aim is to optimise use of resources, where "optimisation" in terms of resource use generally also means "minimisation" in terms of cost. The trick is to do cost-minimisation without compromising essential requirements, such as productivity, safety, and quality. Unfortunately, these objectives are somewhat incompatible, and achieving them can feel like a juggling act where all the balls must be kept in the air all of the time.

Time and Motion (T&M) studies offer a mechanism for doing quantitative analyses of operational demining programmes. The primary objective of any T&M study is to develop a snapshot of the programme being studied within a defined time frame. Using that snapshot, the details of operational procedures can be explored; for example, in order to assess how resources are being used, to make comparisons among different operational situations or teams, or to test new procedures.

The concept of using T&M studies to explore a programme is not new, and in reality most managers use some kind of T&M approach when they make decisions about alternative options. The aim of this guide is to explain the principles and procedures behind T&M studies in order to make them more user-friendly and systematic.

The sorts of questions normally asked using a T&M approach include:

- Which are the most time-consuming elements of a specific demining procedure?
- Why does one team work faster than another?
- Why does productivity vary in different operational conditions?
- Will a suggested procedural adjustment improve productivity?
- Will a suggested safety adjustment affect productivity?
- Why does this supervisor appear more efficient than that one?
- And so on.

In practice, only the imagination of the questioner limits the extent to which T&M studies can contribute to improving productivity.

The above are specific examples. Stated more generally, T&M studies are used to improve understanding of the complex behavioural systems represented by a demining programme.

Why do a Time and Motion study?

Perhaps the simplest way to answer this question, is with another question: Why be ***subjective*** when it is easy to be ***objective***?

It is normal to make some kind of assessment before a new procedure is introduced, or a new idea is trialled. The assessment is likely to be comparative, in that the observer has considerable experience with the procedure being replaced and can therefore mentally compare the new procedure with the old. Some simple measurements might even be made, such as area cleared after 2 hours work using the old and new systems. However, the assessments tend to be subjective, being based on limited observations of people at work or using the new tool.

T&M studies do all of the above, and much more. Through using simple sampling procedures to gather data, they provide an objective or quantitative analysis of the system under study. In addition, they allow the observer to explore fine details of procedures, e.g. when fine-tuning an idea. Surprising things may be discovered that cannot be seen using subjective or qualitative analyses.

For example, say a new procedure reduces the time required for marking from 10 per cent to 5 per cent. This apparently small difference is unlikely to be detectable using a subjective analysis. But the new procedure has reduced the time required for marking by half (= a 50 per cent time saving). A T&M study will measure that difference, allowing the manager to make precise calculations of time savings and productivity benefits. A time saving of 5 per cent translates into one additional metre for every 20 metres cleared by a manual deminer. Calculated across 20 deminers, that represents a considerable amount of extra land, obtained at no cost. An example where reduced requirements for marking produced a measured time saving of even more than 50 per cent is given in Figure 8, below.

T&M studies are a tool. Any decision to adopt a new tool requires learning, practice, and preferably a reference manual. This document is that manual. It outlines a methodology, gives examples, and explains how to gather and work with the data that are the building blocks of a T&M study. It is not a blueprint for any particular study — every study is different and will necessarily involve local decisions and local constraints. But the principles and procedures are essentially the same for any T&M study, and they are described in detail here.

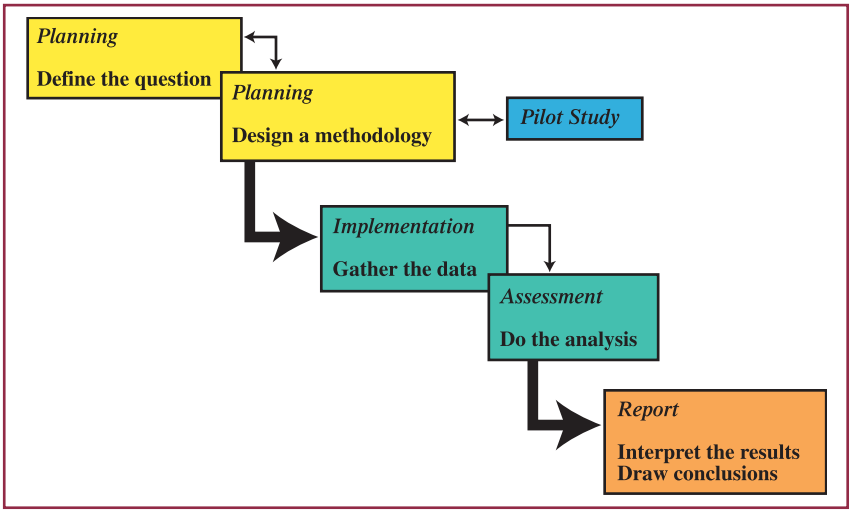
2 Anatomy of a Time and Motion study

Most T&M studies involve a comparison, such as between two teams or between two operational scenarios. However, even the simplest study in which just one operational concept is explored can provide valuable information, for example, in relation to productivity. The key to an effective T&M study is to ensure the following:

- the **question** to be asked by the study is **clearly defined**;
- the right **data** are gathered;
- appropriate **sampling** is done;
- the **analysis** addresses the question being asked;
- **reporting** requirements are part of the planning.

These points can be structured as a series of steps, shown in *Figure 1*.

Figure 1. The steps to a successful T & M study



Although they are not difficult to undertake, T&M studies require careful planning and some knowledge of data sampling and analytical procedures. The required time commitment means that the studies are unlikely to be undertaken by senior management staff (who are routinely under time pressure). Fortunately, they are easily supervised by management staff and the study itself can be undertaken by any competent person with good observation and computer skills, and an ability to understand the notion of sampling.

Planning may include either or both a **written proposal** and a **workshop**. Both is best. Management staff will normally be involved in the planning and an assigned data-gathering person or persons will do the data gathering and the analysis. If possible, a pilot study should be undertaken to check the procedures, identify elements of the study that were missed during planning, and make adjustments to suit the local context (e.g. if an important behavioural action was missed from the list of items to be recorded).

A breakdown of the steps in a Time and Motion study

Planning (1 day, up to several weeks ahead)

- Frame the question.
- Identify resource requirements.
- Design the data-gathering approach.
- Make decisions about appropriate sample sizes.
- Identify assumptions and constraints.
- Choose (and train) the person(s) to do the study.
- Decide whether statistical analysis will be needed and where the expertise will come from.
- Budget the necessary time (including for analysis and reporting).
- Workshop with management and other relevant staff.

Pilot study (1 hour to 1 day)

- Test procedures.
- Gather a small amount of test data.
- Analyse test data.
- Revise procedures.
- Revise question.
- Revise list of actions to be sampled.
- Revise assumptions and constraints.
- If multiple observers, test their agreement on definitions and procedures.
- Workshop pilot results.

Main study (1 day to 1 week)

- Gather data.
- Enter data into computer immediately (daily).
- Write down methodology (at least in note form).
- Keep diary of daily activities.

Data analysis (1 day to 1 week)

- Qualitative description — write a description of what was done.
- Graphing (quantitative description) — explore the data in different ways.
- Statistical analysis (if needed) — quantitative support for the conclusions.

Report

- The report has a standard structure: Introduction, Method, Results, Discussion, Conclusions and Recommendations.
- The **Introduction** can be extracted from a well-written project proposal.
- A description of the **Method** was outlined in the proposal, and a more detailed version was prepared during the implementation phase.
- Analysis will be more efficient if the analyst has a good understanding of an appropriate computer package, such as MSExcel®.
- A seminar or workshop may be given presenting the Results, Conclusions, and Recommendations. The seminar integrates involved parties into the process and gives them ownership of the recommendations.

Simple background rules

Applying the following rules will help to streamline the process and ensure that the project proceeds efficiently.

- **Workshop the project plan** with interested and/or affected parties before the study is undertaken.
- **Budget similar amounts of time** for each of:
  - data gathering,
  - data entry and analysis, and
  - report preparation.

  [*Warning! Do not underestimate or undervalue any of these time requirements*]
- It is best to **do data entry on the same day as the data are gathered**, even if it is necessary to compromise on time spent data-gathering in order to create the needed time. Delaying data entry leads to errors, first because details have been forgotten, and second because recording errors are less likely to be noticed if computer entry is delayed.

Useful equipment

- Portable chair.
- Binoculars.
- Two stopwatches.
- Countdown timer (giving automatic time intervals for sampling).
- Portable computer.
- Notebook and pen.
- Pre-printed data sheets (usually designed during the planning and pilot study).
- Clipboard.
- Weather and insect protection (shade, sunscreen, umbrella).

Ensure that the data gathering person can work in conditions that are as comfortable as possible, and is able to view the action in as much detail as possible!

3 Methodology

On the surface, T&M studies appear to be simple descriptions of defined activities (or behaviours), which will be sampled as defined ***actions***. So why not simply write down everything that is going on? If one does that, then it should be possible to define the question or figure out the best analysis later, because a full record of information will be available.

Better still, just video everything and analyse the video later.

To put it bluntly, **the above is impossible**. Why?

- Behaviour can change very quickly — more quickly than can be recorded.
- Writing something down requires the observer to direct attention away from the subject, and some behaviour is missed.
- Even speaking the behaviour (e.g. into a tape recorder) is too slow; some behaviour will be missed because actions occur faster than they can be spoken.
- Several behaviours of interest can occur at the same time; should all be written down or spoken, or just some?
- An observer has a broader field of vision than a video camera, and can see more detail.
- The camera itself is a distraction, as operating it draws the attention of the observer away from the action.
- It is very difficult to maintain a continuous time base for the data.
- Analysing videos (or spoken recordings) is time-consuming. At the minimum, the total length of time required to collect the data is doubled (once to film and once to analyse), but in reality setting up, finding, replaying, etc., mean that the required time is likely to be tripled. Video can be essential for obtaining fine details of specific behaviours, but cameras should not be used to replace an observer who takes data directly from the subject(s).

More than anything else, T&M studies are therefore about ***sampling***. It may be impossible to record everything, but it is certainly possible to sample

everything. Designing a sampling programme that will provide the needed snapshot is the most creative part of T&M planning. That creativity requires a good understanding of the notions of sampling as a principle, and sampling as a procedure.

Sampling: the principle

A snapshot records a moment in time. In a T&M study, that moment is a week rather than a second, but the same notion applies. The aim of the study is to build a picture of some situation. If the aim is to compare two situations, then two pictures must be built.

A picture is built using ***measurements*** of the behaviours (or activities, called ***actions***) that occur during the study period. Measurements are obtained through ***sampling***.

Measurements are not all of the same type. The question: *"Is your personal Protective Equipment (PPE) comfortable?"* (answer: yes or no) does not give the same type of measurement as a measurement of height, which is quantitative and has units. The types of measurements recorded during scientific sampling are described in Annex 1.

A **sample** is a single representation of something much larger. For example:

- A vial of blood is a sample of all the blood in a person's body.
- The action of a person at a single moment in time is a sample of the range of actions that person performs.
- A single measure of the length of a person's coffee break is a sample of the length of all coffee breaks taken by that person.

Sometimes, a single sample is enough to provide a useful measurement of the larger thing being studied. For example, one vial of blood is enough to provide a measurement of a person's iron level, which would be about the same, no matter how many samples were taken on a given day.

The examples above are of samples taken to represent something about an individual. But a sample can also be taken from an individual to represent a group. Thus:

- The height of one man is a sample of the height of all men.
- The iron content in the blood of a pregnant woman at 30 weeks is a sample of the iron level in all pregnant women at 30 weeks.
- The length of the coffee break of one person is a sample of the length of coffee breaks for the group that person works with.

Figure 2. Sampling a measurement of length

In the above examples, the sample and the measurement are essentially the same thing. Sample the height of a man and what you get is a measurement of height. Take a measurement of length and you have a sample (*Figure 2*).

Here are three key points that will be developed further below

- **That sample of height taken above is representative, but it is not enough if you need an estimate of the typical height of all men (or put another way, of the height of an average man). For that, the measurements for many men are needed, presumably obtained by taking samples from a reasonable number of men.**
- **Sample and measurement might not be the same thing. In the blood example above, an additional step was required between sample and measurement. A blood sample is just that — a sample. It is not a measurement, and further processing will be required before a measurement is obtained. Thus, having obtained the blood sample from a pregnant woman, a lab must then use some procedure to measure iron content. When taking behavioural samples, it is typical that an additional processing step is required between taking the sample and obtaining a measurement.**
- **When designing a study, it is normal to want to compare between two (or more) groups. Those groups could be two teams of deminers. But they could equally be the same team working at different times (often in a "before" and "after" design). E.g. if the measurement required was the change in productivity after introducing a new tool, two samples would be needed, one before and one after the tool was given to the team.**

If the question of interest is: "*Do manual deminers use their PPE correctly?*", try checking one deminer. The answer (yes or no) is a sample; essentially a snapshot of the use of PPE by that deminer.

But this single sample is not very useful because:

1. like the measurement of the height of one man, it is not representative of average use of PPE by all deminers; and
2. it is not a measurement.

The quality of information obtained in this example can be improved in two ways, both of which involve taking multiple samples:

- by taking a series of samples of the behaviour of the same deminer at different times; or
- by taking one sample from each of many different deminers at the same time (or in as short a time as possible).

The first option gives a clearer snapshot for the single deminer. The series of yes and no samples can be used to calculate a percentage of time that the deminer wears PPE correctly. E.g. with 7 "yes" and 3 "no" samples, it is easy to calculate that this deminer wears PPE correctly 70 per cent of the time. Here, 10 samples were used to make one measurement.

But if the aim is to look at what deminers do as a group, ***the second option is better***. When many samples are taken from one deminer, the data will give

a detailed assessment of use of PPE by that person, but the sample size will still be one (one deminer). That deminer may be normal, but s/he cannot be average, because an average must be calculated across a group. When you want to know about the average behaviour of a group, samples must be taken from the group.

In this example, the question above was actually phrased rather badly. Here is a more useful way to ask it:

"What percentage of deminers wear their PPE correctly?"

The percentage was already calculated for one deminer, using 10 samples (70 per cent).

Instead of concentrating on one deminer, why not make the same calculation by taking one sample from all 10 deminers in the team. Say, 7 were wearing PPE correctly and 3 were not. Just as for the single deminer, the obtained 70 per cent is still one measurement. But now it represents the wearing of PPE by the entire team.

An essential difference between a sample and a measurement is that when the data are analysed, it is measurements that are used, not samples. Except, of course, where the sample and measurement are the same thing.

Figure 3. Sampling the weather

Using samples to construct a measurement

Obtaining measurements: sampling from a distribution of actions

The process of sampling involves taking a sample in a systematic way from the available distribution of actions in time, with the aim of representing the behaviour of each subject under study. Two typical actions of manual deminers are shown in Figures 4 and 5.

Figure 4. Manual deminer working with metal detector (code e.g. MD)

Figure 5. Manual deminer prodding (code e.g. Pr)

What does it mean to say that actions have a distribution*?*

Although an action is likely to be recorded as a code (e.g. Ma = Marking), that code is easily transcribed into a number. For convenience and simplicity, the examples below use lists of numbers rather than codes, but codes could have been used. Thus the first list could have been something like:

Ma, Ex, CT, Ma, Ma, …

(where Ma=1=Marking, Ex=5=Excavation, CT=4=Change Tool)….and so on.

The following list of numbers gives an example of a ***sequence*** of actions by one subject (note that these are actions, not samples):

1, 5, 4, 1, 1, 2, 1, 1, 4, 3, 1, 5, 3, 2, 3, 1, 2, 1, 3, 2, 1, 1, 1, 4, 1, 3, 3, 1

Each number is an action, and there is no expectation that each action fills the same length of time. For example, Rest (5) usually fills a great deal of time even though it occurs just twice in the sequence. Whereas Drink (4) occurs 3 times but fills much less time than Rest.

The number "1" is common in the list, and its ***distribution is random***. It therefore occurs frequently and unpredictably. If it is a relatively lengthy action (like Rest), then its representation in the final proportion of total time should be more than the 50 per cent implied here. If it is a relatively short action (like Drink), then its representation in the final measurement should be less than the 50 per cent implied here.

The following list contains exactly the same numbers, but in a different sequence:

1, 5, 1, 4, 1, 2, 1, 4, 1, 3, 1, 5, 1, 3, 1, 2, 1, 3, 1, 2, 1, 3, 2, 1, 4, 3, 1, 3

Here, the number "1" is ***non-randomly distributed.*** It occurs frequently and predictably.

It takes only a moment of inspection to see the difference in the distribution patterns. And immediately, a question emerges — would it be interesting to study the sequence of actions in a distribution?

Yes, it might be — that will depend on the question being asked. But if the sequence of actions is not part of the research question of interest, then the sampling procedure need only be designed to capture sufficient samples of each action in the distribution to create a useful measurement.

The distribution of actions is controlled by the subject, not by the researcher. A distribution is a necessary consequence of the use of different actions by the subject through time. If the subject doing the actions is checked regularly, say once per minute, then that check will hit a regularly defined moment in the distribution. Here, the data gatherer records whichever action is being used at the moment that the subject was scanned in order to produce one overall measurement of proportion of time spent doing each action (e.g. the subject spent 60 per cent of time doing "1"). It makes no difference to the scan-sampling procedure what the distribution of actions is (random or non-random), or that some actions fill more time than others.

Obtaining measurements: deciding what to sample

In the planning and pilot stages of the study, a list of actions to be sampled will be compiled. The process of defining actions for sampling is controlled by the researcher, and is normally done on a fairly fine scale — in more detail than will eventually be needed in the analysis. It is impossible to record everything, so sampling categories need to be defined. The best way to do so is in relation to the question.

Actions may also be broken into smaller categories, called subsidiary actions. For example, excavation is a common action of deminers. But while excavating, the deminer could use a prodder, a trowel, a mattock, a spade, or a rake. Analysis of the subsidiary components of excavating might well be essential to addressing the main questions of the study.

Thus, the observer records Ex/p (Excavation/prodder) rather than just Ex (Excavation).

A subject can do even more than two actions at the same time. When the subject is doing Ex/p, are they standing? Sitting? Kneeling? Lying? Do you need to know? Go back to the question — it will help when deciding if it is important to record posture.

The requirement to break behaviour up into categories applies to anything being sampled. The subject could be a machine, a dog, a handler, a team, a supervisor or a deminer.

No matter how much planning was done, it is possible for a new action to appear at some point during the study. You can choose to begin sampling it

from the point in time when it first appears because it has not been available to sample so far, so has not been missed. It is better not to define a new action for sampling part way through the study. If that action was previously either not being sampled, or more likely, was being sampled as part of a broader action, then introducing it part-way through will compromise the data both for the new action and the other action of which it was previously a part. Identifying such actions is one reason for the pilot study.

Other useful rules:

- Sample on a finer scale than you think you will need in the analysis. You can always combine the data during analysis, but you cannot break it up into smaller components. In other words, subsidiary actions, such as prodding, can always be lumped into major actions, like excavation. If only excavation is recorded, you cannot later look specifically at prodding.
- Use a logical coding system for naming each action (e.g. a 2-letter code).
- Work with about 10 major action categories, and no more than 15.
- Aim to get about 10 times the number of samples as major action categories (e.g. with 10 action categories, aim for about 100 samples; with 15 action categories aim for about 150 samples).
- If some categories have subsidiary actions, apply the same rule of 10 as above (use up to about 10 subsidiary categories).
- Record no more than 2 actions at one moment (i.e. one action and one subsidiary action). Trying to record more often leads to confusion and errors. Thus recording Ex/p, or Pr/St is ok, but recording Ex/Pr/St (Excavating/Prodding/Standing) is not a good idea.

Obtaining measurements: sampling procedures

Actions come in every conceivable form (quick, lengthy, complex, simple, common, rare, etc). Thus no single procedure is suited to sampling all of them. In practice, an observer will often use several different sampling procedures at the same time, and may even switch between techniques during the sampling period. As always, it all depends on the question.

A second consideration is the time required to do the sampling. In general terms it takes an observer about 10 seconds to take and record a sample, and that imposes a limit on the number of samples that can be taken per minute.

What sampling procedures are available?

Time taken

- Record start and end time of an action to obtain the time spent doing that action.
- Normally used for long actions.
- If the action is very short, the between-action interval might be measured.

- Delivers total time spent on an action, and the data are easily converted to proportions or rates.

Example: How frequently does a manual deminer drink water while working?

Wait for the deminer to drink, start a stopwatch, and wait for the next drink. The time, a sample of the inter-drink interval, is a measurement of drinking frequency (e.g. 13 minutes) and can easily be converted to a rate (= 4.2 drinks/hour). While the stopwatch is running, the sampler continues with other sampling procedures while keeping an eye on the subject for whom inter-drink interval is being measured.

Scan sampling

- Delivers frequencies of each action, or proportion of time spent doing each action.

For the above question about drinking, check the deminer **for an instant** on a regular time cycle, say every minute, and record whether s/he is drinking. At the end of 2 hours there have been 120 checks (= 120 samples). Drinking is a quick action and some drinking events will be missed. But that does not matter. If the deminer is drinking regularly, then drinking will be the action on a few of the checks. Using those 120 samples, the analyst calculates one measurement which took 2 hours to obtain:

the deminer was drinking on (e.g.) 3 out of 120 checks (2.8 per cent).

This measurement can be restated as: **the deminer spent 2.8 per cent of time drinking.**

This is not the same measurement as the 13-minute inter-drink interval obtained above, because the measurement is a percentage of time drinking, and is not a frequency or rate of drinking.

The main benefit is that the observer had lots of time left over to gather other data because one deminer was not continuously watched.

There is lots of sampling time available, so why not scan the entire team regularly, say every minute, and record the name of any drinking deminer on each scan. In other words, why not address the question using all 10 deminers at the same time. As above with sampling from one deminer, the problem is that the action "take a drink" is quick, and some drinking events will be missed. But that does not matter because the objective is to measure the proportion of time drinking, not the total number of drinks taken. Drinking by the deminers has now been sampled on a short time cycle. Percentage of time spent drinking can be calculated for each of the 10 deminers rather than just one at no cost in observer time. The observer has obtained 10 measurements (one for each subject) and there is still time left over to record other information.

During the scan above, each deminer was checked to see whether or not s/he was drinking. So why not record whatever they were doing — any action, not just drinking? Choose (e.g.) 6 general activities of deminers. Record which of these 6 actions each deminer is engaged in on each scan. Once sampling is finished, the percentage of time spent doing each of 6 actions can be calculated for each deminer. The data now give a lot more detail about the

activities of deminers (including drinking, if that was one of the recorded actions). If a 1-minute scan time and 2-hour sampling time were used, then the total N is still 10 deminers from which 120 samples were taken (=1,200 samples in total). Six measurements were obtained for each deminer (a proportion of time for each action), giving a total of 60 measurements.[1]

Count

In a defined time period, how many times did an action occur.

- Normally used for quick actions and rare actions.
- Easily converted to rates.

Qualitative description

Describe in words how an action is performed.

- Normally used to describe complex actions and rare actions.

Sequence record

- Normally used to describe the behavioural sequence in a complex action.

A final comment

Time taken and scan sampling are the most frequently used procedures when building a snapshot. Scan sampling is by far the best method to provide a broad picture of the overall proportions of time spent in each action in order to build a snapshot of behaviour. However, scan sampling gives little information on rare or quick actions (e.g. it might miss them completely), some of which may be important in relation to the research question. For these, any of the other procedures might be used.

An example

A broad descriptive question is being addressed:

"How do deminers use their time while working?"

The observer has chosen to use **scan sampling** of 10 subjects (5 pairs of manual deminers) throughout an 8-hour working day, on a 2-minute scan cycle. Twelve actions have been defined for sampling, two of which have subsidiary actions. Subjects are working 1 hour on (working) and 1 hour off (resting). Activities during rest time are not interesting and are not sampled, so the observer will sample from 5 deminers on each scan, leaving some time available in each 2-minute cycle to watch for rare or complex activities.

At the end of the day, up to 120 (4 hours worked x 30 scans/hour) scan samples have been obtained for each of the 10 subjects (1,200 samples) and the proportions of time each subject spent doing the 12 different actions can be calculated. The rule of obtaining at least 10 times the number of samples as actions was satisfied.

1. There is one limitation — it will take a little longer for the observer to record the action of 10 people on every scan (rather than simply recording the name of the occasional drinking deminer). An observer who is familiar with the scanning process and has developed an efficient recording system will normally need about 10 seconds to locate, identify, and record the activity of one deminer. Thus, a scanning cycle of about 2 minutes will be required for 10 people.

Fifteen indications were recorded (**count** — easily obtained because the observer could hear the metal detectors). The observer used a corner of the sampling sheet to make a mark each time an indication was heard.

One mine was discovered (**count** of a rare event), and the observer missed 3 scans during the 6 minutes that the mine was being uncovered and marked, preferring to record all details of that process (**qualitative description**). However, the SOP requires all deminers to stand down during uncovering of a mine, so their actions during those 3 missed scans were known (Resting, Rs), and therefore were not lost from the data. Productivity loss due to standing down can be calculated directly from the **time taken** to deal with the mine (6 minutes lost by 4 deminers), and there is no need to enter standing down as a separate activity from Resting in the computer file, although that decision is up to the observer.

Management suspects that the deminers are chatting a lot during handover, and requested precise records on how long it was taking. Handover of the lane between pairs of deminers was a predictable event because it occurred on the hour. Each time there was handover, one pair of deminers was chosen and the handover timed precisely using a stopwatch (**time taken**). At the end of the day, the time taken for 7 handovers had been recorded precisely for 4 pairs with accuracy to within 2 seconds.

The supervisor circulated regularly and occasionally talked to a deminer. It has been suggested that this supervisor talks to deminers for much longer than other supervisors (wasting time?), and the observer was asked to measure how long these conversations take (**time taken**). Talking to Supervisor (TS) is also one of the 12 actions being sampled, so the study will return two separate measures on this activity (proportion of time talking to supervisor; average length of a conversation with supervisor). The observer recorded the length of conversations opportunistically: if the supervisor was seen approaching a deminer, a stopwatch was started (**time taken**), and the observer checked the talking pair regularly while continuing to record scan samples. The obtained 17 samples of time spent talking were accurate to within 10 seconds. Many conversations were missed (the scan sampling returned TS on 122 scan samples across all the deminers), but that does not matter as the timing records were opportunistic and are a random sample of those conversations.

In this example, the observer used 4 of the 5 listed sampling procedures, and at times was using 2 procedures at the same time (e.g. **scan sampling**, and stopwatch running during a conversation — **time taken**). Depending on how the description of the exposure and marking of the mine was written down, it might also be possible to look at the **sequence record** for that rare event.

The scan samples do not provide a sequence record. Using scan sampling, only one of the many actions performed by a subject is recorded each 2-minute sampling cycle. A sequence record can only be taken during intensive observation of one subject.

Let's be realistic. This is intensive work. The observer will be exhausted at the end of the day, and might need to schedule a 5-10 minute break every hour. Perhaps a small amount of data will be lost during the break, but as

long as the missed data are not biased in some way (e.g. the handover on the hour is always missed); no problem! After all, you are **sampling** — you are not attempting the impossible of recording everything that deminers do.

After sampling is finished for the day

The observer is not yet finished for the day. Now, the data must be entered into the computer.

There were 240 scan-sampling events, each with a sample from 5 deminers. That is 240 lines of data, requiring 2 hours of data entry. Separate files will contain the 17 samples of time spent talking, and 7 samples of time taken during handover. The qualitative description of exposing the mine must be transferred from the field notes into the log book or diary (which might be another computer file). The diary itself needs to be updated with a general description of the day's activities. With backing up data files, the observer has about 3 hours more work to do, and **it must all be done today**.

Don't want to do it today? Any delay raises the chances of errors. Right now, the day's sampling is fresh in your mind. You can still remember details that might be confused in your notes, or where your writing is difficult to read, or where rain smudged the writing. Wait even one day and all those details will be gone — replaced by the details of tomorrow's sampling.

The question?

In practice, it is likely that the question involves making a comparison, for example between manual deminers working behind a flail and with no flail. Thus, tomorrow, the same data will be gathered from a team working in a minefield that was previously cleared with a flail (or alternatively a different observer could have sampled at that other minefield on the same day). Perhaps the research design required that the same team be used under both conditions (flail, no flail), in which case the sampling must be done on different days.

Or perhaps tomorrow a different metal detector will be used by the same team. Or tomorrow this team will receive training on a different procedure, and then on Day 3 you will conduct the same study on them using the new procedure (a before/after study). Or … and so on. You designed the experiment, so you will know what to do.

Analysis

Having obtained the data, many different analyses will be possible. Here are the likely steps:

- Where necessary, **convert the data** from a sample to a measurement. The number of times each action was recorded can be counted for each deminer, which allows the proportion or percentage of time spent doing each of the 12 sampled actions to be calculated (as a proportion of the total number of samples taken). In the example above, 12 actions were sampled, so 12 measurements will be produced for each deminer. There were two subsidiary actions, for which the proportion might also be calculated.

- No conversion is needed for the 17 time taken measures of supervisor-deminer conversations, or the 7 time taken measures of handover, although you may want to calculate the average time spent talking by each deminer with the supervisor.
- **Return to the question**. What are the most useful comparisons? What are the most useful summaries? You might do all possible analyses in the interests of completeness, but there is no need to do so.
- **Check the data**. Any values that are obviously out of line with the others may include a mistake. Twelve actions mean 12 codes. If there are 13 codes then there is an entry error. The 12 proportions calculated for each deminer should add up to 100 per cent. Typical mistakes include a misplaced decimal point, entering the wrong code for one subject, or data entered using different units (e.g. 15 of the 17 time taken measures were entered as seconds, and 2 were entered as minutes). In practice, the opportunity for error is large, even with careful data entry. Assume there are errors, and find ways to search for them.
- **Explore the data**. For example, plot preliminary graphs. Such graphs are unlikely to be the final figures used in the report, but they allow visual inspection of the data so that you can see the trends and check for errors.
- **Statistical analysis**
  - The technical meanings of "significant", "variance", and "$P<0.05$" are described in Annex 2.
  - Statistical analysis includes summarising the data (such as calculating means), plotting graphs, making visual inspection of trends, and statistical testing.
  - In the example above, 10 deminers were observed. Thus the sample size (N) for most analyses will be 10. The variability amongst those 10 deminers is one of the most valuable features of the data, and should not be ignored.
  - Analysis may or may not include use of statistical testing in order to make comparisons among groups. Whether or not such testing occurs is likely to depend primarily on the skills of the analyst (who may not be the observer). In some sets of data, the patterns are obvious and statistical analysis is unnecessary. Statistical testing is all about making comparisons between groups, and it may be that no such comparisons are planned. Statistical testing is the formal version of what you have already done (summarise data, plot graphs, inspect trends).
  - The central point is that by following the principles outlined in this guide, you have gathered data that support statistical testing, potentially making the results of the study much more convincing. Statistical testing can be used to find trends that are not obvious from visual inspection, or to show that apparent differences are not real. Do not be afraid to consult someone with better statistical skills than yours. Statisticians love playing with other people's data!
  - One other point. Politicians might use statistics in misleading ways

to misrepresent trends or patterns, **but statisticians do not**. A statistician will tell you what your results show and what they do not show. No more and no less.

- New ideas often emerge as a result of the above process. The study will both give answers and raise new questions, some of which are interesting and some are not. It is not possible to measure everything, and having measured certain things in relation to the question, you discover something new and interesting that you now wish you had measured. This process is the source of the complaint often heard from non-researchers, that "*research never gives answers, it only ever leads to more research*". But that complaint misses the point. Answers were obtained for the central question of the study. But answers certainly were not obtained for all possible questions that might have been raised, some of which were only recognised because the study was done.
- Complaints often heard in the reporting-back workshop are:
  - "*that is not the question that I would have asked…*", or
  - "*why did you not ask this other question…*".

Avoiding such comments is what the original planning and workshopping was all about! On the other hand, you may genuinely have identified a new and interesting question, and some follow-up work is therefore justifiable.

Reporting

The reporting requirements will depend on the client who has requested the study, and the extent to which others want access to your results, or to which your results have generality.

In principle, any study of this sort should be properly written up, and the report made widely available. It is only by sharing information that the demining community will improve the quality and quantity of its product. If the report is available, then they can look at it and decide if it is useful to them.

The standard structure of a report can be found in the Box on page 20.

A standard report will have the following sections

➢ ***Introduction***

- **This section explains the context of the study and gives a clear statement of the question being addressed.**
- **Most of this is already written in the original proposal document. The terms of reference might be laid out here.**

➢ ***Methods***

- ***How* did you do the study?**

➢ ***Results***

- ***What* did you find? Describe the results in words, with reference to tables or figures. Do not say "*the result can be seen in the figure*", which is the same as saying "*go figure it out for yourself*".**
- **In general, the results section should not include discussion or interpretation comments.**

➢ ***Discussion***

- ***Interpret* the results in relation to the question.**
- **The discussion is not an opportunity to write for multiple pages about anything that seems interesting. It should be kept focused and short.**

➢ ***Conclusion***

- **The *key points* that arise from the study, usually in dot-point form.**

➢ **For a larger study, a *Summary* will normally be provided, usually at the beginning of the report. An *Executive Summary* will give key points in a half to one page. Busy managers love well-written executive summaries!**

➢ **A *Recommendations* section may be included, normally also at the beginning of the report.**

4 Example Time and Motion studies

The following examples are taken from real studies done on demining programmes. The full studies can be found in the cited GICHD reports.

Example 1: What do the raw data look like in a snapshot of a demining programme?

Table 1 shows a small portion of the data that were recorded for Standard drill in the study described in Questions 3 and 4 below. The numbers 1-10 are ten 1-minute scan samples. A-H are individual deminers. The 2-3 letter codes are sampled actions; e.g. MD = using Metal Detector, MKG = Marking.

Table 1. Data taken during scan sampling of Standard demining drill, for 8 deminers (A-H)

	A	B	C	D	E	F	G	H
1	MEC	CV	MD	TS	CV	TS	CV	MKG
2	MD	CV	MD+	TS	CV	MD	CV	CV
3	MKG	MD	MD+	MEV	MD+	MEC	MEC	CV
4	MKG	ISP	CT	CV	MKG	MKG	MKG	MD
5	MKG	ISX	MD+	MD	MKG	MKG	MKG	RV
6	MEC	CT	ISP	MD	MKG	TS	CV	CV
7	MD	ISX	ISX	MKG	TS	CV	CT	CV
8	MD	ISD	MD	CV	CV	MD+	MKG	CT
9	MKG	MKG	ISD	MKG	MD+	MD+	TS	CV
10	CV	MKG	MD	TS	CT	PW	CV	CV

In order to convert these codes to measurements, the count of each code needs to be obtained (the computer can do that), and then a proportion will

be calculated for each code representing the proportion of time spent doing each action. Percentages of time spent doing each action in Table 1 are calculated in Table 2.

Note that in this study, 22 actions were sampled, but some lumping of data was done across actions, resulting in 15 activities for which proportion of time was calculated. Only some of those activities appear in this subset of the data.

Table 2. Calculated proportions (as %) of time spent doing each of 15 activities during Standard drill for the 8 deminers (A-H) in Table 1*

Activity	A	B	C	D	E	F	G	H
TS	0	0	0	30	10	20	10	0
VEG	10	20	0	30	30	10	40	70
MCL	20	0	0	0	0	10	10	0
CT	0	10	10	0	10	0	10	10
MD	30	10	30	20	0	10	0	10
MD+	0	0	30	0	20	20	0	0
WAT	0	0	0	0	0	10	0	0
ISP	0	10	10	0	0	0	0	0
ISX	0	20	10	0	0	0	0	0
ISD	0	10	10	0	0	0	0	0
PPE	0	0	0	0	0	0	0	0
QA	0	0	0	0	0	0	0	0
MKG	40	20	0	20	30	20	30	10
DP	0	0	0	0	0	0	0	0
RST	0	0	0	0	0	0	0	0

* The many zeros are because of the small data set in Table 1.

Example 2: What is the time taken by dogs to search a line[1]?

Most demining dogs are trained to search a line, either on short-lead or long-lead. In this example, the amount of time taken on the line was measured using a stopwatch ("time taken").

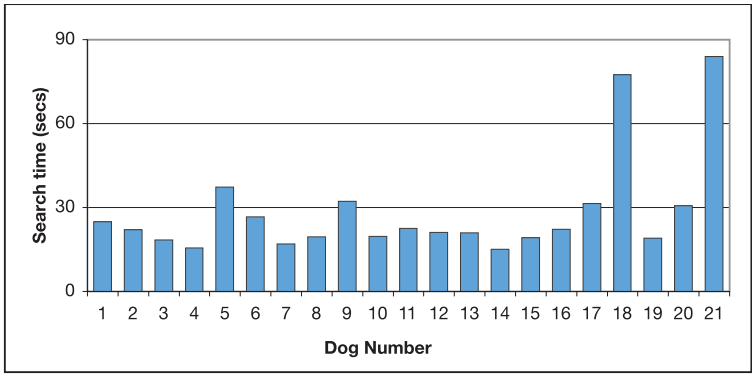
Sampling required records of the start and stop time for dogs at the beginning and end of a search on a line. For each dog, a variable number of samples was obtained and a sample and a measurement are the same thing. Thus, for each dog an average search time could be calculated.

1. This study is reported in GICHD (2005a).

Clearly, there is considerable variability among dogs in terms of the time taken to search a line. However, most dogs used between 15 and 35 seconds. "Dog 21" is a person working with a metal detector, and it is reasonable to conclude that dogs are generally faster than such a person, even without any statistical analysis. Dog 18 used a different search technique, explaining why it was slow.

Figure 5 shows the average time taken across all measures for each dog. Thus the figure could have been drawn with the standard error of the mean on each bar.

Figure 5. Time taken to search a line by mine-detection dogs



Do these data need statistical analysis?

Without more information, there is probably no point in doing statistical analysis, because it is not clear from the data alone why it would be interesting to know that there was statistical (or significant) variability among the dogs. However, if a statistical analysis was done, Dogs 18 and 21 would be removed from the data before the analysis because they are clearly different. Any analysis would then be done on the original raw data from which the means were calculated, and not directly on the means shown in the figure.

For statistical analysis, the question would be:

"Is there significant variation in time taken to search a line by the dogs?"

Note that this is a different question from the broader research question laid out above. The analysis would be a 1-way analysis of variance.

Having reviewed these data, the researcher might now ask why some dogs are faster than other dogs (Qa). Searching faster might not be a good thing, as there is a possibility that fast-searching dogs miss mines (Qb). These two follow-up studies would involve measuring the details of the search procedure used by each dog (to address Qa), and setting up tests where dogs search for mines (to address Qb).

If a second set of data were available which measured dogs' search time on a line using a different technique, then a statistical comparison of the two techniques would almost certainly be desirable and interesting.

Example 3: How much area is cleared by manual deminers using three different drills[2]?

This study involved a comparison of three different demining drills, a Standard drill and two experimental drills (termed Hybrid and Crab). Hybrid and Crab drills both involved working a lane, as for Standard drill, but the lane was placed alongside a safe area, allowing the deminer to step out of the working lane and walk around an indication.

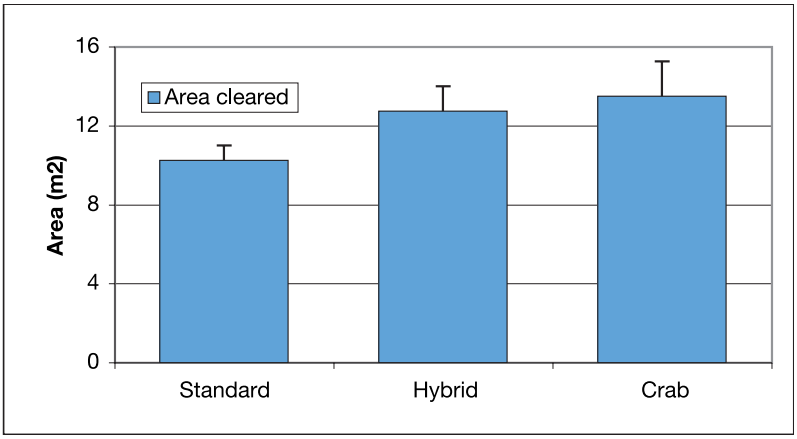
The study was conducted because it was believed that the experimental drills would lead to improved efficiency, indicated by increased amount of land cleared and reduced time spent on activities that slowed the deminer down.

A total of 16 deminers each worked all three drills for 150 minutes of working time in lanes 1 metre wide. The total amount of land cleared was calculated once the 150 minutes weas completed. Thus here, a sample and a measurement are the same thing, and one measurement was made for each deminer.

Figure 6 shows the average area cleared by all 16 deminers (with the variance indicated as standard error of the mean). It appears that on average, the most land was cleared using Crab drill, and least land was cleared using Standard drill.

However, the standard error bars are reasonably large, and it may be that the differences seen here are not statistically different. In this example, a statistical analysis is essential before any conclusion can be drawn about whether Crab is the fastest drill.

Figure 6. Area of land cleared of mines using three drills.
N=16 deminers working for 150 min, and all 16 deminers worked each drill. Bars are mean + standard error.



2. This study is reported in GICHD (2005b).

The statistical analysis showed that Crab was significantly (= statistically) faster than Standard, but Hybrid was not significantly different from either of the other two.

More comments about this Figure can be found in Annex 2, where the technical concept of statistical significance is explained in more detail.

Example 4: How much time do manual deminers spend changing tools when using three different demining drills?

This question was addressed as part of the same study described in Question 3. But here, it is the behaviour of the deminers that was measured. Thus an extra step is required to convert sampled actions to a measurement of time spent doing each action.

While the deminers worked, a scan sampling procedure was used on a 1-minute cycle to record actions of the deminers being observed. Four deminers could be observed from one location, and 150 minutes represented half a day of work. Thus it took 2 days to complete the data gathering for all 16 deminers working each drill, and 6 days in total to record the data for all three drills.

"Change Tool" involved any switching of tools, such as from metal detector to prodder, or prodder to trowel. Tools needed to be put down in a safe place, and retrieving tools sometimes required walking back to the beginning of the lane.

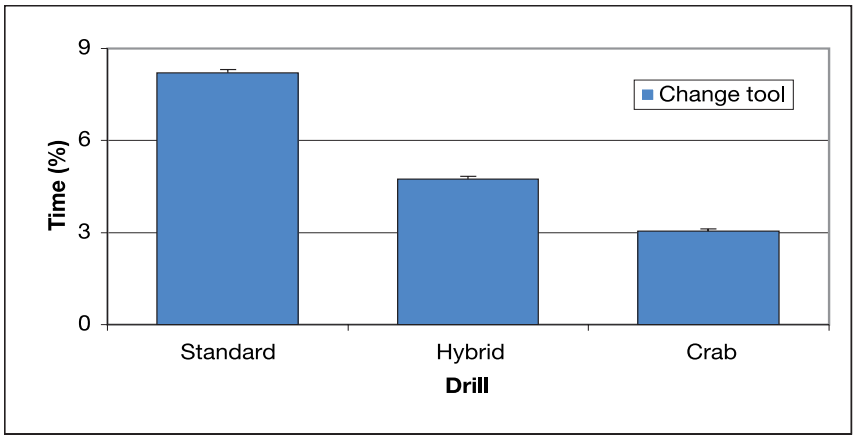
A measurement of time spent Changing Tool was obtained by counting the number of times CT was recorded in the scan samples (for one deminer doing one drill) and converting to a proportion (%) by dividing through by the total of 150 samples obtained.

It appears from *Figure 7* that deminers spent about twice as much time Changing Tools in Standard drill than in the other two drills. However, Hybrid drill may involve even less time spent Changing Tool than Crab drill. The very small standard error bars suggest statistical differences among all three drills.

The statistical analysis confirmed that all three drills were statistically different from one another. It is appropriate to conclude that Crab drill is the most efficient in terms of time spent Changing Tool, with time-saving ratios of about 1.5:1 for Hybrid:Crab, and about 2.5:1 for Standard:Crab.

A demining manager might not be convinced about the difference between 8 per cent and 3.5 per cent, and might not understand or accept results presented with great fanfare as a "statistically significant difference". But they should be convinced by the ratios in the above paragraph!

Figure 7. Time spent changing tools as a proportion (as %) of total time worked by 16 deminers, in relation to different drills.
Bars are mean + standard error.



Example 5. Exploring the data using variance[3]

All of the above examples show counts or means with variances. But it is possible to go beyond these simple statistical measures to explore the data further. In Annex 2, the notion of statistical assessment is described in terms of exploring the relationship between mean and variance.

In the above examples, it was the absolute values of the means (= the height of the bars) that resulted in the apparent differences in the Figures. The reader can also look at the length of the standard error bars to check the variance around the mean.

But the relationship between mean and variance can be explored more directly, and without the distraction of absolute differences between the means. This is achieved by calculating (and plotting) the **mean:variance ratio**. Here are some examples to help with visualising the relationship:

- A mean of 4 with a variance of 8 gives a mean:variance ratio of 0.5.
- A mean of 100 with a variance of 200 gives a mean:variance ratio of 0.5.

In these two examples, the absolute difference between the means (8 versus 100) has been eliminated by conversion to the ratio.

- A mean of 4 with a variance of 2 gives a mean:variance ratio of 2.
- A mean of 100 with a variance of 50 gives a mean:variance ratio of 2.

The higher value for these ratios (2) compared to the examples above (0.5) indicates greater consistency in the second two sets of data.

- And so on.

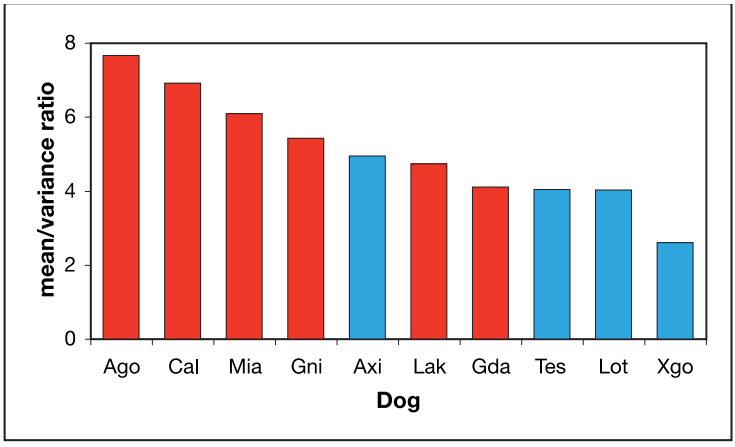
3. This study is reported in GICHD (2005a).

It is easy to see that the relationship is now entirely relative — absolute values have been removed from the result. Now it is possible to explore the relative differences without being distracted by absolute differences between the means.

An example is in *Figure 8*, which shows the consistency of time spent searching a line by two different types of dog (short-lead and long-lead). Most of the red bars are higher than most of the blue bars, indicating that short-lead dogs (red) are more consistent in their line-search behaviour than long-lead dogs.

One long-lead dog (Xgo) was very inconsistent. This male dog was refusing to search some of the time because he was distracted by two females in heat, resulting in the low consistency score. At the time of the study, the demining agency was already aware of the problem and its cause, but had they not been so aware, this graph would have helped them to recognise a problem with one of their dogs. Knowing there is a problem is the first step towards finding a solution.

Figure 8. Mean/variance ratio (calculated as mean/standard deviation for each search period) for line-search times for dogs.
Height of bar provides a relative index of consistency in search behaviour, with higher bars indicating greater consistency. Long-lead dogs = blue, Short-lead dogs = red.



5 Conclusion

There are two levels of value arising from snapshots of demining operations.

- First, any demining organisation can explore its own operational system, test new ideas, and provide objective assessments of ideas suggested to improve productivity.
- Second, the broader community can use a written report of the snapshot to review its own operational systems without having to redo the entire study from scratch.

These potential benefits have been underestimated and even ignored by the demining community, where it is routine for organisations to operate in isolation. More relevant is that much testing is done, but is rarely reported. If the test fails, then it assumed that nobody else will be interested. If the test succeeds, the skills and time for writing a proper report are often not available.

The reality is that a community of exchange can only benefit all parties. Tests that produced a negative result can be just as valuable as successful tests. Unfortunately, there is a tendency to treat a negative result as a failed test. But if the test was properly conducted, it did not fail, and it will be interesting to others.

New ideas are sometimes implemented without proper testing. Worse, is the possibility that new ideas are not implemented at all because no mechanism is in place for considering them properly.

Time-and-motion studies provide that mechanism. It takes time to do a good T&M study, but the number of people involved is small. The benefits in terms of improved efficiency as a result of conducting a convincing analysis are likely to outweigh the costs of the study very quickly.

The GICHD welcomes and encourages requests for advice and support in conducting such studies.

Bibliography

GICHD (2005a)
Mine Detection Dogs: operations, GICHD, Geneva.
GICHD (2005b)
A study of manual mine clearance, GICHD, Geneva.

Annex 1.

Types of scientific measurement

There are four general levels of measurement:

1. **Nominal:** the data consist of names, or labels, and have no order. For example, we might want to measure hair colour. You can categorise the colour of human hair (blond, brown, black, red). A single sample of hair colour tells you very little, except the hair colour of that one person. Many samples are required in order to calculate the proportion of people in a specific population with blond hair, for example. The question "*Do you use PPE correctly?*" is an example of a nominal level of measurement, because the data fall into the categories "yes" or "no", and many samples are required to create a measurement. Obtaining many samples where the data are nominal provides a ratio level of measurement (see below).

2. **Ordinal:** the data have order, but no information on the interval between measurements is available. For example, in a horse race, the winner is ranked 1, the second horse, 2, and so on. These numbers give the order in which the horses finished, but they provide no information about, e.g., the time intervals between horses. Ordinal data would not be used for calculations such as averages. "*In a horse race with 10 horses, the average finishing place was 5*", is clearly a meaningless statement.

3. **Interval:** these data have order and equal intervals between measurements on a scale. However, interval data do not have a true zero. An example of interval data is temperature in Celsius. The zero in Celsius measurements is arbitrary. It does not mean that there is no temperature, for instance, in the same way that a measurement of 0 cm means there is no length. Interval data can be averaged, and subject to most statistical calculations, but they cannot be expressed as ratios. For example, it is not correct to say that 40ºC is twice as hot as 20ºC.

4. **Ratio:** this highest level of measurement is the best for statistical use. Measurements have meaningful intervals between them, and have a

true zero, which means that fractions or ratios can be calculated from ratio data. Examples of ratio level data are: weight, height, length, time (measured in minutes or seconds, not as in time-of-day). A true zero means literally that there is none of the thing present.

Annex 2.

Interpreting statistical analyses

Reports of statistical results use a technical language that is not generally familiar to those reading reports about demining. Thus a short introduction is provided here.

Statistical tests normally compare two or more groups of data. One group of data constitutes a set of measurements of a variable (e.g. proportion of time spent using a metal detector), usually obtained as one measurement per subject. The number of subjects therefore constitutes the sample size (N). The test itself involves applying a mathematical formula to the sets of measurements in order to calculate a ***test statistic*** — a number which represents the variability found within and between the sets of measurements.

In simple terms, if the test statistic is small, that normally means either or both of:

- the variability within each set of measurements is large, and
- the difference between the means is small.

Most people understand a mean (or average), but have more difficulty understanding the concept of variability (or variance) around the mean. Table 1 gives a simple example using data from a study of manual demining in Sudan (the same study as for Questions 2 and 3 above). Two sets of measurements are listed, each giving the proportion of time one deminer (the subject) spent using the metal detector in two drills. Here, the variance is presented as the ***standard deviation*** of the data around the mean. But variance can also be calculated in other ways, and is often presented as the ***standard error*** (as in *Figures 6* and *7* in the Examples section of the Guide). For the purposes of this Annex, the difference between these concepts does not matter.

The means are only slightly different between the two sets of measurements, but the variances are quite different. The reason is easily seen by reviewing the data. In drill 1 (low variance), the measurements range from 8.7 to 17.3. In drill 2 (high variance) the measurements range from 6.0 to 26.7. Just from looking at these data, it is easy to predict that the two sets of measurements will not be statistically different from each other, but that prediction is not made using the rather similar means — it is made by looking at the ranges and variances of the sets of measurements.

Table 1. Two sets of data for seven subjects with means and variances (calculated as standard deviation)

Subject	Use MD, drill 1	Use MD, drill 2
A	15.3	21.3
B	17.3	20.7
C	12.7	6.0
D	13.3	10.0
E	8.7	10.0
F	10.0	26.7
G	15.3	21.3
Mean	12.9	15.8
Variance (s.d.)	**3.2**	**8.2**

When reviewing a set of measurements visually, the range is useful. But statistical tests do not normally use the range in the data. In simple terms, what they estimate is the relationship between the means and the variances. For example, it is quite possible for two means with the values 12.9 and 15.8 to be statistically different — all that is required is that the variances be small (much smaller than in this example). In that case, the ranges of the data would also be much narrower or, put another way, the data would be clustered more closely around the mean.

There is no need to understand the mathematics underlying statistical tests in order to understand the results of a test. The calculations have been subject to a long history of development and testing and are standardised in many computer software packages.

The meaning of "significant"

It is essential to understand the concept of a difference that is "**significant**". This term has a specific technical meaning, and the notion of a statistically significant difference is central to any statistical conclusion.

In essence, ***increasing differences*** between the means, and ***decreasing variances*** around each mean, together imply an increasing likelihood that the two sets of measurements are **significantly different** from each other in statistical terms.

In Table 1, the means of the two sets of measurements were slightly different, but were they different enough to allow a conclusion that the difference was in some sense real? Statistical testing provides an objective mechanism for addressing that question.

The hypothesis being tested here is that drills 1 and 2 are somehow resulting in a different use of metal detectors. In other words, there is something fundamental to drills 1 and 2 that leads to a real (or statistically significant) difference in the way metal detectors are used.

Statistical testing uses a standard rule: if P<0.05, then the conclusion should be drawn that there is a statistically significant difference. P is estimated using the result of the statistical calculation (the test statistic).

P stands for "Probability", and the shorthand P<0.05 can be written out in words as:

the probability of the measured difference being due to chance is less than 1 in 20 (5%, or 0.05).

A probability of less than 1 in 20 is regarded as unlikely enough to support a conclusion that something other than chance factors are at work. The difference between the sets of measurements is real, i.e. is an effect of the different conditions. The notion of "less than one chance in 20" is a standard rule in statistical analysis, and is seen regularly in scientific presentations.

These days, the computer normally reports an exact probability and that probability is then reported as part of the Result, along with the test statistic. Thus a standard statistical report (in this example for a t-test) will be phrased as:

X was significantly bigger than Y (t = 10.9, P=0.004, Table Z).

An enormous amount of useful information is bound up in this simple sentence. But in essence, it simply says that the difference between X and Y can be attributed to something other than chance, and it also gives the direction of difference: X is bigger. It is appropriate therefore to appeal to the different conditions under which X and Y were measured as the likely source (or cause) of that difference. A summary of the data used to make the test can be found in Table Z. Table Z might alternatively have been a graph.

A t-test is the simplest form of an analysis of variance, because only two sets of measurements are compared (as in Table 1). If more than two sets of measurements are available (i.e. more than two conditions are being compared), then a more general test is required: the standard test is analysis of variance (ANOVA). In the Sudan trials, three conditions were compared, so an ANOVA was used to test the data. ANOVA returns an "F" statistic, which is reported along with the result:

There was significant variation among the three conditions, with X being largest and Y smallest (F=7.2, P=0.008).

A P value of 0.008 is lower than the P<0.05 rule, so the appropriate conclusion is that differences among the sets of measurements are due to something other than chance, hence the use of the word "significant" in the sentence.

Where three or more conditions are being compared, the analyst may want to know which pairs of conditions are significantly different from each other. Say the F test gives a significant result and the means are A:2.4, B:5.8 and C:6.3. Just by looking at these means, it seems reasonable to expect that A and C are significantly different, with B intermediate. B might be significantly different from A but it is unlikely to be significantly different from C. This is the situation that arose in *Figure 6* (Example 3, above). The statistical procedure used to assess these pairwise comparisons is called "post-hoc analysis". In *Figure 6*, it turned out that A:C was a significant difference, but A:B and B:C were not significantly different.

Geneva International Centre for
Humanitarian Demining
Centre International de
Déminage Humanitaire - Genève

Geneva International Centre for Humanitarian Demining
7bis, avenue de la Paix
P.O. Box 1300
CH - 1211 Geneva 1
Switzerland
Tel. (41 22) 906 16 60, Fax (41 22) 906 16 90
www.gichd.ch