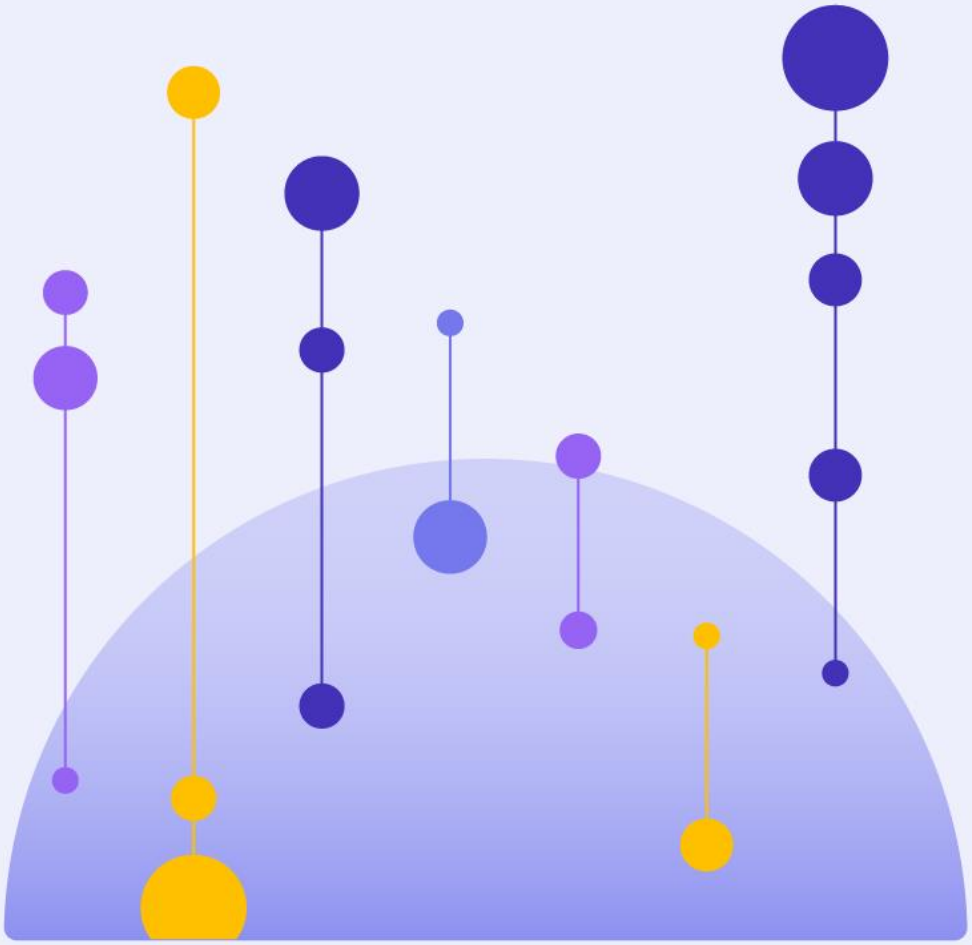
DISPLAYR

RESEARCHSOFTWARE

The Complete Guide to Market Research



displayr.com

Table of Contents

INTRODUCTION 2
Estimate 3
Models… 4
Insights… 6
RESEARCH DESIGN 11
QUALITATIVE RESEARCH 19

QUANTITATIVE SURVEY DATA COLLECTION 31

ADVANCED QUESTIONNAIRES: ATTITUDES, BELIEFS, PERSONALITY AND PRICING 65
ADVANCED QUESTIONNAIRES: NEW PRODUCT DEVELOPMENT 72
DATA PREPARATION 98
BASIC DATA ANALYSIS 133
ADVANCED DATA ANALYSIS 159
REPORTING 179
Q TUTORIAL 189

INTRODUCTION

> Market research is used to figure out how things work. Most commonly, the goal is to understand how a market works, but it can also be used for other problems, such as understanding how people feel about government services and attitudes to social problems. Market research is always a means to an end.

Q Guide to Market Research

Market research is about three things:

1. Estimation
2. Models
3. Insights

Typically, these three things occur as a part of a fairly standard process.

Estimation

We estimate things. What is our market share? How satisfied are our customers? What impact will a 10% price cut have on sales? Second, we explain things. Why do people buy iPhones? What makes customers satisfied? Why did our last product launch bomb? Why do sales increase by 30% when price is cut by 10%? What can we do to rejuvenate our brand?

Lots of market research studies are focused entirely on estimation (although this technical term is not commonly used by practitioners). Market sizing studies work out how many people are in a market. Usage and attitude studies (U&As) measure what people think and do. Governments around the world conduct census studies, estimating the number of people in their countries.

When people think about market research they often think about questionnaires, asking people things like:

```
Thinking about your last flight on Qantas, were you...[] Very satisfied[] Somewhat satisfied[] Neither satisfied nor dissatisfied[] Somewhat dissatisfied[] Very dissatisfied
```

However, there are lots of other ways of estimating facts about markets. *Peoplemeters* sit on top of TVs and count who is watching what at any given time. Point-of-sale scanners in the supermarket measure what has been purchased. Toy manufacturers give children rooms full of new toys and

observe which ones are played with. Some researchers even sort through garbage to find out what people really eat and buy. In focus groups and in-depth interviews, qualitative researchers discuss topics with people and form conclusions about what people think; these conclusions are also estimates (although qualitative researchers would not use this term!).

Estimates can relate to history, such as how many people purchased a product last month. Estimates can be *predictions*, such as the revenue impact of a price increase or the degree of cannibalization that will result when a new product launches. We can even have estimates of things that did not happen, but could have happened, such as the amount of sales that we would have had last year if we had spent an extra million dollars on advertising. Predictions can be quantitative forecasts, such as forecasting that a new ice cream will be purchased by 8% of the market. Predictions can also be more general conclusions about cause and effect, such as concluding that a change in the packaging design of a product will result in an increase in sales.

Estimates are often derided, with good reason, as being mere facts and figures. The real need is often for understanding. Is Microsoft the dominant software brand in the world because it is excellent, because it has blocked its competitors or for some other reason? Why do people buy Coke? What does a new bank have to do in order to encourage customers to switch to it? What can a government do to make people more willing to pay their tax?

Models

The core of all good research is the creation of *models*. Models explain how things work. Any explanation of how a market works is, by definition, a model. We cannot estimate anything without models. Think about an estimate of new sales for a product. The only way we can create such an estimate is by taking into account the factors that determine how many people will buy the product. Less obviously, even a simple estimate, such as the proportion of people that are women, can only be obtained by the application of numerous models (we return to this later).

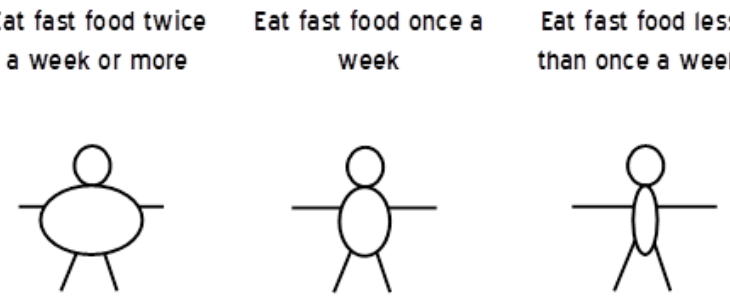
The term 'model' can conjure fear in the minds of the uninitiated. Some people imagine that models are complex sets of equations that can only be understood by statisticians and mathematicians. This is a misunderstanding. The technical meaning of the word 'model' is identical to the understanding you had as a child. Consider a model of Copenhagen's harbor, built out of Lego. This is a model in exactly the same way that we use the term 'model' on this site: it is something designed to look like the real things, but it is not actually the real thing.

A model is a simplification of reality. Good models of markets capture the key aspects of how markets work. Bad models misrepresent reality.

We can build a model out of wood, Lego, metal, clay and any number of other materials. Models can be expressed in words:

```
The amount of advertising determines the level of sales.
```

Models can be combinations of images and words:

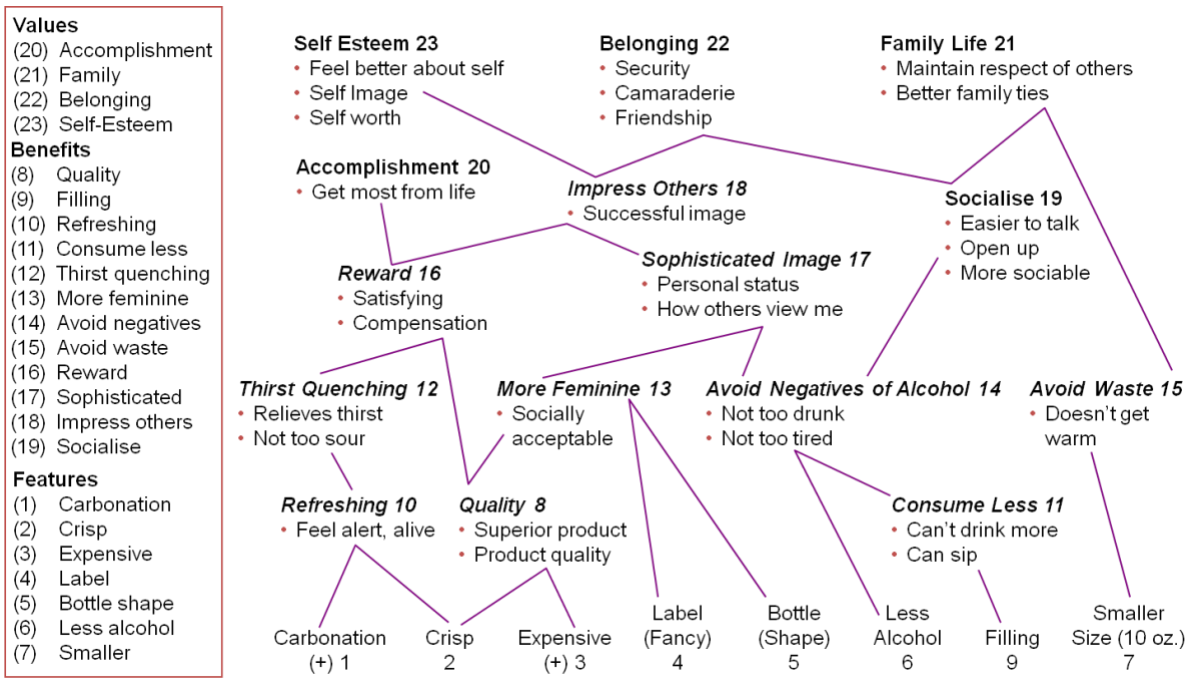


And as equations:

```
$Sales = 121 + 4.1 × $Advertising expenditure
```

Each of these examples of models differs in their precision. The equation is generally the most precise type of model (although the precision will be misleading if the equation is wrong). Models also

differ markedly in terms of their complexity. The models we have seen so far are all simple, involving only a few variables. In real-world applications much more complex models are often required, such as the model shown below, which is discussed in Laddering).



Insight

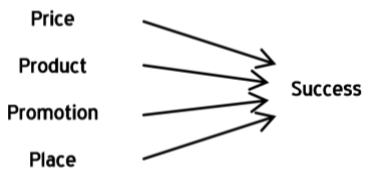
Although models are at the heart of research, the term model is a technical term and few people doing market research use the term when describing what they are doing. Sometimes people will instead talk about theories. A theory is just another name for a model.

Sometimes people talk about *hypotheses.* A hypothesis is a tentative part of a theory. Often research is conducted to check a particular hypothesis. By checking and refining a model, we better understand how a market works. For example, if somebody says "we have a hypothesis that we need to reduce the price of our product to below $4 in order for people to notice our product", they are saying that a stakeholder believes that this may be the case and they want to check if it is true.

Similarly, the hypothesis that "price is more important than brand" is a speculation about the strength of the relationship between variables in a model (i.e., they are speculating about estimates).

Most often, users of research talk about wanting to conduct research so that they can find insights. An insight is a conclusion about how a market works that is useful. As discussed, a model describes how a market works, so an insight is a description of a model, or a part of a model. However, to qualify as an insight, the conclusions about how a market works need to be useful. Thus, an insight is a conclusion about how a market works which results in a company doing something that they would not have otherwise done. Research that does not lead to some improved outcome, such as a better strategy or policy, has thus, by definition, failed to produce insight.

Consider the 4P model of marketing, which says that the success of a product is determined by price, product, promotion and place:



This is a popular model in introductory marketing courses (particularly once 'promotion' is defined to include advertising). Many students in introductory courses know little about the subject, and as a result, find this model useful. For some, this simple framework is an insight. However, this model would never be a source of insight to an experienced marketer.

To qualify as an insight, we need to have something that is a statement about how a market works which is non-obvious to the end-user of the research. A conclusion like “if we advertise at the same time as offering discounts, we get twice as many sales as when we offer discounts without advertising” may qualify as an insight, provided that this is correct and not already known by the end-user.

Throughout this site we will discuss lots of quite technical aspects of research. However, it is important to never forget that they are just means to an end, and the end is insight with potential to create competitive advantage. Research that provides no insight is useless. The challenge of obtaining insight is at the heart of the tension that exists between research consultants, and their clients, as illustrated in this joke:

A man flying in a hot air balloon suddenly realizes he's lost. He reduces height and spots a man down below. He lowers the balloon further and shouts to get directions, “Excuse me, can you tell me where I am?”

The man below says: “Yes, you're in a hot air balloon, hovering 30 feet above this field.”

“You must work in market research,” says the balloonist. “I do” replies the man. “How did you know?” “Well,” says the balloonist, “everything you have told me is technically correct, but it's of no use to anyone.”

The man below replies, “You must work in marketing.” “I do” replies the balloonist, “But how'd you know?” “Well”, says the man, “you don't know where you are, or where you're going, but you expect me to be able to help. You're in the same position you were before we met, but now it's my fault.”

How to find insights

The best chance of obtaining insights from research is to:

- Create models.
- Use the right type of data.
- Work out what is known before attempting any research.

Create models

Models need to be created. Simply reporting results (i.e., estimates) without attempting to create models guarantees that no insights can occur.

Use appropriate data

The data that is used needs to be capable of containing insights. Although it is certainly the case that some people are more gifted than others at extracting insights from data, it is also true that some forms of data are more useful in helping to find data than others.

Remembering that an insight is a new conclusion about how the world works, it follows that our best chance of finding insights is to use data that we have not previously examined. The more varied the data we examine, the better our chance of finding insight. If we have gained most of our

understanding about how a market works by analyzing sales data, when trying to crack a difficult problem, we should focus on using some completely different type of data (e.g., qualitative research, surveys). Similarly, if a firm has typically used qualitative research, the likelihood is that it will gain better results from a quantitative survey. The interesting result in the first row is Animal which, for Dog, sits at 100. But, the residual is only 5, indicating virtually no association between being an animal and being a dog. Why? All rows of the data are animals (and four, like the Dog, mammals). So, while a Dog is an animal, this does not distinguish it from all the other things in the analysis. As a result, this association becomes very weak, which is what is reflected in the residuals.

Identify what is known before doing any research

If we do not know what models are currently being used to make decisions, it is extremely difficult to design research which will uncover insights. When we analyze data, there are a thousand and one results that may appear to be interesting. As most people do not have the patience nor the capacity to review thousands of findings, researchers focus on presenting the findings that seem to best explain how the market works. If these are already known, it is almost inevitable that nothing ends up being found out from the research.

Unfortunately, many people who commission research fail to understand this last point and, when asked for their theories and hypotheses, say things like "I want the research to be objective ... I don't want to taint the research with my prejudices … you just do the research, I'll work out it means." However, these people fail to recognize that in doing so they are maximizing their chance of getting useless research (in other words, finding out they are in the hot air balloon).

Standard research process

The essence of market research is about creating models and estimating things. Commercial market research usually takes place in the following process:

Briefing: the client articulates what they want to know (i.e., their objectives for the research). Most of the time, they will either be conducting research to gain one or more of specific estimates (e.g., market share, advertising recall, predictions) or insights. This briefing can be done verbally or in writing; the written document is called a *brief* in England and Australia and a *request for proposal* (or, *RFP*) in the US.

Project scoping: the researcher works out what the client needs. The project scoping needs to find one or more of: all the estimates the client requires; any hypotheses or theories of the client, and the behavior(s) in the market that the client wishes to change (most commonly, brand choice).

Research design: the plan by which estimates are going to be computed and models refined. Research designs are sometimes referred to as research methodologies. The 6Ws is a good aid memoir for making sure that all the detail of the research design has been thought through:

Who should be interviewed? What information should be obtained? When should the data be collected? Where should the data be collected? Why should the data be collected? What way should it be collected?

Proposal: this describes the research design to the client, along with pricing and timings. It is primarily a sales document. It is typically about six to twelve pages in length, but it can be as short as a one-line email or as long as several hundred pages.

Instrument design: questionnaires are written for quantitative research, data capture specifications for observational studies and discussion guides for focus groups and depths (in-depth interviews). See Basic Questionnaires and Advanced Questions and Questionnaires for more information.

Data collection: interviewing and/or observation (e.g., recording where people go in a shopping center). This will typically involve qualitative research, conducting a quantitative survey with a questionnaire or a both of these activities.

Data processing: erroneous data is identified and either deleted or corrected (cleaned), and text data is turned into categorical data (coded).

Weighting: the distribution of the sample on key variables is compared to what is believed to be the correct distribution; if the distributions differ, a weight is computed to correct the sample (this is discussed in more detail in the next chapter).

Analysis: estimates are computed and models are refined. This can take many months, in the case of large strategic pieces and government reports, or may be as limited as a bit of reflection in the shower for a fast turnaround qualitative study. This is discussed in more detail in Basic Data Analysis and Advanced Data Analysis

Reporting: sharing the results with the client. Most of the time this is done with a presentation and a PowerPoint document, and perhaps an online dashboard.

Research Design

A research design is a blueprint describing how to conduct a research project. It is a plan describing which estimates are to be computed, how they are to be computed and how models are to be tested and refined. A good research design is one that identifies all the things that need to be estimated and works out the best way to measure them.

- Information Needs
- Data Collection Plan
- Analysis Plan

Research Design

Information Needs

The starting point for creating a research design should be an understanding of which decisions need to be made. Examples of the type of decisions that research is used to address are:

- Should a pack of chewing gum change from 10 to 12 pieces?
- Should a vegetable juice replace its glass bottles with plastic bottles?
- Should an airline replace its economy seats with stand-up beds?
- Who is the best Republican candidate for President?
- How many phone plans should a company offer?
- Which of 10 new product ideas should be further investigated?
- Should a company replace its American call centers with call centers in India?

It is rare that research can actually provide a definitive answer to such questions. Rather, it provides data for a business case. For example, research can provide an estimate of the impact of sales that occurs when a pack of gum is increased by two pieces, and this is combined with the company's cost data to work to create a *business case* for the change in pack size.

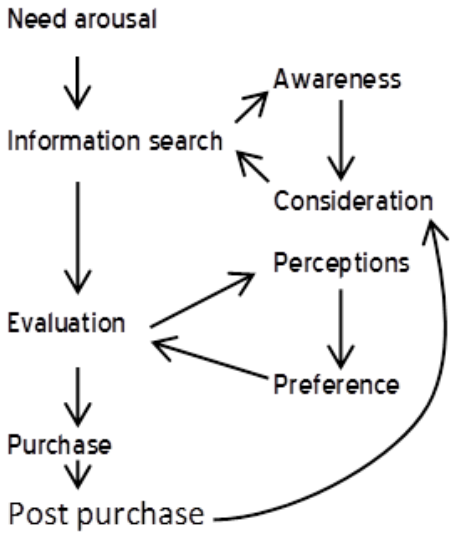
A key thing to keep in mind when working out which business decisions need to be addressed by research is that the more precisely a firm can identify the key decisions, the more precise the resulting estimates will be. A firm that is able to specify that it wants to understand the impact on sales of changing its pack of gum from 10 to 12 pieces has a good chance of conducting research that is sufficiently accurate to help make this decision. However, if the same firm says that it wants to "understand the consumer value equation for gum, in terms of how consumers trade off price, pack size, multi-packs, flavors and health benefits" it will end up with a much less precise estimate of the impact of a change in the number of pieces and thus risks making worse business decisions.[note 1]

Once the relevant business decisions have been identified and a list of things to be estimated has been created, the next step is to create any appropriate models, as these will lead to the identification of further estimates that are required. For example, the figure below shows a model and information needs for a banking study aimed at increasing acquisition and retention. Where the model has not

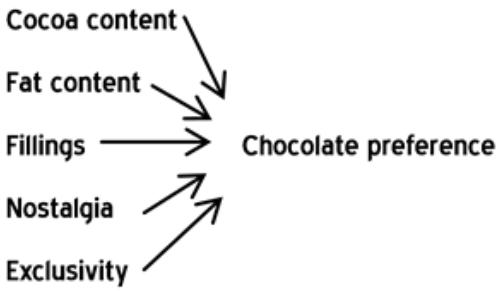
been proposed by the client (which is usually the case), the researcher then needs to create or find useful models. Usually, one or more of three different types of models are required: models of behavior change, measurement models, and decompositions.

Models of Behavior Change

Research is often desired when a problem has been identified but there is no clear idea of how to solve the problem. For example, a company may wish to identify opportunities for increasing sales. Thus, the behavior to be changed is brand choice, and a model is required which explains what causes this behavior. Although it is possible to conduct qualitative research to try and create such a model from scratch, often models which already exist can be used. For example, when trying to understand how to increase market share, it is common to use a model such as this:[1]



Alternatively, the model may be something much more specific:



Whatever model is used needs to be able to accommodate the theories and hypotheses that are held by the users of the research. And, it should take into account which outcomes the client is trying to affect (e.g., if the the business objective is to try and reduce the level of defection of its customers, then the model should mention defection as an outcome and have arrows pointing to this outcome).

Experienced researchers often do not write their model down, instead creating it in their heads. Inexperienced researchers often do not know they need to create a model, and instead design their research by repeating aspects of projects they have seen or previously worked upon. Nevertheless, their research is still structured around models. As mentioned, it is impossible to do research without a model. When you do not explicitly create a model, you risk implicitly using an inappropriate model as the basis of your research design.

Decompositions

A decomposition involves breaking something apart. It is a useful way to estimate many things. Let us start with a simple estimation problem. How many Japanese people are there in Australia with dentures (fake teeth that can be taken in and out)? And yes, this was was a real-world consulting project.

The simplest and laziest approach to this problem is to do a survey. For example, you might email 1,000 people in Australia and ask how many people in each household are Japanese and have dentures. The result of this would be an estimate of the proportion of Australians that say they are Japanese and have dentures. To get to our required result we would multiply this by the number of Australians. Thus, we have used the following model:

```
Number ProportionJapanese = Japanese × Numberdentures dentures Australians
```

This formula is a decomposition. We have decomposed the thing we are trying to estimate into two separate estimates - the proportion and the population size.

As research designs go, this is a poor one. The proportion we are trying to estimate is likely to be a very small one (e.g., less than 0.1%). Thus, we would expect that we might need to interview well over 1,000 people before we identified a single one of them and perhaps 50,000 or 100,000 people before we obtained a sufficiently precise estimate, for a cost of millions dollars. People researching bizarre topics are usually short of funds, so we can feel confident that this research design is inappropriate. Furthermore, we probably would not get a very precise answer anyway, as Japanese speakers are relatively less likely to participate. So, how can we resolve this? There are lots of other possible decompositions. For example:

```
Number Number Japanese = people with × ProportonDentures dentures Japanese
```

This decomposition requires two completely different inputs: the number of people with dentures and the proportion of people that are Japanese. Both of these numbers may be available from publicly available sources (e.g., trade associations, government statistics), so we might be able to do this very cheaply. Of course, the result may also be quite inaccurate, because this decomposition implicitly assumes that people of Japanese origin are neither more nor less likely to have dentures than the rest of the population.

So, we can use different decompositions to solve the same problem. The trick is to trade-off which will be cheapest with which will be most precise.

Now, let's solve a more traditional problem. Consider the problem of trying to predict sales of a new brand of laundry detergent. A standard decomposition for this is:

```
Sales = Market Share × Market Size
```

There are lots and lots of ways to estimate market share. One of them is to present people with a screen showing a picture of a supermarket shelf, including the new brand, and ask people to choose one; the proportion of people who choose the new brand is then an estimate of the market share. The market size can usually be obtained from historic sales data.

An alternative decomposition of sales of a new product is:

```
Intention Sales = to × Purchase × Population purchase frequency size
```

Intention to purchase is estimated by showing somebody a picture of the proposed new product, and asking them whether they will buy it (this is called Concept Testing). Purchase frequency can be estimated by asking people how many times they will buy it (although looking at historical purchase rates of similar products will often be more valuable). The population size can be obtained from government statistical agencies.

When designing research, we need to find the most cost-effective way of producing sufficiently precise estimates. Consider the decomposition of:

```
Sales = Market Share × Market Size
```

Let us say we are trying to forecast laundry detergent and we are trying to produce a forecast for next year. To produce our sales estimate we need to estimate market share and market size. It is inevitable that the market size will be broadly similar to the sales from the previous year, with a little growth. So, if the market size last year was $13 billion, the market size next year will probably be between $13 billion and $14 billion.

Now think about the market share estimate. If the new product is a 'dog', it will get 0% market share. If it is wonderful, perhaps it can get 20% of the market. So, if we multiply the lower bound estimates of market size and market share we compute a lower bound estimate sales of 0% of $13 billion = $0 and an upper bound of 20% of $14 billion = $2.8 billion. Our range of estimates for market size make comparatively little difference to our forecast. If we assume that the market size is $13 billion, this drops the upper bound from $2.8 billion to $2.6 billion. By contrast, changing the estimated market share from 20% to 0% drops the upper bound all the way down to $0. Thus, the estimate of sales is most sensitive to the estimated market share analysis. This process, of working out which bit of the

research design will most impact upon the precision of an estimate, is known as sensitivity analysis. It follows, from this sensitivity analysis, that if we are decomposing sales into market share and market size, our focus in producing an estimate of sales should be on estimating the market share.

Measurement Models

If you have ever watched House or used alternative media, you will be familiar with the idea that people do not always tell us the truth. Expressed in a slightly more academic way: what people say is a mixture of the truth and error:

```
Observed = Truth + Error
```

This model is known as a *measurement model*. And depending on context, it is often described with slightly different terms (e.g., using the terms 'Measured', 'Claimed', 'Manifest' or 'Estimate' in place of 'Observed'). Measurement models are most commonly used in more academic studies. More detail on measurements models is in Measuring Abstract Concepts.

Data Collection Plan

A data collection plan is a document that illustrates the who, what, when, where, why, and how of data collection. This is covered in more detail in the data collection section of this guide.

Analysis Plan

An analysis plan is a list of all the intended analyses that the study will need to address. For example, if the purpose of the study is to work out whether a new product will be successful, then the analysis plan will indicate which analyses will be conducted to answer this question.

Different types of analysis are discussed in the Basic Data Analysis and Advanced Data Analysis sections.

Common things that are usually listed in an analysis plan

1. Any particular steps required for data preparation, such as new variables to create, NETs, and weighting.
2. The key sub-groups that should be used when exploring the answers to all the questions (e.g., age, gender, heavy vs light buyers).
3. Specific analyses that are required (e.g., specific hypotheses to be tested).

Example

The following analysis plan is for a simple Concept Test:

- Create a *top two box* NET on the purchase intent question.
- Compare the top two box score with top two box scores from similar studies.
- Conduct a Segmentation using the attitude questions.
- Compare top two box scores by segment.
- Code the score's reasons for liking and disliking the product using the same code frame.
- Crosstab all the questions by:
    - Age: Under 30; 30 to 49; 50+
    - Gender: Male, Female
    - Income: Under $100,000; $100,000 and above

Qualitative Research

- What is Qualitative Research?
- In-depth interviews and focus groups
- Moderators and observers
- Analysis
- The subconscious
- Sample Sizes for Qualitative Research
- Discussion Guide
- Techniques

Qualitative Research

What is Qualitative Research?

Qualitative research is a catch-all term which refers to all forms of research that do not involve storing data in data files. Most qualitative research, however, involves a combination of:

- Crude quantitative research, where the researcher has insufficient knowledge to conduct a quantitative study so instead conducts a discussion, observes behavior or does something else in an attempt to gain more knowledge. Such research is often a precursor to a quantitative study.
- Probing into underlying needs and wants that are unable to be discovered using quantitative research.

In the US and Asia the focus tends to be more on the first aspect, whereas in the UK, Europe and Australia the focus is more on the second.

In-depth interviews and focus groups

The vast majority of commercial qualitative research involves interviews. A discussion guide is used to structure the interview. Where there is one interviewer and one participant the interview it is called a 'depth', which is short for 'in-depth interview'. As the name suggests, such interviews are typically conducted when it is believed that a detailed discussion is required. Such interviews typically go for about 45 minutes, but this can vary.

Where there is a desire to involve participants in discussion among themselves, or to achieve economies of scale, multiple participants may be involved. Where the interview involves two

participants it is a called a 'paired depth'. With three to five participants the interview is sometimes known as a 'mini-group'.

A 'group', which is short-hand for a 'focus group' and 'group discussion', is an interview which typically involves one interviewer and from four to sixteen respondents. Six-to-eight is typical when the focus is on probing. Larger group sizes are sometimes used when conducting a crude form of quantitative research.

Moderators and observers

The person that conducts the interview in a group discussion is known as the 'moderator'. Where the qualitative research is focused on probing to understand underlying needs and wants, the moderator is typically also the chief person responsible for observing and interpreting the discussions. Where the study is more of a crude quantitative study, it is common that the primary responsibility for observing and interpreting the discussions resides with an observer who is watching either by one-way mirror or a video-link.

Analysis

There is no agreement among practitioners about how best to analyze qualitative research data. Common methods include:

- Reflecting upon the interviews and drawing a conclusion (i.e., no formal process).
- Reading through transcripts and identifying salient quotes.
- Coding, sometimes with the aid of computer programs developed specifically for this purpose.

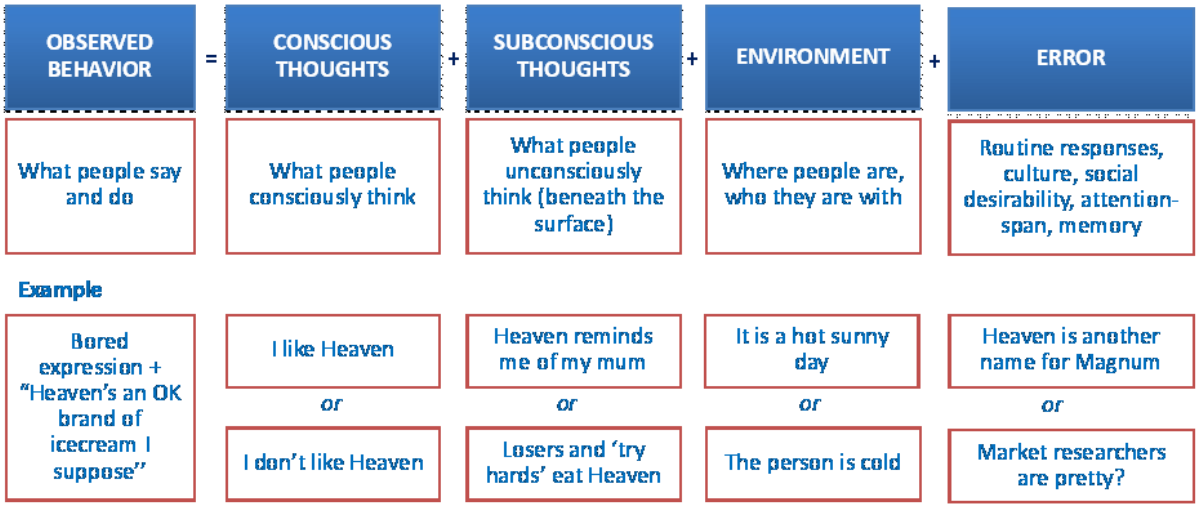
The subconscious

There are many situations in which people are unable to answer the questions of interest. For just about anything that people eat and drink, if people are asked why they buy what they buy, they say that it is because they "like it". Although this answer is obviously true, it is incomplete. Consider, for example, choice of brands of beer and cola. When blind taste tests are conducted, people often prefer brands different to those they regularly buy, but when the taste tests are done with the labels on packaging, people invariably say the brand they normally buy, or aspire to buy, tastes best.

A basic framework for trying to disentangle motivations is view behavior - where an initial answer to a question is an example of a behavior - as being explained by a combination of:

- Conscious thought.
- Subconscious thought.
- The environment.
- Error (i.e., random factors which do not really have any meaning).

Two examples are in the following diagram:



Sample Sizes for Qualitative Research

Sampling for qualitative research is completely different and it is not unusual for qualitative studies to have tiny samples. To understand the impact of sample size on qualitative studies we will briefly pretend they are quantitative (see Determining The Sample Size for a discussion of sample sizes for quantitative surveys). Imagine a study where we are evaluating two new products, and the first of these products is a clear winner with 75% of people preferring it. With a single focus group of eight people we can compute the confidence interval as 35% to 97%. If we conduct two focus groups the confidence interval becomes 48% to 93%, so we still cannot even be sure that the product that was the clear winner is actually most preferred. With four focus groups, and a sample of 32, the confidence interval is 57% to 89%.

The true confidence intervals for qualitative research are actually much larger than this, as researchers rarely count the number of people to prefer a concept, instead relying upon their perceptions of the overall vibe of respondents, which adds more noise. Further, the samples that are recruited for qualitative research can be quite unrepresentative, which adds yet more uncertainty to any result.

The poor precision of qualitative studies means that a greater level of discipline is required when interpreting qualitative research. Qualitative studies should not be used for estimation. That is, they should not be used to pick winning concepts or to understand what people like (which are both estimation questions). Qualitative studies can, however, identify bad ideas: if nobody likes a concept then it is very unlikely to become popular with a bigger sample.

As discussed in the introduction, market research involves estimation and explanation. What qualitative research lacks in rigor for estimation it makes up for in its improved ability to explain things (i.e., to create new models). Of course, estimation and explanation are not independent and a degree of caution is required when obtaining explanations from qualitative studies as the problems of precision make it impossible for there to be much in the way of certainty about the accuracy of any models derived from qualitative research. For this reason, it is common to use quantitative research to confirm any important findings from qualitative research (although many firms rely solely on qualitative research when conducting branding and advertising research, as often quantative methods is seen as being poor for such research).

Discussion Guide

A list of topics to be discussed during. In some circumstances this can be almost as detailed as a questionnaire. In other examples the discussion guide may be a few dot points.

Health care example

I. Welcome and introductions
II. Current health status and medical problems
III. When do you go see a doctor?
IV. Relationship with health care providers
V. Kinds of contact/time spent
VI. Difficulty in seeing doctor/making appointments
VII. Cost/ease in getting to provider
VIII. Physical examinations
IX. Does doctor explain what is happening?
X. Can you influence what tests are performed?
XI. How are results communicated?
XII. How is the decision made to get a physical exam?

Techniques

Whereas in quantitative research extensive attention is paid to creation of the questionnaire, in qualitative interviews effort is put into the selection of appropriate techniques to permit detailed exploration of the points in a discussion guide.

Techniques commonly used include;

Probing

Probing involves an interviewer asking follow-up questions designed to obtain more and deeper information. It is conducted in a conversational style. Some people believe that probing can elicit people's sub-conscious motivations. An example of probing is:[1]

Respondent: When I bought my new Mazda, I felt, “What a bargain!”

Interviewer: I hear what you thought, but what did you feel?

Respondent: Well, I felt that the price was fine.

Interviewer: And what did you feel?

Respondent: You know, I felt a little cheap.

Sorting

Sorting, also known as card sorting, simply involves asking participants to sort objects (e.g., brands) according to which are most similar and to explain what makes different objects similar and dissimilar.

Laddering

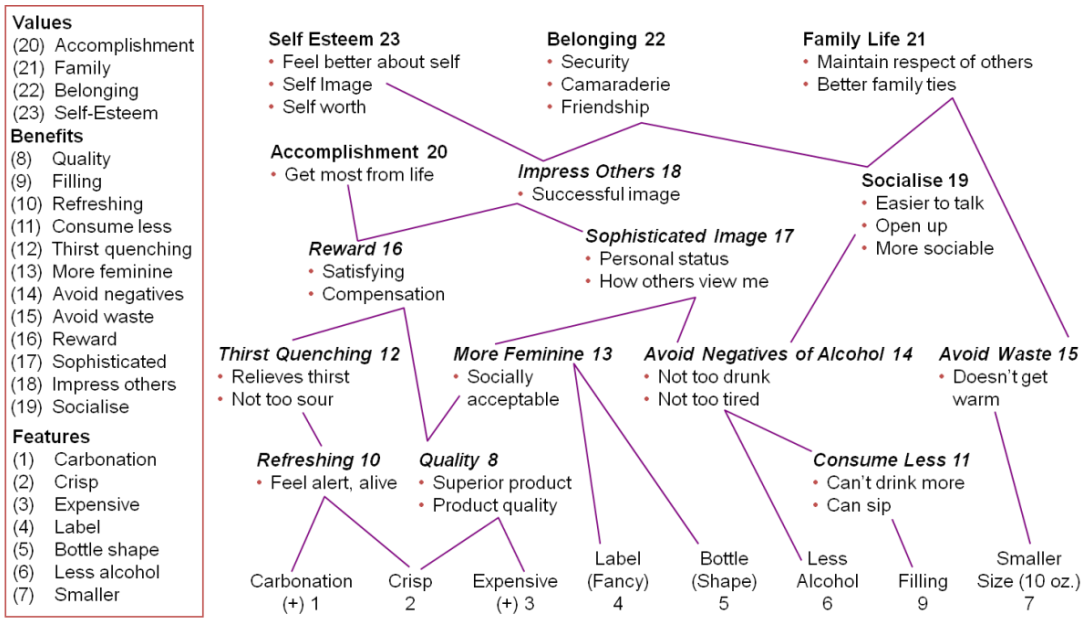
Laddering is a type of probing which is based around the idea that values determine the benefits that we seek, which in turn determines which features we seek when buying products: Values -> Benefits sought -> Features

Example

Moderator: Why do you like Lemon Ruski?
Jenny: I like it because it’s less alcoholic than having a straight spirit
Moderator: What makes that important to you?
Jenny: It stops me getting too drunk and tired
Moderator: What benefit does this provide to you?
Jenny: Otherwise I might look like a slag – it’s important to appear sophisticated
Moderator: Sophisticated?
Jenny: Well... I like to impress others. It’s good for my self-esteem and it’s important to have the respect of your friends

Hierarchical value map

A key output that can be created by laddering is a *hiearchical value map*. The example shown below is from the wine cooler market (the precursor to today's alcopop and RTD markets).[1] The basic idea behind such research is that consumers' demand for particular products is driven by their basic value system. We start from the bottom of the ladder, which lists the product features of the products, link these to benefits which are, in turn, linked to values.



Projective Techniques

Projective techniques seek to get to the subconscious. They work as follows:

1. Participants are asked to project their feelings and thoughts onto other things.
   For example: If Coca-Cola was an animal, which animal would it be?
2. Participants are then asked to explain their answers. This 'why' question is the important part of using projective techniques, as the projective techniques are designed to release the sub-conscious though rather than to be, in themselves, revealing. Probing is used to try and uncover the real explanations. For example, if a Coca-Cola was seen as a cow, the explanation may be that the respondent sees it as fat, slow moving and uninspiring.

Projective techniques are fun. They are widely used, with clients, respondents and researchers all finding them a welcome change from the humdrum of traditional market research questions. However, considerable care needs to be used when asking such questions. As with measuring abstract things, our inability to phrase a perfect question causes confusion between truth and error. Measuring the subconscious is obviously much more difficult than measuring something abstract, as the problem of abstraction is that it is hard to find the right words, whereas the problem with the subconscious is that people do not even think about it.

Projective techniques were developed in psychology. Concerns about their reliability have caused them to be, by and large, the provenance of 'alternative' psychology practitioners. In market research they remain mainstream, as other techniques often fail to generate any insight at all into brand buying. Of course, being perceived as useful does not solve the problem of validity. Three things are done to prevent spurious conclusions being reached using projective techniques:

1. Probing is used to confirm any interpretations.
2. Multiple techniques are used and results are concluded to be reliable when the same conclusions are drawn from multiple techniques (which is equivalent to the approach to abstract things, where multiple questions are asked).
3. Combining the data of multiple people (although almost all qualitative research makes this assumption it is, somewhat inexplicably, rarely mentioned). If the same responses are obtained from lots of people, then it follows that the responses likely contain a kernel of truth. From a theoretical perspective, combining the answers of multiple people to one question is equivalent to getting one person to answer multiple questions if it is assumed that the differences between the people are random and the differences between how a person answers multiple questions are also random.

Sentence Completion

Sentence completion is a *projective technique* that involves asking participants to complete sentences. For example:[2]

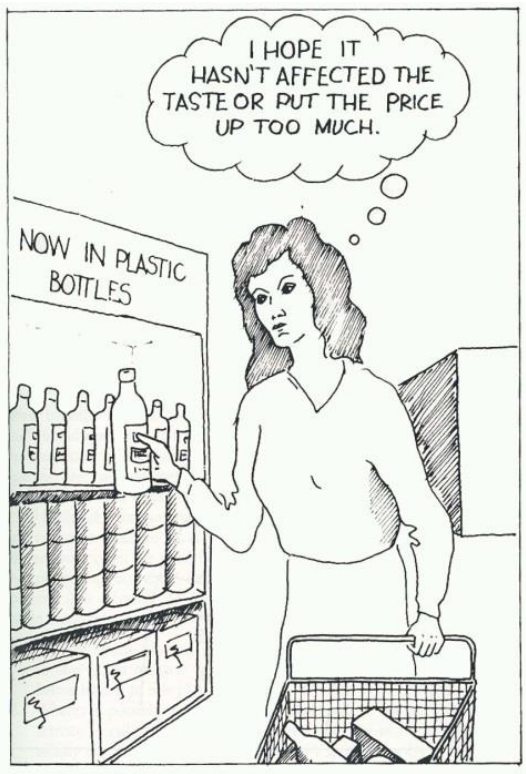
People who buy Coke are _____I only buy Diet Coke when _____If my wife brought home a case of VB I'd say _____People who are concerned out additives and

preservatives in food are _____I think women who buy frozen pastry are probably_____

They are then asked to explain why they have completed the questions in the way that they have completed them.

Cartoon Completion

Cartoon completion is a Projective Technique which asks participants to write in the bubble of a cartoon:[1] and then to explain what they have written.



Stereotyping

Stereotyping is a Projective Technique that involves presenting a description of people, either using words or images, and asking people questions about the topic of interests. For example:

- What soft-drinks would these families have in the house?
- What brand of bubble gum is she chewing?

This is then followed up with a probing question in which the participants are asked to explain their answers (e.g., `What makes you think that?`)

Example word descriptions

Family A: Both are about 30 years of age with two young children and are very much concerned about owning things. They desperately want a better lifestyle, so he works hard, often for overtime pay, in order to be able to achieve this. She is planning to return to work, part-time, as soon as possible. They spend time furnishing and decorating their home as they are very house proud. When it comes to buying things, they are more influenced by looks than anything else. They watch TV and read an average amount and tend to follow trends once they have already become fashionable. They think of themselves as modern and up-to-date, but are really quite conventional with regard to family life. They fall into the medium income bracket.

Family B: Husband and wife are in their early fifties with one married son and a 19-year old-daughter. Mr B. works at a boring job and is grateful for the money, which pays for the basic necessities of life. Mrs B does not work, so they do not do much in their leisure time because they cannot afford it. He enjoys football and gambling; she bingo. They watch TV a great deal. They feel that life is hard nowadays, and owe a lot of money. They buy on price – the cheapest possible. Their attitude to their family is traditional, and so are their political and social views. They are climate sceptics.[3]

Brand Personification

Brand personification is a Projective Technique that asks people to think about brands as if they were people and to describe how the brands would think and feel. For example:

`If these three brands were at a party, what would they be saying to each other?`

Brand personification questions are followed up by questions designed to understand the personification: `Why do you say that?`

Alternative versions of this method ask for the brand to be described as a person, animal, car, celebrity or house. Some researchers ask respondents to create collages of photos to represent the brand, or, present people with a set of photos and ask which photo shows a person most likely to buy the brand.

Some researchers use versions of the question where respondents are shown photos of the same or a similar-looking person with different hairstyles and clothes (e.g., one as a yuppie, one has a housewife), and these photos are believed to represent archetypal personality types. There is no published evidence to suggest that these methods work, but it is unclear if this is because they do not work, or because the researchers who use these questioning methods wish to protect their intellectual property.

1. Sudman, Seymour and Edward Blair (1998), Marketing Research: A Problem-Solving Approach. New York: McGraw Hill, p. 187.

2. Adapted from Gordon, Wendy and Roy Langmaid (1988), Qualitative Market Research: A Practitioner's and Buyer's Guide. Aldershot: Gower, p. 105.

Quantitative Survey Data Collection

- Defining the Population
- Selecting a Sample
- Deciding How to Conduct the Interviews
- Determining the Sample Size
- Setting Quotas
- Code Frame
- Fielding the Survey
- Programming and Testing
- Data Collection Software

Quantitative Survey Data Collection

Defining the Population

A sample in a market research study is drawn from a population (sometimes called a universe). For example, when conducting political polls, our population is the people that can vote in the next election. When conducting studies of fast moving consumer goods (e.g., breakfast cereal, laundry detergent), our population is grocery buyers.

The unit

Defining the population for a study involves two separate decisions. The first decision is what unit to study. The unit is the 'what' that is to be counted. Commonly-used units include people, consumers, users, households, main grocery buyers, businesses, business decision makers and voters.

The boundaries

The second decision is to define the boundaries of the population: which units are included and which are excluded. Most obviously this is done via geography, where a decision is made about which countries (or states within countries) should be included in the research. Typically, boundaries are also defined based on behavior, such as consumption of a particular product. Examples of population definitions for studies include:

- Main Grocery Buyers that have purchased instant coffee in the past four weeks in Australia.
- People that will be eligible to vote in the next US Presidential Election.

The classic mistake

A common mistake when defining population is to over-define the population such that it is too small. One chip manufacturer undertook a study where the market was defined as “Males aged 25 to 34 that have consumed premium chips in the past two weeks”. When it was suggested that the market was too narrow to be sensible, the brand manager (male in late 20s) indicated that this was not a problem as the age group was “aspirational”. The reason that such a narrowly defined market is problematic is

that inevitably most of the people who are really in the market for the product are not represented by the study and thus the study likely becomes biased.

There are some simple rules of thumb that can be used to see if a population has been too narrowly defined:

- If the biggest selling products in the market are getting more than 20% of their sales from people outside the defined market, the definition is likely too narrow.
- If the definition involves any demographic variables (e.g., age, gender, income), it is probably a Target Market rather than a sensible population definition, unless the demographics are causally related to aspects of people's physiology or life circumstances (e.g., if the study is about retirement, then age is causally related to this, whereas if the study is about chip buying, age may be correlated but is not causally related).
- The definition must be operational. For example, rather than “women capable of bearing children” a more operational definition is “women aged 12 to 50”.

Selecting a Sample

Once the population has been defined, we need a way of finding the people in the population. To do this we need either a list of people in the population, or, a list of places where we can find people.[1]

Surveys involve either selecting everybody from a list or at a specific location (which is known as a census), or, sampling, which involves randomly selecting people from lists or locations.

Lists

Common examples of lists include:

- Online panels, which are databases of people willing to participate in market research in return for pay or prizes.
- Telephone directories.
- Delivery Point IDs, which are lists of all the postal addresses in Australia.

- Lists of customers' email addresses.
- Lists of people willing to participate in focus groups and depth interviews (these lists are held by recruiters).

Such lists are sometimes referred to as the sample. However, it is important not to confuse this with use of the word sample with its normal meaning (where the normal meaning is respondents that have completed the questionnaire).

Places

Common examples of lists of places where we can find people include:

- Shopping centres.
- Websites' banner advertisements (e.g., "click here to earn money").
- List of telephone prefixes (which can be used to randomly generate phone numbers).
- Suburbs, where face-to-face interviews may knock on, say, every fifth door.
- People that phone in to phone-in-polls run by TV stations.

Checking to see if the list is good

The coverage of a survey refers to the extent to which people in the population (e.g.,consumers in the market) could, potentially, have been included in the study. A challenging aspect of all surveys is that most people are unwilling to participate. This means that all surveys tend to have poor coverage. For example, most surveys are conducted using online panels and around 1% of the population is on these panels.

Coverage error is error that occurs when the list is not representative of the population in terms of the things being measured. Consider, for example, using an online panel to understand attitudes to the internet. People that do not use the internet will not be on such a panel and, consequently, any online study seeking to understand attitudes to the internet in the general community will be massively biased.

In an ideal world for research, we would have access to lists containing everybody in the population. This rarely happens in practice. Even companies' lists of their customers are generally incomplete, due to people having provided incorrect contact details, having changed contact details or having advised the company they do not want to participate in market research. Consequently, when evaluating whether coverage error is a problem for a survey, our focus is not on working out whether

the list contains 100% of the population. Rather, our focus is on determining whether the factors that cause somebody to be omitted from a list will be correlated with how the data they would have provided if they had been included. For example, it is problematic to use the a telephone directory as a list if conducting a study of high net worth individuals, as wealthier people are more likely to have unlisted numbers.

Selecting from the list

Often you will not want to attempt to interview everybody on your list. Typically, if you have a large list, the process is:

- Work out your required sample size.
- Take a guess at your response rate, which is the proportion of people that complete an interview relative to the proportion that you invite to the interview. For example, if you invite 10,000 people and 1,000 complete the interview then you have a 10% response rate. It is best to be pessimistic when guessing your response rate.
- Select the required number of people from your list. E.g., if you need 1,000 interviews and your response rate is 10% then you need to select 10,000 people from the list. The process of selecting these people from the list should be random (i.e., it should not favor any particular group of people). A simple way to achieve this outcome is to take, say, every 20th person from your list (where the '20' is adjusted based on the size of the list).

Deciding How to Conduct the Interviews

The vast majority of quantitative market research surveys are conducted online. Nevertheless, there are many other ways of data collection, including in-store interviews, phone interviews, online qual, SMS interviews, door-to-door interviews and mail interviews.

There are numerous different considerations that need to be traded off when choosing how to collect data, including cost, time taken to conduct the research, the ability to show stimulus (e.g., videos, images). In the main, these trade-offs are straightforward to make. The only difficult issue in most real-world surveys relates to coverage.

Coverage

Different ways of conducting interviews 'cover' different proportions of the population. For example, if you work out what the market works by asking your best friend, you have only covered a minuscule proportion of the market. The proportion of the market that can potentially be included in your survey is its coverage. For example, if you decided to conduct a survey by phoning 200 people on a Wednesday evening, the coverage will consist of all the people who:

1. Were at home at the times that you could have called them.
2. Will have answered the phone.
3. Were in the list of phone numbers that you selected your sample from.

Note that the idea of coverage relates to people that you could potentially have contacted, as opposed to those that you did contact.

Coverage error

Where the people in the population that are not able to be contacted by your way of collecting the data are different in important ways from those that can be contacted then the survey will suffer from coverage error.

All surveys suffer from some degree of coverage error. The problem is that we usually have no easy way of working out the likely coverage error and need to apply commonsense.

Minimizing coverage error

It is good practice to choose to collect data in the way that will reduce coverage error. For example, many service companies still conduct important market research by phone, even though it costs more than ten times as much per interview as an online interview. They do this because they often have their customers' phone numbers but are less likely to have all off their customers email addresses (i.e., the list of email addresses has worse coverage than the list of phone numbers).

Coverage error is an endemic problem in market research. As a general rule, the cheapest forms of data collection suffer from the worst sample selection biases. Consider the phone in polls used by TV stations and the web polls run by online newspapers.

Example

Consider a chain of gyms wanting to understand how to increase gym attendance. It seems evident that the people that have the most active lifestyles will be spending less time using computers and are

thus less likely to participate in online surveys, meaning that if an online survey is conducted it will over-represent unfit people and thus using an online survey will result in coverage error'.

While it is clear that conducting an online survey is far from the ideal way of conducting a survey of looking at the exercise market, all of the other options suffer from similar problems and some, such as phone and door-to-door interviewing are considerably more expensive, and thus, unless hundreds of thousands of dollars are available for the survey, conducting such a survey online is usually the only practical solution. Importantly, this does not make the online sample 'valid'. Rather, it simply makes it the tallest dwarf and we need to keep this in mind when interpreting the data from such a survey (and from most online surveys for that matter).

Selecting a Sample

Once the population has been defined, we need a way of finding the people in the population. To do this we need either a list of people in the population, or, a list of places where we can find people.[1]

Surveys involve either selecting everybody from a list or at a specific location (which is known as a census), or, sampling, which involves randomly selecting people from lists or locations.

Lists

Common examples of lists include:

- Online panels, which are databases of people willing to participate in market research in return for pay or prizes.
- Telephone directories.
- Delivery Point IDs, which are lists of all the postal addresses in Australia.
- Lists of customers' email addresses.
- Lists of people willing to participate in focus groups and depth interviews (these lists are held by recruiters).

Such lists are sometimes referred to as the sample. However, it is important not to confuse this with use of the word sample with its normal meaning (where the normal meaning is respondents that have completed the questionnaire).

Places

Common examples of lists of places where we can find people include:

- Shopping centres.
- Websites' banner advertisements (e.g., "click here to earn money").
- List of telephone prefixes (which can be used to randomly generate phone numbers).
- Suburbs, where face-to-face interviews may knock on, say, every fifth door.
- People that phone in to phone-in-polls run by TV stations.

Checking to see if the list is good

The coverage of a survey refers to the extent to which people in the population (e.g.,consumers in the market) could, potentially, have been included in the study. A challenging aspect of all surveys is that most people are unwilling to participate. This means that all surveys tend to have poor coverage. For example, most surveys are conducted using online panels and around 1% of the population is on these panels.

Coverage error is error that occurs when the list is not representative of the population in terms of the things being measured. Consider, for example, using an online panel to understand attitudes to the internet. People that do not use the internet will not be on such a panel and, consequently, any online study seeking to understand attitudes to the internet in the general community will be massively biased.

In an ideal world for research, we would have access to lists containing everybody in the population. This rarely happens in practice. Even companies' lists of their customers are generally incomplete, due to people having provided incorrect contact details, having changed contact details or having advised the company they do not want to participate in market research. Consequently, when evaluating whether coverage error is a problem for a survey, our focus is not on working out whether the list contains 100% of the population. Rather, our focus is on determining whether the factors that cause somebody to be omitted from a list will be correlated with how the data they would have provided if they had been included. For example, it is problematic to use a telephone directory as a list if conducting a study of high net worth individuals, as wealthier people are more likely to have unlisted numbers.

Selecting from the list

Often you will not want to attempt to interview everybody on your list. Typically, if you have a large list, the process is:

- Work out your required sample size.

- Take a guess at your response rate, which is the proportion of people that complete an interview relative to the proportion that you invite to the interview. For example, if you invite 10,000 people and 1,000 complete the interview then you have a 10% response rate. It is best to be pessimistic when guessing your response rate.
- Select the required number of people from your list. E.g., if you need 1,000 interviews and your response rate is 10% then you need to select 10,000 people from the list. The process of selecting these people from the list should be random (i.e., it should not favor any particular group of people). A simple way to achieve this outcome is to take, say, every 20th person from your list (where the '20' is adjusted based on the size of the list).

Determining the Sample Size

Precision and sample size

Consider a simple survey question such as:

```
How likely would you be to recommend Microsoft to a friend or colleague?
Not at all Extremely likely likely 0 1 2 3 4 5 6 7 8 9 10
```

This question was asked of a sample of 312 consumers and the following frequency table summarizes the resulting data:

Likelihood to recommend Microsoft	%	n
Not at all likely 0	7.1	22
1	3.8	12
2	4.8	15
3	3.5	11
4	4.8	15
5	16.0	50
6	9.3	29
7	20.5	64
8	14.4	45
9	8.7	27
Extremely likely 10	7.1	22
NET	**100.0**	**312**

From this table, we can compute that the average rating given to Microsoft is (0×22+1×12+…+10×22)/312=5.9 out of 10. However, only 312 people have provided data. As there are around seven billion people in the world, it is possible that we would have computed a different answer had we interviewed all of them. But how different? This is the question we seek to understand by evaluating precision.

The most common way of communicating the precision of an estimate to a non-technical audience is using confidence intervals. If we use the simplest formula for computing confidence intervals, we compute that the 95% confidence interval for likelihood to recommend Microsoft is from 5.6 to 6.2, which can be interpreted loosely as the range in likely true values that would have been obtained had everybody been interviewed (i.e., based on the survey, we would be surprised if the true figure, obtained by interviewing everybody in the world, was not in the range of 5.6 to 6.2).

Formulas for computing confidence interval take the sample size as an input. With a total of 312 respondents in the sample the confidence interval is 5.6 to 6.2. With a larger sample size there is a smaller confidence interval. Similarly, with a smaller sample size a bigger confidence interval is computed. For example, if we halve, double or quadruple the sample size (n) we can compute the resulting 95% confidence intervals shown in the table below (if this table was not rounded to the first decimal place a more gradual change in the confidence intervals for different sample sizes would be evident).

n	Lower bound	Upper bound	Interval
156	5.4	6.3	5.87±0.43
312	5.6	6.2	5.87±0.31
624	5.6	6.1	5.87±0.22
1,248	5.7	6.0	5.87±0.15

The right-hand side of the table shows the confidence interval in a slightly different way, represented as the mean with a figure ±. If you have ever read a political poll you will have seen this before, where it is described as the margin of error (e.g., a political poll may report a result of 50% with a margin of error of 3%).

A key thing to note about the confidence intervals is that although they get smaller with larger sample sizes, there are diminishing returns. In the example, the largest sample size of 1,248 is 8 times as big as n = 156, but the resulting confidence interval for n = 156 is less than three times as big as that for 1,248.

Guidelines for determining sample size

A practical challenge when designing a study is working out how large a sample is required. Although larger sample sizes reduce the sampling error – that is, the width of the confidence intervals – larger sample sizes cost more money. From time-to-time it is possible to work out the required sample size by working backwards from a desired confidence interval. For example, if Microsoft indicated they wanted to measure the average likelihood of being recommended to within ±0.22, we could determine that this required a sample size of 624. In practice, it is extraordinarily unusual rare for such computations to be used to determine sample size. Instead, market researchers use a combination of gut feel, rules of thumb and understanding of their clients' need for accuracy when working out sample sizes. Some rough rules of thumb for sample size in quantitative studies are: n = (Total budget - fixed costs)/cost per interview n ≥ 300 for basic studies n ≥ 600 for segmentation studies n ≥ 1000 for important strategic studies n ≥ 100 per subgroup (e.g., segment)

Sample sizes with small populations

It is intuitively obvious to most people that if sampling from a small population then you can use a smaller sample size. This intuition is wrong. Unless the sample size is going to be 20% or more of the size of the population, it actual size of the population is pretty-much irrelevant in terms of determining confidence interval and thus is not a factor in working out the required sample size.

That is, the estimated confidence interval of 5.6 to 6.2 is the same width regardless of whether the study had been conducted in a population of 10,000 or 10 billion people. This is mostly good news. It means that there is virtually never a need for super-large samples (e.g., 10,000 or more). Even in China, the sample size used to measure the TV ratings is only 14,650,[1] and the reason it is so large is not because the population is so large, but because there is a need to divide the country up into regions and have relatively narrow confidence intervals within the regions. A flipside of this is that in Australia sampling market research is comparatively expensive, because even though the population is relatively small, the sample sizes still need to be broadly equivalent to those in much larger countries (the Australian TV ratings sample is about one-third the size of the Chinese sample, even though the Chinese population is more than 50 times larger).

Nevertheless, when sampling from relatively 'small' populations (e.g., in business-to-business studies), it is commonplace to use a smaller sample size than in a survey of consumers. However, this is principally because the cost-per-interview is substantially greater in business-to-business surveys.

Setting Quotas

Most surveys employ quotas, sometimes referred to as strata or stratification, to ensure that the data is representative on key variables.

Example

A study of 2,000 energy households used the following quotas. The key logic behind the quotes in this example is that the structure of a household is a key determinant of its energy usage patterns and levels and thus ensuring that the sample is consistent with these quotas provides some assurance that the survey will accurately capture the true variation between households in terms of their energy usage.

Quota Group	Quota	Proportion
Couple - youngest child aged 11 or under	466	23%
Couple - youngest child aged 12 to 17	186	9%
Couple - children at home aged 18 or more	128	6%
One parent family - youngest child aged 11 or under	120	6%
One parent family - youngest child aged 12 to 17	56	3%
One parent family - children at home aged 18 or more	46	2%
No Children - 18 to 24	140	7%
No Children - 25 - 34	238	12%
No Children - 35 - 44	158	8%
No Children - 45 - 54	208	10%
No children - 55 or more	254	13%

Total	2,000	100%

Interlocking and non-interlocking quotas

The quotas shown above involve multiple variables: household structure and age. Such quotas are referred to as interlocking quotas. An alternative approach is to have separate quotas by each variable, non-interlocking quotas, for example:

Quota group	**Quota**
Mainly use PC	100
Mainly use Macintosh	100
Aged under 30	66
Aged 30 to 50	67
Aged more than 50	67

If using these quotas and randomly selecting respondents from an online panel, there is a good chance that by the time 150 interviews have been conducted, the PC quota will be filled (because many more people in the population use PCs so the quota will fill faster) and the two younger quotas will be filled (because most people in online panels are younger). Thus, these quotas may result in the last 50 respondents being Mac users aged more than 50. The resulting data will seem to reveal a strong correlation between age and Mac usage, whereas the reality is that the correlation has been created by the use of overlapping quotas.

A natural conclusion to reach from this example is that we should have large tables of interlocking quotas. For example, hundreds of quota groups defined by age, gender, household structure, geography, and any other key variables. Unfortunately, this is impractical. We generally do not have data to permit us to work out how many customers we have aged 18 to 24 living in NSW in a group household with no children, which makes it impossible to create good quotas. Further, even if we could create the quotas, large numbers of quotas are very expensive, as there always ends up being a few quota groups that are extraordinarily hard to fill.

Proportional versus non-proportional quotas

Most quotas are proportional, which is to say that the number of respondents required in each group is determined by the number of people in the population believed to be in each group (e.g., based on ABS published statistics). Sometimes the quotas may be non-proportional to ensure enough people in a key segment are in the study (e.g., if doing a study for Apple, you may want 50% of the sample to be Apple customers); if employing non-proportional quotas the data needs to be weighted and this requires that you know the correct population proportions for weighting.

Online interviewing and send quotas

Developing quotas for online interviewing requires an additional step to that described above. In addition to creating quotas relating to how many people complete the survey (which is what the above descriptions relate to) there is additionally also a need to create quotas relating to the number of people invited to participate. These should always be interlocking and it is generally advisable to create a large number of quotas (e.g., creating quota categories of 5 year age bands by gender by geography).

Some online panel companies refer to such quotas as weighting.

A good quality online study requires both quotas of the sends and quotas on the number of completes. Where quotas are only provided on the number of completes it can lead to bizarre results. For example, in one study which used the age bands of Under 30, 30 to 50 and 51 or more, almost all of the respondents in the 51 or more category were 51 years old, meaning that the survey had no representation of older consumers. When quotas are used on the sends but not on the completes, the result tends to be an over-representation of older consumers, because they are more likely to respond to an invitation to participate.

Code Frame

The list of Values and their associated interpretations in a question. For example, in a question asking about gender, the code frame is:

Value	Value Label
1	Male
2	Female

The specific terms of Value and Value Label are the terms used most commonly (e.g., in SPSS, Q and Displayr). However, there is no standard terminology (e.g., some researchers refer to the Value Label as a 'code' and others refer to the Value as a 'code'.

The code frame will typically be created in one of the following ways:

- It will exist in the questionnaire.
- When the questionnaire is programmed the categories will be entered (i.e., the Value Labels) and the Values will automatically be generated by the program (typically, with a 1 for the first category of a single response question, a 2 for the second category, etc.).
- It will be created at the time of coding the text data.
- It will be created using the first two methods and then modified during the coding of Other (specify) options.

Multiple response questions

The meaning of the code frame in a multiple response question is slightly ambiguous. Most commonly, it will refer to the list of options and there will not be any specific values assigned to each of these values. However, the ambiguity is that the variables that constitute a multiple response question will typically have their own code frame, which will usually be something like:

Value	Value Label
0	Not selected
1	Selected

or

Value	Value Label
1	Yes
2	No

Fielding the Survey

Fielding is the industry jargon for conducting the actual survey.

The soft send

In online surveys it is good practice to conduct a soft send once the survey has been programmed and tested. This involves sending enough invitations such that, say, 50 to 100 interviews, are completed. The resulting data is then used to test and check various assumptions about the research design. In particular:

- Does the resulting sample appear to be sensible (see Checking Representativeness).
- How much effort will it take to achieve the desired number of interviews (e.g., how many people will be invited to participate).
- Will it be possible to achieve the desired number of interviews.
- Is the resulting data sensible and useful. Generally this can only be assessed by creating and reviewing a summary report.

Checking quotas and sample size

Additional interviews are conducted until the sample size and the specified quotas have been achieved. Where there is a particular problem filling some of the quotas this means that either there is a problem with the list being used, or, that the quotas are wrong. Either way, it is important to carefully review any difficulties in achieving quotas rather than to continue to collect data until the quotas are all filled as a difficulty in filling a quota can indicate that the resulting sample has a coverage bias.

Programming and Testing

Programming refers to entering the questionnaire into the Data Collection Software and testing refers to checking that it has been programmed correctly. Key things to check are that:

- The skips in the questionnaire make sense. That is, that there are no combinations of options that people select which cause them to receive an error message or to be shown the wrong questions.
- That anybody can do the questionnaire. On old or very young family member is one good way of checking.
- That the data exports correctly. This is done by creating and reviewing a Summary Report.

It is good practice to conduct a pilot survey to a dozen respondents and check that the resulting data is as anticipated. An even better practice is to sit with these respondents as they are doing the survey and ask them to think aloud, explaining how and why they are answering the questions, as this can given insight into any problems with question wording.

Data Collection Software

There is a large variety of online services for survey design and deployment. Prices range from free to thousands of dollars for a yearly subscription. Generally, more expensive packages allow for greater levels of customization with regard to survey design, improved features for obtaining respondents, and more flexible output options.

Comparison table of services

The table below presents some of the services that offer online survey creation and deployment. For analysis, the most important factor is the availability of an appropriate data file. SPSS data files are usually the best of the file types available in the cheaper data collection programs as they contain

metadata that allows software to interpret the meaning of the data. CSV and other spreadsheet formats require more work to set up data in analysis software packages. Companies marked with an "N/A" for data files either:

1. Don't supply SPSS data files
2. Did not have data files available to inspect for quality

If you wish to suggest improvements/changes to the table below please provide some detail and a contact email in the comment form at the bottom of this page.

Company	Cheapest Package	Cheapest package with complete data export	Interesting Features
SurveyMonkey	**Basic** - Free	**Gold** - $300/year • SPSS data files • Unlimited responses • Surveys deployed by email, website link, Facebook, Twitter • Export reports to Microsoft Office and online	• Word Clouds • Question randomization
Survey Gizmo	**SG Lite** - Free	**Survey Gizmo** - $810/year • SPSS data files • Unlimited responses	• TURF reports • Customized/branded reports

		• Surveys deployed by email, website link, Facebook, Twitter • Online and PDF reports	
QuestionPro	Free	**Corporate** - $1008/year • SPSS data files • Surveys deployed by email, website link, Facebook, Twitter • Online, PowerPoint, Excel reports	• Support for multimedia in surveys • Integration with third-party panel/sample providers
Qualtrics	**Free account**	**Research Suite** - N/A • SPSS data files • Unlimited surveys and responses	• Create detailed online reports • Panel management
Snap Surveys	N/A	**Single PC Installation with hosting** - $1,995 + $395/year • Up to 5000 responses • SPSS data files	• Include multimedia in surveys • Produce paper version of survey, allows

		• Design and analyze surveys from desktop software	for manual data entry • Purchase panels from Snap
Toluna QuickSurveys	Free	NA	• No subscription fee, pay to recruit respondents • Smartphone-friendly surveys
Super Simple Survey	**The Basic** - Free	NA	• Intuitive survey building • Smartphone-friendly • Multiple users per account • No limit on accessibility of features
Polldaddy	Free	N/A	• iOS app for collecting responses
Obsurvey	Free	N/A	• Completely free survey creation and hosting
Wufoo	Free	N/A	• Simple service for building online forms

			• Multiple users per account • Wide range of templates and themes
Insightify	Free	N/A	
Survey Crafter	$495 for a single license	N/A	• Desktop survey-creation tool • Basic analyses include summary data and cross-tabulation

Basic Questionnaires

- Question Types
- Phrasing the Question
- Programming Instructions
- Ordering the Questions in the Questionnaire

Basic Questionnaires

Writing a questionnaire involves the following steps.

Identify everything that you need to measure

- Questions that you need to answer. For example, whether people are interested in buying a new product.
- Diagnostics that will help you understand the answers to your key questions. For example, reasons for liking or disliking a new product.
- Questions that will be useful for profiling. For example, age and gender.
- Questions that are useful for checking the representativeness of the study.

Writing one or more question for each thing that you need to measure

This involves:

- Selecting the Question Type for each question
- Phrasing the Questions.
- Writing Programming Instructions.

Ordering the Questions in the Questionnaire

Questions in a questionnaire are ordered so as to reduce the time taken to complete the questionnaire and to reduce order effects.

Question Types

Open-Ended Questions

```
Please enter the name of the last soft drink you bought.

_____________
```

Most data collection programs will let you constrain the type of responses. For example, when collecting numeric data, forcing respondents to provide a numeric response and perhaps limiting the range of possible responses:

```
How many glasses of wine did you drink last night? ____
```

Open-ended questions are the easiest type of question to ask. However, they should generally be used sparingly,[1] as they take longer for respondents to answer and many respondents will put so little effort into answering the questions that their data becomes unhelpful.

Single Response Questions

```
Are you...

o Male

o Female
```

When writing a good single response question, it is important to:

- Ensure that each person will only fit into one and only one category.
- Ensure that each person can find an appropriate category. Often this is achieved by leaving an "Other" option. It is typically a good idea to require that the respondent type what they mean into a box, as otherwise you can end up with a lot of responses that are not intepretable.
- Where the question forces people to compare the options (e.g., Which of these best describes your personality?", provide a maximum of seven responses, as otherwise by the time people have gotten to the end of the list they will have forgotten the earlier responses).
- Where the question is only requiring a simple act of recall (e.g., Where did you go for your last holiday), order the responses alphabetically so they can quickly find the appropriate response (this is particularly important when there are large numbers of options).

Multiple Response Questions

This question type is also known as a pick any or a 'multi' or 'multicode' question.

```
Which of the following have you bought in the past week? Tick all that apply.

[] Coke

[] Pepsi

[] Fanta

[] None of these
```

Grid Questions

There are a variety of questions which are referred to as 'grid' questions. They differ in terms of how they permit the respondent to choose options, with common alternative being to limit respondents to choosing a single option in each row, allowing them to tick multiple options in each row, sliders and text boxes for entering numeric responses.

One response per row

```
Please rate your satisfaction with the following restaurants
```

	Low	**Medium**	**High**
McDonald's	o	o	o
Burger King	o	o	o
Wendy's	o	o	o

Multiple responses per row

```
Which of these brands are
        Fun  Sexy  Masculine
Coke    []    []    []
Pepsi   []    []    []
Fanta   []    []    []
```

Typically, such questions are best instead asked as a series of multiple response questions, as respondents can be daunted or confused by such a grid layout.

Numeric responses

```
In the past month, how many economy flights did you take on... Qantas United SAS
Economy ______  ______  ______ Premium Economy ______  ______  ______
Business Class ______  ______  ______ First Class ______  ______  ______
```

Date Questions

Date questions can either be similar or open-ended questions, such as in the example below, or can use a data selector (i.e., a calendar).

```
What is your date of birth?

____ / ____ / ____
```

Ranking Questions

The simplest ranking questions use a series of open-ended questions:

```
Rank the following brands according to how much you like them... Please place a 3
next to the brand you like most, a 2 in your next preferred brand and a 1 next to
your least preferred brand.

Coke ____

Pepsi ____

Fanta ____
```

More sophisticated versions use dragging and dropping.

Constant-Sum Questions

A constant-sum question asks the respondent to allocate a specific number of points, tokens, dollars or some other quantity across different alternatives. For example:

```
If you had $10 to spend on soft drinks, how much would you spend on each of the
following? __ Coca-Cola __ Diet Coke __ Coke Zero __ Fanta
```

More advanced question types

There are many more advanced question types. Some of these are described in Advanced Questions and Questionnaires.

Phrasing the Question

Writing a question involves selecting a question type and phrasing the question. For the novice researcher, the secret to wording questions well is to copy others' work. Plagiarism is at the heart of asking good questions, as there are many, many ways of wording a question poorly. The easiest way to get across the basics is to look at a few examples of bad questions (all from real studies): What is your principal brand of soft-drink? What problems can you see with this question? 'Principal' is an overly technical word. Sure you know what it means, but what about the guy who collects the garbage? A better word than principal is 'main'. However, this does not remove the ambiguity from the question: is it asking which one is liked the most, or, bought the most often? Such a question will likely end up measuring brand salience (i.e., which brands come to mind) rather than anything else.[1]

And another example:

```
Do you prefer…  [] Beer  [] Wine  [] Spirits  [] Non-alcohol drinks
```

There are lots of problems with this question. What about cider drinkers? They are ignored (i.e., the categories are not exhaustive). Furthermore, people that do drink are likely to drink different types of alcohol at different times, such as beer after sport, wine with dinner and spirits after dinner; how should such a person answer the question?

Key principles when phrasing a question

Principle 1: The question must be answerable

The least-educated dumb person that is likely to have to answer the question needs to be able to figure out how to answer the questions accurately.

Principle 2: The question must not be ambiguous

The issue is not one of respondent confusion, as this is already addressed by the first principle and respondents are remarkably adept at answering incomprehensible questions (presumably because this is the only way they can finish many questionnaires). The problem with ambiguity is to avoid situations where, once we have collected the data, we cannot discern what it means. For example, one study asked flyers about the importance of getting 'discounts' and found that it was important. However, the research did not clarify if the flyers were happy with a 2% or 20% discount, the research was ambiguous and could not be used to derive valid insights. To avoid being ambiguous, questions

need to focus on the current, the specific and the real.[2] A trick to achieving this is to focus on writing questions around the 5Ws – who, what, where, when and why.[3]

A good way of checking for ambiguity of questions is to use think aloud interviews, in which real respondents are asked to answer the question but are required to verbalize all their thoughts while answering the question. A shortcut to this can be to get the dumbest person you know to complete the questionnaire and comparing their answers with what you believe to be true.

Principle 3: Incentive compatibility

The third principle is that questions must be incentive compatible, which is a term of art in economics, and it means that questions need to be written in a way that people have an incentive to provide honest data. Consider the question:

```
Are you aged under 35? [] Yes [] No
```

If asked as a screener at the very beginning of a questionnaire, respondents conclude that an answer of No will cause them to be screened out of the study (i.e., asked no more questions). Consequently, when this question is asked at the beginning of questionnaires where people are being paid more if they complete the whole questionnaire, some respondents lie and pretend they are aged less than 35, when they are actually older. By contrast, if asked at the beginning of a questionnaire where people are not being paid to do the study, people aged under 35 pretend to be older, as this becomes a polite way of refusing to participate. An incentive compatible way of screening people based on being aged under 35 instead asks:

```
How old are you? [] Under 18 [] 18 to 24 [] 25 to 34 [] 35 to 44 [] 45 to 54 [] 55 to 64 [] 65 or more
```

and then screens people out if they select one of the older age groups. It is “incentive compatible” because there is no incentive to answer the question dishonestly.

Imagine yourself charged with the problem of pricing an iPad before they were launched. How much should you charge? There were no competitors, and thus no straightforward way of working out even a ballpark price point. A simple approach to this problem is to ask people how much they might pay:

```
What is the most you would pay for the product?
```

The amount that people nominate is referred to as either the reservation price or their willingness-to-pay for the product. The analysis then proceeds by assuming that people will buy a product if its price is less than or equal to the amount nominated by the respondents.

Such questions are not not incentive compatible. Imagine that you loved the concept and would be prepared to pay $2,000 for an iPad. Would you tell this to an interviewer? If you did, it could result in the product being sold at $2,000. Whereas if you lied and said $1,000, you may end up with the product being sold for less and you not having to spend so much. Sure, some respondents may be honest, either because they are by nature honest or because they believe that by being honest they will maximize their chance of the product they wish to buy being sold, but others may lie. Observing that respondents have an incentive to lie does not prove that they will lie (unless one is an economist). However, it is better to figure out ways to conduct research that are incentive compatible, as then the risk is removed (this is discussed in Advanced Questions and Questionnaires).

Frequency questions

Perhaps the most frequent mistakes made by novice researchers relate to frequency questions. Consider this question:

```
How often do you go to the cinema?[] Never  [] Rarely  [] Sometimes  [] Often
```

What do the scale points of Rarely, Sometimes and Often mean? One person's often may be another person's sometimes. When measuring quantities it is always important to give people precision about time and frequency, such as:

```
Thinking about the last twelve months, how many times did you go to the movies?[]
None [] Once or twice  [] 3 to 6 times [] 7 or more times
```

A problem with such a question is that respondents may not be able to answer it accurately because they will not remember the precise number of trips to the cinema from the past twelve months. Such a reservation is justified. A solution is to reduce the time interval:

```
Thinking about the last two weeks, that is, everything from today back to
Wednesday two weeks ago, how many times did you go to the movies?[] None  [] Once
[] Twice  [] 3 to 6 times [] 7 or more times
```

Asking about such a short time period creates a different problem. Attendance at the cinema is sporadic for most people. There is a big difference between somebody who goes six times a year and

somebody who does not go at all, but asking about the last two weeks can result in such people being grouped together and, thus, it is better to ask about the longer time interval.

So, what then is the best way to ask about frequency of going to the cinema? A question like:

```
Thinking about the last twelve months, how many times did you go to the movies?[]
None  [] Once or twice  [] 3 to 6 times [] 7 or more times
```

perhaps remains the best approach, as even though many people will not be able to answer with complete accuracy, most of the people who have not gone will be able to select None and most that have gone more than seven times will be able to select 7 or more times, whereas with the vaguer Rarely, Sometimes and Often categories, we can not make any deductions about what they actually mean.

Another example:

```
How often do you visit the web pages of this site?[] Several times a day[] Once a
day[] Once a week[] Once a month[] Less often than once a month[] This is the
first time I have visited
```

These categories are not mutually exclusive (anybody who can tick the last category can also tick the second last category). And, they are not exhaustive. What should somebody answer if they visit two or three times a week?

Programming Instructions

Once the basic question wording has been finalized it needs have instructions added so that the person who has to program the questionnaire in the data collection software can ensure that it works as intended. By tradition these instructions are entered in CAPITALS, italics or another color. In addition to the instructions sometimes people enter specific values for each of the options. In the example below a Student is represented by a 1; this information is used to determine how the categories are represented in the resulting data file.

Example

```
Q2. ASK IF AGED 15 OR OVER: To make sure we have a cross section of people in this
study, we need to know your work status. Which of these best describes your ''work
status''? SINGLE RESPONSE Student 1 GO TO Q4 Homemaker 2GO TO Q4 Retired
3 GO TO Q4 Not working 4GO TO Q4 Part-time worker5GO TO Q4 Full-time worker 6
Don’t know/refused 7 END SURVEY
```

Ordering the Questions in the Questionnaire

Questions in a questionnaire are ordered so as to reduce the time taken to complete the questionnaire and to reduce order effects.

Minimizing the time to complete the questionnaire

Screeners

A screener is a question that can be asked at the beginning of a questionnaire with the objective of determining eligibility. For example, if conducting a study on cola drinkers, then a question about cola consumption is asked at the beginning of the questionnaire so that non-cola consumers are not asked the rest of the questionnaire (i.e., are screened out to use the industry jargon).

Skips

A skip is an instruction in a questionnaire which makes respondents skip questions for which their answers are not required. For example, if a person has no children then they should skip a question asking about the ages of their children.

Filters

Options that are not relevant to some respondents should not be shown (i.e., filtered). For example, if a consumers has never heard of a product they should not then be shown that product in a following question asking which products they consume most often.

Multiple response questions and grids

Questions with similar structures should be grouped together. For example, true/false questions can be grouped together as multiple response questions. And, questions which share a response scale can be grouped into grids.

Deleting questions

The shorter the questionnaire the better. Consequently, questions should only be included that are believed to be essential.

Minimizing order effects

An order effect occurs when people give answers in questionnaires that are influenced to an extent by the order with which the questions, or options in a question, are presented to respondents.

Knowledge and salience

The simplest type of order effect occurs when earlier questions unintentionally educate a respondent, or, make them think about a particular thing. For example, if a question that asks people if they have heard of Qantas and then, later in the questionnaire asks them to type the names of all the airlines for which they can recall recently seeing TV advertisements, there is a good chance that Qantas will get a higher result in the latter question than otherwise.

Leading questions

Leading questions cause respondents to change their opinions, which can effect either the answer of the question that is leading, or, the answers to following questions. Consider the following from the classic British sitcom Yes Prime Minister:[1]

Sir Humphrey: "You know what happens: nice young lady comes up to you. Obviously you want to create a good impression, you don't want to look a fool, do you? So she starts asking you some questions: Mr. Woolley, are you worried about the number of young people without jobs?"

Bernard Woolley: "Yes"

Sir Humphrey: "Are you worried about the rise in crime among teenagers?"

Bernard Woolley: "Yes"

Sir Humphrey: "Do you think there is a lack of discipline in our Comprehensive schools?"

Bernard Woolley: "Yes"

Sir Humphrey: "Do you think young people welcome some authority and leadership in their lives?"

Bernard Woolley: "Yes"

Sir Humphrey: "Do you think they respond to a challenge?"

Bernard Woolley: "Yes"

Sir Humphrey: "Would you be in favour of reintroducing National Service?" [A requirement that people join the army after leaving school.]

Bernard Woolley: "Oh...well, I suppose I might be."

Sir Humphrey: "Yes or no?"

Bernard Woolley: "Yes"

Sir Humphrey: "Of course you would, Bernard. After all you told me can't say no to that. So they don't mention the first five questions and they publish the last one."

Bernard Woolley: "Is that really what they do?"

Sir Humphrey: "Well, not the reputable [market research firms] ones no, but there aren't many of those. So alternatively the young lady can get the opposite result."

Bernard Woolley: "How?"

Sir Humphrey: "Mr. Woolley, are you worried about the danger of war?"

Bernard Woolley: "Yes"

Sir Humphrey: "Are you worried about the growth of armaments?"

Bernard Woolley: "Yes"

Sir Humphrey: "Do you think there is a danger in giving young people guns and teaching them how to kill?"

Bernard Woolley: "Yes"

Sir Humphrey: "Do you think it is wrong to force people to take up arms against their will?"

Bernard Woolley: "Yes"

Sir Humphrey: "Would you oppose the reintroduction of National Service?"

Bernard Woolley: "Yes"

Sir Humphrey: "There you are, you see Bernard. The perfect balanced sample."

Guiding and educating respondents

The opposite to leading a respondent is to guide them, as occurs in Probing and Laddering. In studies for new products it is routine to progressively introduce more information about the products throughout the questionnaire, so that the respondent is gently introduced to the information and also so that the researcher can understand the incermental impact of the additional information.

Fatigue

Harder questions should generally be asked towards the beginning of the questionnaire, so that if respondents become fatigued while completing the questionnaire, the fatigue occurs after they have completed the more challenging questions. And, the order of questions can be used to maximize the energy of the respondent, mixing boring questions with more interesting questions (e.g., projective questions).

Advanced Questionnaires: Attitudes, Beliefs, Personality and Pricing

> Although most questionnaires are written using the approach described in Basic Questionnaires, more advanced approaches have been developed for solving some specific problems. When these approaches are used they typically involve the creation of one or more additional questions which then form a section of the questionnaire.

Advanced Questionnaires: Attitudes, Beliefs, Personality and Pricing

Attitudes, beliefs and personality

Measuring Abstract Concepts

Most questions in questionnaires measure concrete things like age. A much harder challenge is measuring abstract things. For example, how can we measure involvement, innovativeness, cynicism, sophistication, or susceptibility to peer group pressure? The solution is to use a multi-item scale which asks multiple similar questions and computes an average response.

The logic of multi-item scales

If you've ever done a personality test or an IQ test you will be familiar with the solution to this problem: ask a whole lot of related questions. Consider the problem of trying to work out which people are susceptible to the influence of other people. We could try and understand this by asking people how strongly they agree or disagree with the statement: `I rarely purchase the latest fashion styles until I am sure my friends approve of them.`

However, a problem with this question is that many people may never purchase the "latest fashion styles", making it difficult to interpret what their answer may mean. We can express this as a simple measurement model:

Where `Truth` refers to the true degree of susceptibility to influence of other people, `Measurement` is the answers that people give to this question and `Error` refers to all of the factors that prevent the measured degree of social susceptibility from being equal to the true value. Discrepancies' between the truth and what is measured are known as measurement errors.

This equation, which is referred to as the measurement error model, tells us that there will be a correlation between `Truth` and `Measurement`, but there is also a correlation between Error and Measurement, meaning that any Measurement will reflect a combination of `Truth` and `Error`.

Another question wording for trying to understand susceptibility to social influence is: `To make sure I buy the right product or brand, I often observe what others are buying and using.`

This suffers from the problem that some people's susceptibility may come about through other channels, such as Facebook and by looking at YouTube. Expressing this as a formula gives:

```
Truth + Error_2 = Measurement_2
```

where the 2_ indicates that this is a different Error to that in the earlier equation.

If we add together both of the equations, we get:

```
2×Truth + Error + Error_2 = Measured + Measured_2
```

Dividing both sides by 2, so that it reflects the average of the two measures gives:

```
Truth + (Error + Error_2)/2 = (Measured + Measured_2)/2
```

Where two measurements are averaged, the resulting measure is referred to as a multi-item measure, so, we can rewrite the equation as:

```
Truth + (Error + Error_2)/2 = Multi-item measure
```

Now, you may be thinking at this stage, "so what", but the above equation is insightful if we make an assumption. The assumption that we need to make is that there is a negligible relationship between Error and Error_2 (i.e., there is close to a 0 correlation between the variables). If this assumption is true, it means that the correlation between the multi-item measure and the truth will be greater than when a single item (i.e., question) is used for measurement. This is because the errors do, to an extent, cancel each other out.

The more wordings of a question that we use, the greater the correspondence between the estimated value and the truth, provided that each different variable has similar levels of error. If it is the case that some of the questions have discernibly more error than others, or, some of the variables' errors are closely correlated with each other, then it is better to omit these variables.

Example: Consumer susceptibility to interpersonal influence[1]

Each respondent's average rating of the following 12 items on a seven-point scale of:

```
Strongly Strongly Disagree Agree   1     2     3     4     5     6     7
```

is an estimate of the role of interpersonal influence on brand and category choice decisions.

1. I often consult other people to help choose the best alternative available from a product class*
2. If I want to be like someone, I often try to buy the same brands that they buy
3. It is important that others like the products and brands I buy
4. To make sure I buy the right product or brand, I often observe what others are buying and using*
5. I rarely purchase the latest fashion styles until I am sure my friends approve of them
6. I often identify with other people by purchasing the same products and brands they purchase
7. If I have little experience with a product, I often ask my friends about the product*
8. When buying products, I generally purchase those brands that I think others will approve of
9. I like to know what brands and products make good impressions on others
10. I frequently gather information from friends or family about a product before I buy*
11. If other people can see me using a product, I often purchase the brand they expect me to buy
12. I achieve a sense of belonging by purchasing the same products and brands that others purchase

The items with a * next to them are measures of the role that other people have as a source of information, while the remaining items measure the direct influence of other people. As only a third of the items have an asterisk this means that the scale more heavily weights other people as a direct influence than other people as a source of information.

How to create multi-item scales

1. Dream up many possible items for each concept that is to be measured (e.g., dozens or hundreds), where an item is a specific statement that is rated.
2. Conduct a small survey of, say, 100 people, and get them to provide a rating on a 5, 7, 9, 10 or 11 point scale for each item.
3. Compute the correlations between each of the items (or, equivalently, use Principal Components Analysis).
4. Discard any items that are not correlated with all the other items.
5. Work out how many items you need to retain. If you are just after a very rough measurement, two or three may be sufficient. If you are wanting to make very precise statements about differences between people you will likely need at least 10. Special formula's have been developed for working out how many items you need and how accurate your ratings will be in multi-item scales, including Cronbach's Alpha.

Finding existing multi-item scales

Thousands of multi-item scales have been developed by academics, and their use can save many hours of time in questionnaire development. There are a number of handbooks containing many of the scales, and they are worthwhile investment for market researchers (sometimes there is a need to reword the questions to make them more consumer friendly). For example:

Beardon, William O., Richard G. Netemeyer, and Mary F. Mobley (1993), Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research. Newbury Park: Sage Publications

Bruner, Gordon C. and Paul J. Hensel (1995), Marketing Scales Handbook. Chicago: American Marketing Association.

The academic discipline of psychometric focuses on the problem of how to create multi-item

Measuring Needs

abstract and the basic process for measuring needs is the same as that described above.

Multiple different multi-item scales have been developed for measuring needs. One of the more popular is Schwarz's Human Values Scale, which is used to understand differences between people in their values (i.e., their central goals in life). The scale contains ten different dimensions, as Schwarz's theory is that almost all differences between cultures can be explained using these ten dimensions. Each dimension is measured with two or more items, where people are asked to indicate whether the people described are:

```
Very much like me Like me Somewhat like me A little like me Not like me Not like me at all
```

The items below are a version of the scale used in the European Social Survey:

VALUE and central goal	Items that measure each value
Power: Social status and prestige, control or dominance over people and resources.	It is important to him to be rich. He wants to have a lot of money and expensive things.It is important

	to him to get respect from others. He wants people to do what he says.
Achievement: Personal success through demonstrating competence according to social standards.	It is important to him to show his abilities. He wants people to admire what he does.Being very successful is important to him. He hopes people will recognize his achievements.
Hedonism: Pleasure and sensuous gratification for oneself.	He seeks every chance he can to have fun. It is important to him to do things that give him pleasure.Having a good time is important to him. He likes to "spoil" himself.
Stimulation: Excitement, novelty, and challenge in life	He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.He looks for adventures and likes to take risks. He wants to have an exciting life.
Self-direction: Independent thought and action choosing, creating, exploring.	Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.It is important to him to make his own decisions about what he does. He likes to be free and not depend on others.
Universalism: Understanding, appreciation, tolerance and protection for the welfare of all people and for nature.	He thinks it is important that every person in the world should be treated equally. He believes everyone should have equal opportunities in life.It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them. He strongly believes that people should care for nature. Looking after the environment is important to him.
Benevolence: Preservation and enhancement of the welfare of people with whom one is in frequent personal contact.	It is very important to him to help the people around him. He wants to care for their well-being.It is important to him to be loyal to his friends. He wants to devote himself to people close to him.

Tradition: Respect, commitment and acceptance of the customs and ideas that one's culture or religion impose on the individual.	It is important to him to be humble and modest. He tries not to draw attention to himself.Tradition is important to him. He tries to follow the custom handed down by his religion or his family.
Conformity: Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.	It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong.He believes that people should do what they are told. He thinks people should follow rules at all times, even when no-one is watching.
Security: Safety, harmony and stability of society, of relationships, and of self.	It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.It is important to him that the government ensures his safety against all threats. He wants the state to be strong so it can defend its citizens.

Ratings with Rankings to Split Ties

A common problem with the use of rating scales is that respondents often will assign many and perhaps most of their ratings to a single scale point. For example, if asking respondents to rate the importance of various product features, such as price, quality and variety of flavors, it is not uncommon for some respondents to indicate that many of the items score a 10 on a rating scale such as the one shown below:

```
0 Not important 1 2 3 4 5 6 7 8 9 10 Very important
```

A solution to this problem is to get respondents to rank any tied ratings. For example, if the respondent indicated that price, quality and variety of flavors all received a rating of 10 the next question would require that the respondent ranked these three items. If relevant, similar rankings can be used to split items which received the same rating for other scale points.

Max-Diff

Max-Diff is an alternative to rating scales which seeks to solve the same problem.

A Max-Diff experiment presents respondents with a series of questions like the one shown below where the set of alternatives shown is varied from question-to-question so that it is possible to isolate the relative appeal of the different alternatives.

```
What is the MOST and LEAST important factor to you when you choose a car? MOST
LEAST [] Price [] [] Top speed [] [] Fuel economy [] [] Emissions [] [] Paint
color []
```

For a detailed description of the design and analysis of max-diff experiments, see Max-Diff.

It is sometimes described as a Choice Modeling method, as similar tools are used to create the experimental designs and to analyze the data.

Pricing

Price Sensitivity Meter

The price sensitivity meter is a popular set of questions designed to provide assistance in understanding the range of prices that can be relevant for a completely new product. There are a few minor variants, but the standard wordings are:

Q1. At what price would you consider this PRODUCT/BRAND to be so inexpensive that you would have doubts about its quality? [“Too cheap”]?

Q2. At what price would you still feel this PRODUCT/BRAND was inexpensive yet have no doubts as to its quality? [“Cheap”]?

Q3. At what price would you begin to feel this PRODUCT/BRAND is expensive but still worth buying because of its quality? [“Expensive”]?

Q4. At what price would you feel that the PRODUCT/BRAND is so expensive that regardless of its quality it is not worth buying? [“Too expensive”]?

The third of the questions can be interpreted as collecting the reservation price (see Phrasing the Question), with the other questions framing this last question in much the same way that asking

people which animal best represents a brand frames the “why” question with Brand Personification question.

From-time-to-time the price sensitivity meter is used to generate price-demand curves and make predictions about sales at different price points. As the questions are clearly not incentive compatible and completely ignore the availability of substitutes, and predictions made using such a technique seem likely to have little prospect of being remotely accurate, except by chance.

Also known as

Van Westendorp's Price Sensitivity Meter.

References

[1] Beardon, William O., Richard G. Netemeyer, and Mary F. Mobley (1993), Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research.

Advanced Questionnaires: New Product Development

- Ideation
- Gap Identification
- Concept Test
- Choice Task
- Experiments
- Choice Modeling
- Decompositions
- Pre-launch Forecasting

Advanced Questionnaires: New Product Development

Ideation

The basic idea

Once a firm has identified the general area on which it wants to concentrate its new product development efforts, the next step is to come up with ideas. Sometimes this is done by technical specialists, such as food technologists and engineers. However, in other situations it is done using ideation. When market researchers are involved this ideation tends to involve consumers, but the same basic method can be used with stakeholders and other employees of the company. Ideation is short for idea generation. Ideation itself is just an application of brainstorming.

Brainstorming

The basic logic of brainstorming is this: the best way of coming up with a good idea is to come up with lots and lots of ideas. A team that creates 50 ideas and chooses the best will, in general, end up with a much better idea than a team that that comes up with only 20 ideas. A way of coming up with lots of ideas is as follows:[1]

1. Get a group of people together, but get each to initially work on their own. The widespread belief that people get ideas by bouncing off each other is not supported by evidence.

2. Give people a specific target in terms of the number of ideas to come up with. For example, give them a target of 10 new ideas in 10 minutes. The trick is to make the target hard – the harder the target, the more ideas people will create. Furthermore, the specific goal is to come up with lots of ideas, not a small number of good ideas, and people should be instructed that quantity, not quality, counts.

3. Then, as a group, review the ideas to identify winners.

Encouraging creativity

Ideation projects usually contain a couple of twists on traditional brainstorming. One of the twists is to give people a warm up task to encourage their creativity. The other is to give them tasks designed to guide them towards better ideas, as is done in SCAMPER below.

Example of a warm up task[2]

- Generate two lists each of four random words (e.g., by randomly opening pages of a dictionary). For example:

```
Toast runway prophet formula chip lollipop tail pyramid
```

- On some basis, pair words in each of the lists
- Define the basis for the pairing. For example:

```
“toast” and “prophet” are both edible “tail and "formula” can be very long
“profit” and “runway” suggest departure for the future
```

SCAMPER

SCAMPER is a checklist of activities designed to spur creative thinking, all of which start by thinking about how to modify an existing product or service:

- **S**ubstitute something
- **C**ombine it with something else
- **A**dapt something to it
- **M**odify or Magnify it
- **P**ut it to some other use
- **E**liminate something
- **R**everse or Rearrange it

Gap Identification

The most common role that market research plays in helping companies to create new products is to identify gaps in the market, where a gap is an unmet need, want or identification of a problem with existing products. This page describes three related approaches to this problem; there are numerous other approaches.

Jobs-to-be-done

Any purchase can be thought of in terms of the jobs that need to be done which motivate the purchase. When buying a car, we may have in mind the jobs of getting to work, transportation to holidays and keeping up with the Joneses. When buying yogurt, the jobs to be done may include functional jobs, such as losing weight and satisfying hunger, emotional jobs such as making us happy and social jobs such as impressing people. The jobs may need to be done regularly, such as filling our stomachs, or they may be one-off jobs. They can be driven by a specific situation, such as having had to skip a meal, or be intrinsic to the person, such as a desire for a snack at the same time every day. The following table shows jobs to be done in the yogurt category.

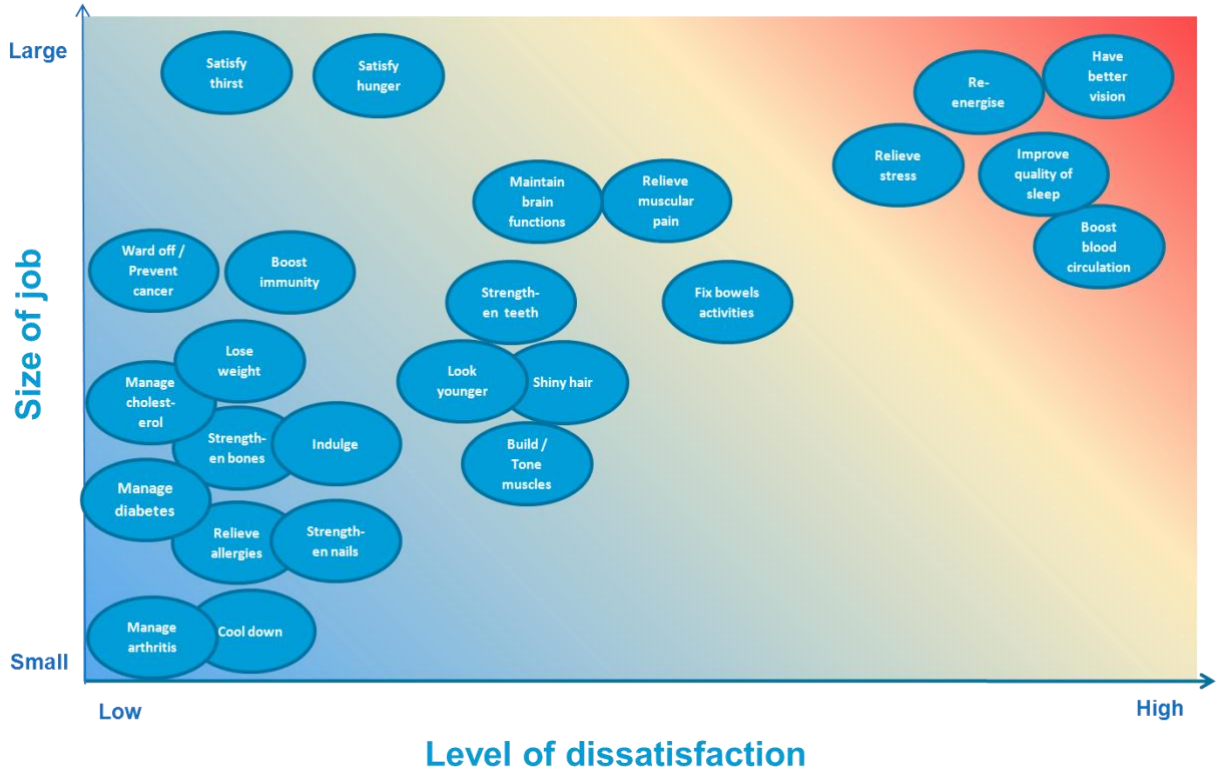
Satisfy hunger	**Lose weight / Maintain current weight**
Boost blood circulation	Manage arthritis
Boost immunity	Manage cholesterol
Build / Tone muscles	Manage diabetes
Cool down	Relieve muscular pain
Ease allergies	Relieve stress
Fix / Maintain digestion / bowels activities	Re-energise
Improve brain functions (e.g. Memory, alertness)	Satisfy thirst
Improve vision	Strengthen bones

Improve quality of sleep	Strengthen nails
Have shiny hair	Strengthen teeth
Indulge	Ward off / Prevent cancer
Look younger	

Note how different these are to generic statements of benefits and needs; they are all precise descriptions of activities that people do, or things they may want a yogurt to provide. When trying to come up with a list of jobs, it can be useful to focus on ancillary jobs, which can be split into related and unrelated ancillary jobs. When Apple launched the iPod, it not only provided a device for playing music, it also solved the related problems of buying and cataloging music. Similarly, Nespresso solves the problem of buying, storing and grinding coffee. Non-related jobs are jobs that can be done at the same time, such as taking photos with a camera or drinking coffee while driving (which is facilitated by cup-holders).

Constraints

Once we have identified the jobs-to-be-done, there are usually two ways that we can get more detailed information about the jobs. We can focus on constraints that prevent jobs from being done well and outcomes which that are desired from jobs. It is often helpful to prioritize jobs to be done in terms of how many people need to do these jobs, how important they are and how dissatisfied people are (each of which can be measured using standard questions in a questionnaire (e.g., `How important...`). If a job is common, important and people are dissatisfied with the products they currently use for the job, then the job presents more of an opportunity. For example:



A constraint is something that prevents a job from being done or adversely impacts the outcome. The identification of constraints can be particularly helpful with new product development. When we focus on what consumers want, we are leaving it to the consumer to tell us what innovations they desire and they often do not have the insight or understanding of possibilities to be able to do this. Think about the coffee market five years ago. Research continued to show the important emotional and social role of coffee. The true aficionados wanted to be baristas in the comfort of their own homes. And what new products were targeted at this segment? Mini industrial coffee makers, designed to replicate the gourmet café experience, but with an emphasis on simplicity and ease of cleaning.

The Nespresso machine has been the standout success in this segment in recent times. Interestingly, it fails in terms of most of the key needs were identified if research into coffee before the Nespresso was launched. The user is nothing like a barista; they just drop a pod into a machine and press a button. How could such a functional product meet the market's aspirations? It does so because it deals with all the constraints of the coffee making jobs. It is very fast. It is easy to clean. People can have much greater control over getting their perfect brew, with different varieties of coffee for every person. They can order online.

Outcomes

It can be useful to break down a job into a series of outcomes. An outcome is an assessment of how well a job has been done. For example, these outcomes relate to the 'satisfy hunger' job performed by yogurt:

Decrease price / cost	**Increase enjoyment**
Decrease time taken to plan meal	Decrease guilt
Decrease time taken to shop	Decrease saturated fats
Decrease serving difficulty	Decrease calories
Decrease consumption difficulty	Decrease food intolerance
Decrease time to prepare	Increase ease of eating 'on the go'
Decrease time to clean up	Increase environmental sustainability
Increase time until I become hungry next	

Identifying and measuring jobs-to-be-done, constraints and outcomes

Jobs, constraints and outcomes are relatively easy to identify through any or all of qualitative research, introspection, discussions with stakeholders and reviewing previous studies. Various Ideation techniques, such as SCAMPER, can be useful.

If an outcome is both important and consumers believe that current products do badly on the outcome, this suggests that there is a gap in the market relating to this opportunity. One method for quantifying the resulting opportunities is to:

1. Get consumers to rate the importance of each outcome on a scale of 1 to 10.
2. Get consumers to rate their current level of satisfaction with the products they currently buy in terms of addressing these outcomes on a scale of 1 to 10.
3. Prioritize the opportunities with the highest score, where the score is computed as:

```
Opportunity = 2 × Importance - Satisfaction
```

There is no great weight of scientific evidence behind this formula, but it is a useful place to start.

Concept Test

A concept test is used for measuring and understanding the appeal of a new product.

The concept board

Concept testing generally involves creating concept boards, which explain the new product to consumers, giving them enough information so that they can have a meaningful reaction when asked if they will or will not buy it. For example, this concept board was used to test the appeal for Magnum Gold prior to its launch.

Copyright Unilever Australia (used with permission)

The success of a concept test hinges on the quality of the concept's description. It is not unknown for 20 minutes to be spent educating the respondent about a concept, but in most studies consumers get 10 to 20 seconds (which is likely sufficient for simple concepts).

If a concept board is insufficient to accurately get consumers to evaluate a concept then consumers may also be given an opportunity to trial the product. Most commonly, products will be shipped to consumers' houses (this is called -in-home placement), or, consumers will go to a central location

where the concept testing may also involve some sensory research (research designed to get specific feedback on how to improve the sensory aspects of the food, such as smell, taste and viscosity). When presenting radical new technologies, the process of respondent education is called information acceleration.

The impact of respondent education can be great. One study found that after describing a concept, the proportion of respondents who said they would 'definitely' or 'probably' buy was 59% at $1.59 and 59% at $2.29 (i.e., the price had no effect). However, when the same study was conducted with an in-home use component (i.e., leaving the product with the consumers for a few weeks), price had a large effect, with the stated intention to buy being 64% at the lower price and 48% at the higher price.

If presenting complex concepts to respondents it is advisable to test for comprehension in some way, such as in Q1 of the questionnaire below.

In addition to respondent education, questions should be incentive compatible, and this is harder to achieve. Our concern is not simply one of respondents trying to mislead. There is also the issue of whether respondents will put as much effort in to answering the question as they would in a real-world purchase situation (as if their efforts are different, then their answer in the questionnaire will be unlikely to reflect what they do in reality). Most concept tests use the same structure, although questions are modified and added as required:

Concept testing questionnaire

SHOW CONCEPT BOARD

Please look at the description of the product. Please take as much time as you need to look at it. We will then ask you some questions about the product.

Q1. Is there anything in this description that you do not understand or would need more information about in order to decide whether or not to buy this product? If so, what? OPEN-ENDED

'**Q2.** What, if anything, do you think you would particularly like about this product? OPEN-ENDED

Q3. What, if anything, do you think you would particularly dislike about this product? OPEN-ENDED

Q4. Which phrase from those below best describes how likely you would be to buy the product for yourself?

```
I would definitely buy it I would probably buy it I am not sure whether I would
buy it or not I would probably not buy it I would definitely not buy it
```

Q5. On average, how often do you think it would be eaten?

Every day Two or three times a week Once a week Once every two/three weeks Once a month Once every two/three months Once every four/six months Less often than every six months Never

Q6. Compared with similar products how different do you think this product is?

Very different Different A little different Not very different Not at all different

Q7. How well do you think the idea for this particular snack fits what INSERT BRAND NAME means to you?s not fit at all
1
2
3
4
5
6
7
8
9
10 Fits extremely well

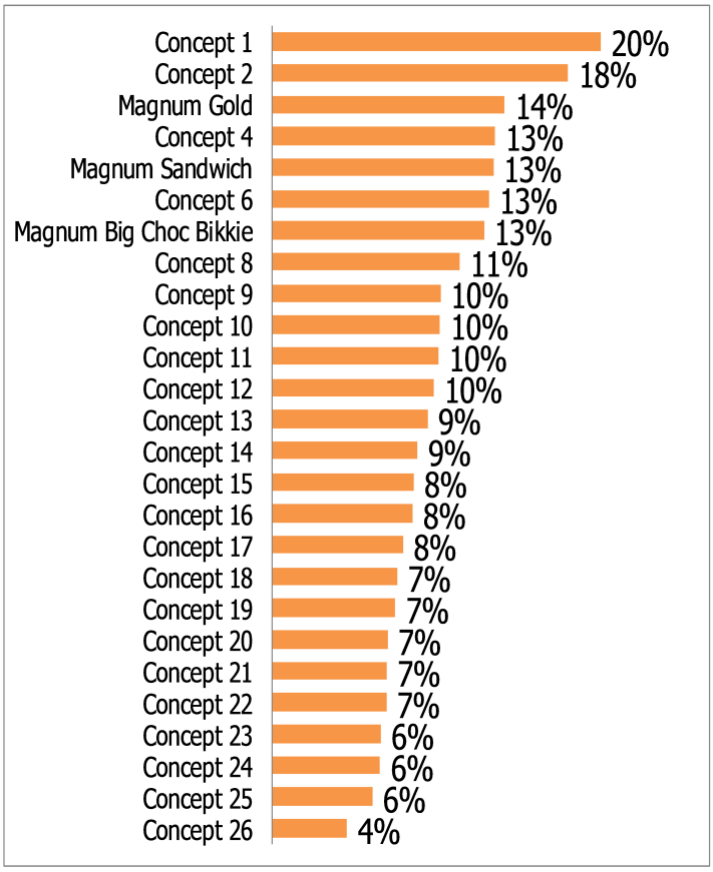
Q8. Do you have any further comments about this product? If so, what? OPEN-ENDED

Qualitative concept testing

Where the concepts are being tested using qualitative research, a similar structure is usually followed, although commonly with a general exploration of needs and wants at the beginning of the study (this can also occur with quantitative studies). Where the purpose of the concept test is to refine concepts, rather than work out which are most appealing, qual is usually more appropriate due to the potential for more depth from the diagnostics.

Analysis

Typically a concept test will involve multiple separate concepts and concepts are evaluated based on a combination of relative performance against each of the other concepts and internal benchmarks. Usually the key question that is used to evaluate whether the concept is a winner is Q4, which is referred to as the purchase intent question. The Magnum concept shown above was one of 26 concepts tested. The resulting concept scores were:



Large companies typically keep records of previous concept tests and develop benchmarks for new products. Common benchmarks include getting a top two box score of 60%, or a top box score, which is also referred to as a DWB ("Definitely Would Buy") of 25%. Such benchmarks are rarely valid.[3] This is because people are not able to accurately state their future behavior (how can they know how much they will like an ice cream before they have eaten it) and because many other factors determine the success of a new product, such as distribution, word-of-mouth, and so on. For example, Magnum Gold's top box score (the proportion of people to say they would 'definitely' buy) was 14% and its top 2 box score was 38%, which are below most commonly-used thresholds, yet it went on to become the top selling ice cream in Australia. And, many other products have passed these thresholds and failed.

Although it is not unusual for companies to take the approach of launching product(s) that 'win' the concept testing, this can be dangerous as consumers have a tendency to give higher ratings to concepts that are similar to successful products already in the market. Consequently, the products with the highest purchase intent often face the stiffest competition if launched. One solution to this is to prioritise concepts based on a combination of purchase intention and uniqueness (e.g., as measured in Q6).

Choice Task

The purchase intention questions in a traditional Concept Test likely measure desire rather than true intention. Desire is infinite and resources are finite and we should expect that purchase intention questions will over-state demand. A solution to this problem is to ask consumers to choose from multiple products. Such choice tasks have the important advantage of most closely replicating the real world tasks that shoppers perform. For example:

If you saw the following products at these prices when you were next shopping for babyfood, which of these would you buy?

Most commonly, consumers will be presented with a single screen (or piece of paper) containing descriptions of the new product and competitor products in the market segment and their prices. Consumers are asked to select only one of the brands. As they have had to make a choice between all of the available alternatives, the data implicitly takes into account the finite nature of consumers' resources, and we expect (and often observe) a high correspondence between the proportion of respondents who select a product and the true proportion who buy it in the real world.

Of course, there is something a bit unnatural about such choice tasks, containing one new product and existing products. The poorer the correspondence between the real-world and a question being used to predict real-world behavior, the less accurate we should expect the question to be. This is referred to as *ecological validity*. A solution of sorts to this is to Choice Modeling, as with a choice model all of the alternatives are artificial and thus the problem of one being new is not a source of bias.

Where consumers are likely to change their purchasing in different situations (e.g., meal occasions), or, due to variety seeking, these can be addressed by occasion-cuing the questions (e.g., “thinking about your last trip”, or, “thinking about your last time you went to the football”) or by using Constant-Sum questions (e.g., “thinking of your next 10 purchases, how many of each of these products would you buy?”).

Many choices tasks include an “I would not buy any of these” option. Such descriptions should only ever be included when it is a realistic outcome. As an example, if modelling airline choice, it is clearly realistic to permit a consumer to say they would choose not to fly given the available alternatives. By contrast, if modelling choice of beverage on a hot day, the inclusion of the “none” option is problematic, as in reality the consumer is likely to be forced to choose between the brands in the choice task and may select “none” as a way of signalling that they do not like these brands (even though if presented with only these brands they probably would have actually chosen one of them).

Experiments

Discrete Choice Experiments

There is a relatively simple approach to dealing with the problem of pricing the iPad which was discussed in Basic Questionnaires. The solution is to conduct an experiment, showing different people the iPad at different prices and asking them if they will or will not buy the product. For example, some people could be shown a price of $500, some $750, some $1,000 and so on. As the respondent cannot know that the experiment is designed to understand the effect of price, they have no way of providing misleading answers in order to gain a benefit and thus the question becomes incentive compatible.

Common types of experiments used in market research

- Completely Randomized Designs. Each respondent is shown one and only one thing (e.g., the example above).
- Max-Diff
- Conjoint Analysis
- Discrete Choice Experiments

Choice Modeling

A common problem when developing a new product is that you have a general view on what the new product should be but need to refine some of the details, such as the price point, the size or the specific combination of features to offer. Or, you may have an existing product or service and wish to work out how best to improve it. One solution to test multiple versions of a product is to conduct multiple concept tests or choice tasks. The alternative is to use choice modeling, which is essentially the same as choice tasks, except that each respondent completes multiple choice tasks and the descriptions of the alternatives change from task to task.

The basic premise of choice modelling is that rather than researching preferences for concepts, we should instead research preferences for product attributes.

The theoretical assumptions of choice modeling

1. Each product product can be described by its attributes

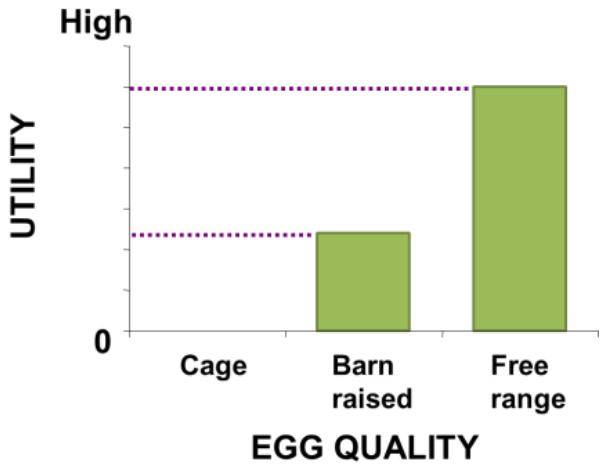
The first key assumption of choice modeling is that products can be described by their attributes. The table below shows attributes in the packaged eggs market. It describes packaged eggs in terms of seven attributes. Each attribute consists of a series of attribute levels. We can see that the attribute Weight has four levels: 55g, 60g, 65g and 70g.

Attribute	Level 1	Level 2	Level 3	Level 4
Weight	Average Egg Weighs 55g	Average Egg Weights 60g	Average Egg Weights 65g	Average Egg Weights 70g
Organic	BLANK [nothing shown]	Antibiotic and hormone free		
Charity	BLANK [nothing shown]	10% of Revenue donated to RSPCA		
Quality	Fresh Eggs (Caged)	Barn Raised	Free Range	

Uniformity	All eggs appear the same	Some eggs appear different (e.g., shell colour)		
Feed	BLANK [nothing shown]	Fed on grain and fish (high in Omega)	Fed on vegetables	
Price	Range of price points from $3.35 to $6.50			

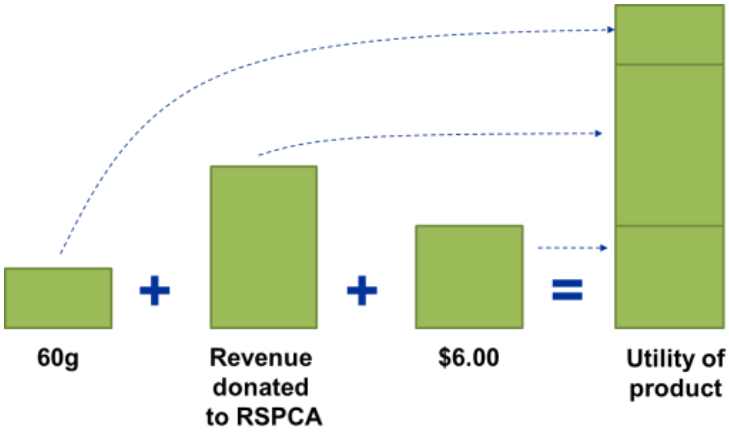
2. Attribute levels differ in their appeal

The second key assumption of choice modeling is that different attribute levels have different levels of appeal. Choice modelling studies tend to use the word utility instead of appeal, but they mean the same thing. For example, with the Egg Quality attribute, consumers may have a utility for caged eggs, barn raised eggs and free range eggs, and a purpose of the choice modelling is to estimate these utilities. Choice modeling studies only calculate the relative utility of different attribute levels. For example, we never estimate the actual appeal of free range eggs; rather, we estimate the appeal of free range eggs relative to some other attribute level, such as caged eggs or barn raised eggs. For this reason, we set one of the levels as having a utility of 0, and then the utilities of the other attribute levels are estimated relative to this attribute's level. Most commonly, the least appealing attribute level is chosen to have a utility of 0.



3. The appeal of a product is the sum of the appeal of the attribute levels

The third key assumption is that the appeal of a product is the sum of the utility of its product attribute levels. This is an example of the use of a decomposition.



4. People are most likely to choose the alternative with the highest appeal

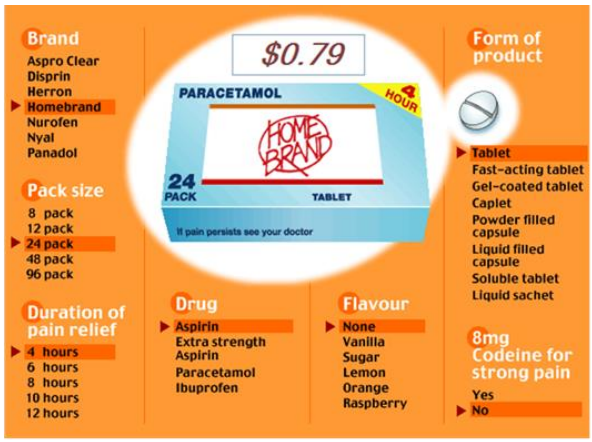
The final key assumption of choice modeling is that people choose, or are most likely to choose, the product with the highest utility. This assumption can be weakened to deal with differences in distribution.

Choice modeling methods

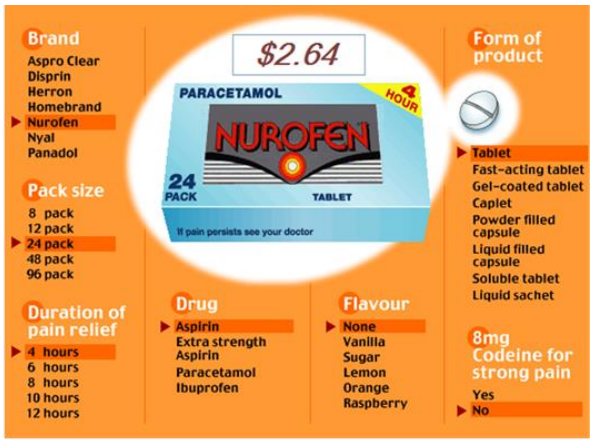
A large number of techniques have been developed for choice modeling.

Build-Your-Own

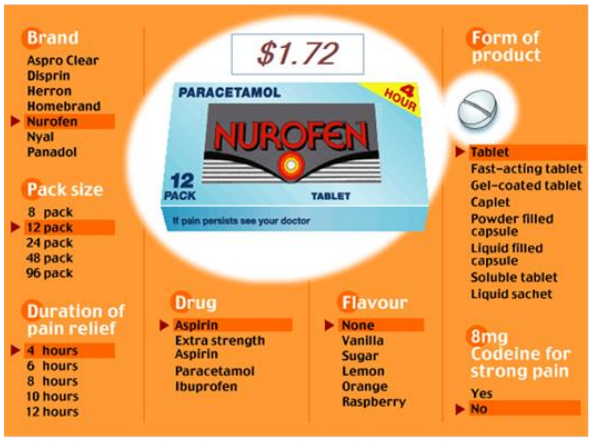
Build-Your-Own (BYO) choice models ask consumers to build their own ideal product. The example below shows four screens from an analgesics study. It shows how a respondent changed the different aspects of the design of a pack of analgesics (e.g., by first clicking on Nurofen). When the respondent changes the product design, the price also changes. Consequently, the respondent builds their ideal product by trading off between the features they desire and the cost of purchase. This is variously known as BYO (Build Your Own), menu-based choice modeling (although this can mean other things), SMART and SIMALTO.



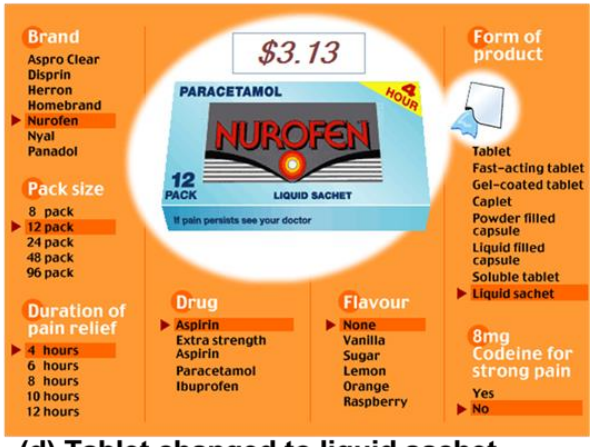
(a) Initial screen



(b) Homebrand changed to Nurofen



(c) 24-pack changed to 12-pack



(d) Tablet changed to liquid sachet

Conjoint Analysis

Conjoint analysis is a Choice Modeling technique which involves presenting people with a series of hypothetical products, called profiles, and asks the people to rank or rate the products in order of appeal. The basic idea is that by looking at which alternatives get a higher rating (or ranking), we can understand what is important to them. For example, if somebody always rates the alternative with the lowest price the highest we can conclude that price is most important to them.

The table below contains descriptions of sixteen different soft drinks. Please rank them according to how likely you would be to buy them, writing a "1" next to your most preferred product, a "2" next to your second most preferred product, and so on until you have given ranks of "1" to "16" to the products.

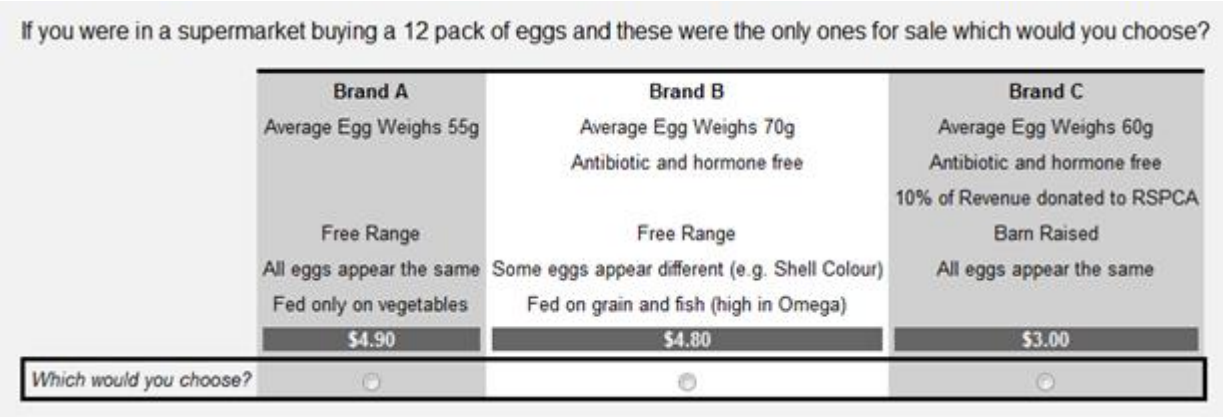
Product	Price	Brand	Size	Material	Flavour	Competition	Rank
1	$2	Virgin	300 ml	Glass	Cola	>No	_____
2	$2	Virgin	300 ml	>Glass	Diet Cola	Yes	_____
3	$2	Virgin	>300 ml	Plastic	Cola	Yes	_____
4	$2	Virgin	300 ml	Plastic	>Diet Cola	No	_____
5	$2	Coke	400 ml	Glass	Cola	No	_____
6	$2	Coke	400 ml	Glass	Diet Cola	Yes	_____
>7	$2	Coke	">400 ml	Plastic	>Cola	Yes	_____
8	$2	Coke	400 ml	Plastic	Diet Cola	No	_____
9	$3	Virgin	400 ml	Glass	Cola	No	_____
10	>$3<	Virgin	400 ml	Glass	Diet Cola	Yes	_____
11	$3	Virgin	400 ml	Plastic	Cola	Yes	_____

12	$3	Virgin	400 ml	Plastic	Diet Cola	No	_____
13	$3	Coke	300 ml	Glass	Cola	No	_____
14	$3	Coke	300 ml	Glass	Diet Cola	Yes	_____
15	$3	Coke	300 ml	Plastic	Cola	Yes	_____
16	$3	Coke	300 ml	Plastic	Diet Cola	No	_____

Discrete Choice Experiments

The most popular form of choice modeling are discrete choice experiments (also known as choice-based conjoint, CBC and sometimes just as choice modeling).

Discrete choice experiments present people with a series of hypothetical purchase situations, and require people to make a choice in each. Typically, consumers are given from six to twenty different Choice Tasks, with each question containing from two to eight alternatives. For example, the choice task below is from the egg study described earlier. In this experiment each respondent completed eight choice tasks (also known as choice sets, choice questions and scenarios), where the attribute levels of each alternative changed from question to question.



The basic idea is that by looking at which alternatives people choose, we can understand what is important to them. For example, if somebody always chooses the alternative with the lowest price we can conclude that price is most important to them.

Decompositions

A decomposition involves breaking something apart. It is a useful way to estimate many things. Let us start with a simple estimation problem. How many Japanese people are there in Australia with dentures (fake teeth that can be taken in and out). And yes, this was was a real-world consulting project.

The simplest and laziest approach to this problem is to do a survey. For example, you might email 1,000 people in Australia and ask how many people in each household are Japanese and have dentures. The result of this would be an estimate of the proportion of Australians that say they are Japanese and have dentures. To get to our required result we would multiply this by the number of Australians. Thus, we have used the following model:

```
Number Proportion Japanese = Japanese × Number dentures dentures Australians
```

This formula is a decomposition. We have decomposed the thing we are trying to estimate into two separate estimates, the proportion and the population size.

As research designs go this is a poor one. The proportion we are trying to estimate is likely to be a very small one (e.g., less than 0.1%). Thus, we would expect that we might need to interview well over 1,000 people before we identified a single one of them and perhaps 50,000 or 100,000 people before we obtained a sufficiently precise estimate, for a cost of millions dollars. People researching bizarre topics are usually short of funds, so we can feel confident that this research design is inappropriate. Furthermore, we probably would not get a very precise answer anyway, as Japanese speakers are relatively less likely to participate. So, how can we resolve this? There are lots of other possible decompositions. For example:

```
Number Number Japanese =  people with  ×  Proporton Dentures dentures Japanese
```

This decomposition requires two completely different inputs: the number of people with dentures and the proportion of people that are Japanese. Both of these numbers may be available from publically

available sources (e.g., trade associations, government statistics), so we might be able to do this very cheaply. Of course, the result may also be quite inaccurate, because this decomposition implicitly assumes that people of Japanese origin are neither more nor less likely to have dentures than the rest of the population.

So, we can use different decompositions to solve the same problem. The trick is to trade-off which will be cheapest with which will be most accurate (i.e., most precise).

Now let us solve a more traditional problem. Consider the problem of trying to predict sales of a new brand of laundry detergent. A standard decomposition for this is:

```
Sales = Market Share × Market Size
```

There are lots and lots of ways to estimate market share. One of them is to present people with a screen showing a picture of a supermarket shelf, including the new brand, and ask people to choose one; the proportion of people that choose the new brand is then an estimate of the market share. The market size can usually be obtained from historic sales data.

An alternative decomposition of sales of a new product is:

```
Intention Sales = to × Purchase × Population purchase frequency size
```

Intention to purchase is estimated by showing somebody a picture of the proposed new product, and asking them if they will buy it or not (Concept Testing). Purchase frequency can be estimated by asking people how many times they will buy it (although looking at historical purchase rates of similar products will often be more valuable). The population size can be obtained from government statistical agencies.

When designing research we need to find the most cost-effective way of producing sufficiently precise estimates. Consider the decomposition of:

```
Sales = Market Share × Market Size
```

Let us say we are trying to forecast laundry detergent and we are trying to produce a forecast for next year. To produce our sales estimate we need to estimate market share and market size. It is inevitable that the market size will be broadly similar to the sales from the previous year, with a little growth. So, if the market size last year was $13 billion, the market size next year will probably be between $13 billion and $14 billion.

Now think about the market share estimate. If the new product is a 'dog', it will get 0% market share. If it is wonderful, perhaps it can get 20% of the market. So, if we multiply the lower bound estimates of market size and market share we compute a lower bound estimate sales of 0% of $13 billion = $0 and

an upper bound of 20% of $14 billion = $2.8 billion. Our range of estimates for market size make comparatively little difference to our forecast. If we assume that the market size is $13 billion, this drops the upper bound from $2.8 billion to $2.6 billion. By contrast, changing the estimated market share from 20% to 0% drops the upper bound all the way down to $0. Thus, the estimate of sales is most sensitive to the estimated market share analysis and, it follows from this sensitivity analysis, that our focus in producing an estimate of sales should be on estimating the market share if we are decomposing sales into market share and market size. This process, of working out which bit of the research design will most impact upon the precision of an estimate, is known as sensitivity analysis.

Pre-launch Forecasting

When we predict a product's sales before it is launched we are doing a pre-launch forecast. As a general rule, pre-launch forecasts are wrong, often massively wrong. Nevertheless, the alternative to having no pre-launch forecast is to have nothing at all, so pre-launch forecasting is popular and useful.

As discussed earlier, there are two main decompositions used in prelaunch forecasting:

```
Sales = Market Share × Market Size
```

and

```
Intention Sales =  to × Purchase × Population purchase frequency size
```

The first of these decompositions can be combined with choice modeling, as choice modeling is particularly popular in that the preference share estimate that is provided can be interpreted as an estimate of market share. Thus, as most companies have a good idea of the market size, it becomes easy to use this decomposition to predict sales.

The second of the decompositions seems sensible on face value. However, it tends to produce forecasts of sales that are much too high. This is because people are unable to gauge either their likelihood to purchase or their purchase frequency. If you think about it for a while, you will realize that the questions themselves are nearly impossible to answer. How can you rate your purchase intent, for example, if you do not know what other products will be available?

Some research companies have methods for calibrating the purchase intention and purchase frequency data with the goal of making the decomposition more accurate. These are the rules of six companies:[4]

- Company A assumed that 100% of those who say they “Definitely will buy” actually will buy (and that nobody else will buy).
- Company B assumed that 28% of those who say they “Definitely will buy” actually will buy.
- Company C assumed that 80% of those who say they “Definitely will buy” actually will buy and that 20% of those who say “probably” will buy.
- Company D assumed that 96% of those who say they “Definitely will buy” actually will buy and that 36% of those who say “probably” will buy.
- Company E assumed that the buying proportions for each of the five categories are 70%, 54%, 35%, 24% and 20%.
- Company F assumed that the buying proportions for each of the five categories are 75%, 25%, 10%, 5% and 2%.

A key point to appreciate about these different rules of thumb is the wide discrepancy in their predictions. This highlights the poor validity of purchase intention questions. Furthermore, all of the methods down-weight the purchase intentions, which highlights that purchase intentions are systematically over-stated.

Another problem with the calibration of purchase intent is that the extent of calibration required will likely relate to the number of competitors and the uniqueness of the product, suggesting that any simple approach to calibration is unlikely to be successful.

Despite the inaccuracy of questions asking purchase intentions, concept tests are very popular. In part this is because the diagnostics they provide (e.g., likes and dislikes) can be highly informative and in part it is because they are simple. Furthermore, they can be used to produce forecasts that are more valid than those of the decomposition. The trick is to study the results of previous concept tests. If, for example, you find that a top box score of 20% results in sales of $10,000,000 on average while 30% results in sales of $15,000,000 then you can estimate that a score of 25% will lead to sales of $12,500,000.

Researchers can often use substantially more complex decompositions than these for producing pre-launch forecasts. For example, some take into account factors like distribution, advertising expenditure, shelf placement and word-of-mouth.

Once a pre-launch forecast has been computed it should then be substituted into an economic analysis which takes into account the extent of cannibalization and the likely profit margin of the product.

References

[1] Rossiter, John R. and Gary L. Lilien (1994), "New 'Brainstorming' Principles," Australian Journal of Management, 19 (June), 61-72.

[2] Bono, Edward de (2008), Creativity Workout: 62 Exercises to Unlock Your Most Creative Ideas. Berkeley: Ulysses Press, page 30-31.

[3] For a case study see: Bass, Frank M, Kent Gordon, Teresa L Ferguson, and Mary Lou Githens (2001), "DIRECTV: Forecasting diffusion of a new technology prior to product launch," Interfaces, 31 (3), S82-S93.

[4] Jamieson, Linda F. and Frank M. Bass (1989), "Adjusting Stated Intention Measures to Predict Trial Purchase of New Products: A Comparison of Models and Methods," Journal of Marketing Research, 26 (August), 336-45.

Data Preparation

- Basic Process
- Getting a Data File
- What a Data File Looks Like
- Correcting Metadata
- Missing Values
- Merging Categories and Creating NETs
- Recoding Variables
- Multiple Response Data Formats

Data Preparation

Cleaning, tidying and weighting are activities that are performed before trying to work out what the data in a survey means:

- Data cleaning refers to checking and correcting anomalies in a data file. The goal is to identify data that is, in some way, clearly incorrect. The unglamorous world of data cleaning can be a key determinant of the quality of data analysis, particularly when the data is from a messy source (e.g., customer records, collected using a cheap data collection program).
- Data tidying involves manipulating the way that data is set up to make it easier to interpret. For example, changing birth dates into age categories, removing 'don't know' categories.
- Weighting is a technique which adjusts the results of a survey to bring them into line with what is known about the population.

These steps are also sometimes referred to as data processing

It is possible to first clean the data, then tidy the data and, then, if necessary, weight the data. However, in practice it is much more efficient to simultaneously clean and tidy the data and then weight the data.

Basic process

Choosing Survey Analysis Software

The easiest option for analyzing survey data is usually to use the analysis tools that come for free in the Data Collection Software that has been used to collect the data. Where only rudimentary analysis is required, such as working out the proportion of people to select each option, this is usually the best approach. However, if there is a need to share reporting with others, to fix problems with the data or to conduct more sophisticated analyses, it is generally necessary to use software specifically designed for the analysis of survey data.

General-purpose software for analyzing surveys

Most people that analyze surveys do so using general-purpose survey analysis software. Some of the main programs for doing this are described in the table below. The next table provides an overview of more specialized types of software used in survey analysis.

The best way to evaluate whether the software is likely to be useful is via having a trial. This link takes you to tutorials which you can use to evaluate the different programs.

Program	Typical user	Key features	Key limitations
Displayr[note 1]	Little or no experience in analyzing surveys	Ease of use and sharing resultsAdvanced analysis	Creation of complex crosstabs is less flexible than in WinCross, Survey Reporter and MarketSightDoes not support Tracking studies
MarketSight	Market research consultants and the internal research departments of large companies	Easy to learn/useQuickly create large numbers of crosstabs	No advanced analysis
Q[note 2]	Market research consultants and the internal research departments of large companies	Easier to learn and use than SPSSAdvanced analysis Weighting Automation (e.g., exporting to office, updating trackers)	Creation of complex crosstabs is less flexible than in WinCross, Survey Reporter and MarketSight
Survey Reporter	Market research consultants and the internal research departments of large companies	Well-integrated with SPSS data collection products	No Advanced analysisAssumes you have access to a DP department that uses SPSS data collection products

WinCross	Market research consultants	Efficient and flexible creation of crosstabs	No advanced analysisRequires the user to do 'programming'
SPSS Statistics	Market research consultants, internal research departments, statisticians and 'advanced quant' teams	More than 10 times as many features as the programs listed above	Difficult-to-learn and generally less efficient to use than the other programs listed above
R	Statisticians and 'advanced quant' teams	AutomationCustomization More features than any of the other programs on this table	Poor at creating crosstabsAlmost everything has to be done by 'programming' which makes it very slow to use relative to the other programs

Overview of specialty survey analysis software

Types of data analysis software	General-purpose survey analysis software	Data Processing Software	Variance Estimation Software	General-purpose statistical software	Data collection software that can can conduct basic analyses
Typical users	Consultants, marketers, market researchers	Data Processors (DP) people that create large numbers of crosstabs for their clients (where the clients are	Statisticians	Statisticians and similar data analysis professionals	Students, industry associations, HR, small businesses conducting their own research

		typically market research consultancies)			
Examples	DisplayrMarket Sight Q SPSS Custom Tables SPSS Survey Reporter Wincross	MRDC Quantum SurveyCraft Uncle	AM Software Bascula CENVAR CLUSTERS Epi Info IVEware Generalized Estimation System (GES) Rpackages: survey and pps SAS Stata SUDAAN VPLX WesVarPC	SPSS StatisticsStatistica R Stata SAS	Confirmit Insightify Polldaddy Obsurvey Qualtrics QuestionPro Snap Surveys Super Simple Survey Survey Gizmo SurveyMonkey Toluna QuickSurveys Wufoo Survey Crafter

Getting a Data File

A data file contains the individual responses to a survey in a format that permits them to be analyzed by a program specifically designed for the analysis of survey data (e.g., SPSS, Q, Displayr, Stata). Almost all programs that are used to conduct surveys are able to export data files.

In much the same way that some meals are great while others are inedible, some data files are great and some are unusable. The quality of a data file is perhaps the biggest cause of problems experienced by people when learning how to analyze surveys, as a common mistake is to put

insufficient effort into obtaining 'good' data which results in the analysis being much harder than it needs to be.

What survey data looks like

Raw data

When we conduct market research we usually (but not always) collect data about individual people, households or businesses.The data provided by respondents is called the raw data. The table below shows raw data for ten households (from a larger data file). Each row in a table of raw data represents the data from an individual respondent, and there are no blank rows. In this case, each respondent was a household. Each column is referred to as a variable. Each variable is a measure of some characteristic of the respondents.

ID	CARRIER	INCOME	MOVES	AGE	EDUCATION	EMPLOYMENT	NON-USAGE	Q1a	Q1b	Q1c
1	2	1	1	3	2	1	9	1	1	1
2	1	4	3	2	2	1	2	1	1	1
3	2	NA	1	NA	NA	NA	6	1	1	1
4	2	NA	3	6	1	5	7	1	1	1
5	2	NA	1	6	2	4	0	1	1	1
6	2	NA	1	6	NA	NA	0	0	1	0
7	2	3	1	4	1	1	3	1	1	1
8	2	2	1	5	2	5	1	1	1	1
9	2	NA	1	5	1	1	0	1	1	1
10	2	3	1	4	1	4	2	1	1	0

Metadata

Data such as that above is not, on its own, readily interpretable. To interpret such data it is necessary to also have metadata, explaining what it all means. The metadata, which is sometimes referred to as a data dictionary, is shown below. So, returning to the table above, the database indicates that the fourth household was not a customer of AT&T, there is no data indicating the household's income, the household moved twice in the last 10 years, the respondent who completed the survey was 65 or older, and so on.

Where a variable is categorical this means that the values stored in the raw data can only be interpreted by looking at the metadata. In particular, with the MOVES variable, a 1 indicates that a household has not moved, a 2 indicates it has moved once, etc. By contrast, with the USAGE variable, which is numeric, a 1 indicates it was used once, a 2 indicates it was used twice, etc.

Variable	Variable Label	Value Labels	Variable Type
ID	A unique identification number assigned to each respondent	1 = first respondent, 2 = second respondent,...	Categorical
CARRIER	Phone carrier of household	1 = AT&T, 2 = Other	Categorical
INCOME	Household income bracket (in thousands)	1 = <7.5, 2 = 7.5-15, 3 = 15-25, 4 = 25-35, 5 = 35-45, 6 = 45-75, 7 = >75	Categorical
MOVES	Number of times the household has moved in the preceding 10 years	1 = 0, 2 = 1,3 = 2, ..., 8 = 7,11 = >10	Categorical
AGE	Age of the respondent	1 = 18-24,2 = 25-34, 3 = 35-44, 4 = 45-54, 5 = 55-64, 6 = 65+	Categorical
EDUCATION	The highest level of education achieved	1 = Did not finish school, 2 = High School; 3 = College, 4 = Postgraduate	Categorical

EMPLOYMENT	Employment status of the respondent	1 = Full-time, 2 = Part-time; 3 = Student; 4 = At home; 5 = Retired; 6 = Unemployed	Categorical
USAGE	The typical monthly number of longdistance telephone calls by the household		Numeric
Q1a	Aware: AT&T	0 = No, 1 = Yes	Categorical
Q1b	Aware: Verizon	0 = No, 1 = Yes	Categorical
Q1c	Aware: CenturyLink	0 = No, 1 = Yes	Categorical

This table shows the minimal metadata necessary to analyze a survey. However, better data files will contain more information. In particular:

- Question Type. For example, note that the last three variables, Q1a, Q1b and Q1c are related and form a part of a single question (which asked people which of the companies they had heard of); a good data file will contain metadata showing that these are linked together.
- Versioning. For example, changes to question wording that occurred during the data collection process and different translations.

A good data file will contain both the raw data and the metadata together in a single file. If you have two files, one which contains the raw data and another which contains the metadata, then you do not actually have a 'data file', you instead have the material you need to create a data file, but still have to create it. Many data analysis programs will provide tools that allow you to import the raw data and then enter the metadata but it will generally need to be done manually (i.e., by retyping it or cutting and pasting each field of information); this is a time consuming and error-prone process which should be avoided where possible.

Data files formats

Data collection programs export data files in a specific format. Most programs provide multiple formats for exporting, but these formats can differ markedly in terms of their usefulness.

Text files

The simplest data files are called 'text data files'. It is generally a very bad idea to obtain the data from a survey as a text file. This is because when data is obtained as a text file there will be one of two problems:

1. It will either contain no metadata, which makes it at best difficult to analyze and at worst impossible (e.g., if you do not know that a value of 2 represents an age of 25-34, then there is no way to interpret the data).
2. It will contain text instead of numbers for all the data. Initially this may appear to be useful, but in practice is a massive problem, as:

    - Most programs for the analysis of survey data do not permit you to do analysis with data in this format, and so you will read the data into the program and then discover that you either cannot do even the most basic analysis, or, need to spend a lot of time re-formatting the data to make it useful.

    - Many of the important features of the survey will not be evident in the data file. For example, if you have asked a question getting people to give ratings from 0 to 10, when you create a table in a text file they will be ordered as: 0, 1, 10, 2, 3, ... Similarly, Grid and Multiple Response questions will generally need to be treated as if they were multiple Single Response questions.

ID	CARRIER	INCOME	MOVES	AGE	EDUCATION	EMPLOYMENT	USAGE	Q1a	Q1b	Q1c
1	Other	<7.5	0	35-44	High School	Full-time	9	Yes	Yes	Yes
2	AT&T	45-75	2	25-34	High School	Full-time	2	Yes	Yes	Yes
3	Other	NA	0	NA	NA	NA	6	Yes	Yes	Yes
4	Other	NA	2	65+	Did not finish school	Retired	7	Yes	Yes	Yes
5	Other	NA	0	65+	High School	At Home	0	Yes	Yes	Yes

6	Other	NA	0	65+	NA	NA	0	No	Yes	No
7	Other	25-35	0	45-54	Did not finish school	Full-time	3	Yes	Yes	Yes
8	Other	15-25	0	55-64	High School	Retired	1	Yes	Yes	Yes
9	Other	NA	0	55-64	Did not finish school	Full-time	0	Yes	Yes	Yes
10	Other	25-35	0	45-54	Did not finish school	At Home	2	Yes	Yes	No

CSV Files and Excel files

This is generally the best of the text file formats (although this is very much a case of being the tallest dwarf). It uses a comma to separate each variable.

Tab delimited files

This is similar to a CSV file, except that a tab character is used instead of a comma. Generally, if data is in this format it is appropriate to open it in Excel and then save it as a CSV file.

Fixed width files (ASCII) files

A fixed width file is one where each column of numbers has a specific meaning. For example, in the data below the first column may represent the first variable, the second and third variable together may represent the second variable, and so on. This format was invented because it took up little hard-disk space, which was an important consideration in the 1960s and 1970s. It is rarely used today and is the worst of all of the file formats as it cannot readily be used with Open-Ended questions and most modern programs will not read this file format. Generally, if data is in this format it is appropriate to open it in Excel and then save it as a CSV file.

```
00001 01200 01203
```

Good data files

The good formats

The gold standard data file is an IBM SPSS Data Collection Model data file (also known as a Dimensions, MDT or MDD data file). This file format contains all the different types of metadata. This data file is only created by the top-of-the-range IBM data collection programs and can only be read by IBM data products and a small number of other products (Q and Displayr).

The next-best format is the Triple S format. It is a little more widely used than the IBM SPSS Data Collection Model format, but it is generally only available in the more expensive data collection programs.

The industry standard 'good' file format for data is an SPSS .SAV data file (usually called a 'dot sav' file). This is not quite as good as the other two formats, as it does not contain the versioning information and it only contains very limited Question Type information (it does not support the various Grid type of questions). However, all good data collection programs can export in this format. Refer to SPSS Data File Specifications for details on how these files are best set up.

Occasionally data collection programs will export both a text file and an SPSS .sps file (also known as a syntax file). The syntax file is actually a program which contains instructions for turning the text file into an SPSS .SAV file. SPSS is the only program that can always read these files, but Q can read these in some circumstances.

Appropriate set up in the good formats

Obtaining a good data file is not just a case of specifying the desired format. In particular, in the case of the SPSS .SAV data files, it is quite common to have them created with either incorrect values and incorrect metadata. The most common problems are:

- Incorrect values for options not selected in multiple response questions. That is, the files use the same value (commonly a 0 or a special indicates missing value category usually called SYSMS, NA, or NaN) to indicate that somebody was not asked a question as they use to indicate that somebody did not select an option in a Multiple Response question.
- Labels that have been truncated (e.g., saying Please rate your satisfaction with the following ba), making it impossible to determine what the data means (except by reviewing the questionnaire).

Most survey analysis programs will have some facilities in them to clean such poor data, but it is generally advisable to try and instead obtain a data file that does not contain such problems.

What a Data File Looks Like

There are many ways to lay out data in an Excel-style file, but some are more useful than others. You typically want to have an Excel file which contains a single worksheet, as if you introduce an additional worksheet there is no way for software to work out what the relationship between the worksheets is, or which worksheet contains the data you are trying to analyze. Further differences between good and bad data layout are illustrated below.

Useless: Typical Tables of Survey Results

The term 'data file' is a bit misleading, as files are only useful if they contain data in the appropriate format. For example, the data shown below is largely useless. Yes, it shows some results from the survey. However, it is impossible to use this data to compute many useful things. For example, this data cannot be used to work out if males were more likely to be satisfied than females.

	A	B	C	D	E	F	G	H	I	J
1	**1. How satisfied are you with the our product?**							**2. Are you…**		
2	**Value**	**Count**	**Percent**					**Value**	**Count**	**Percent**
3	Very Satisfied	101	57.4%					Male	165	94.3%
4	Somewhat Satisfied	74	42.0%					Female	10	5.7%
5	Neutral	0	0.0%							
6	Somewhat Unsatisfied	0	0.0%					**Statistics**		
7	Very Unsatisfied	1	0.6%					Total Responses	175	
8										
9	**Statistics**									
10	Total Responses	176								

Useful: Tables Containing One Row Per Respondent

By contrast, the data below is appropriate for data analysis. This table contains:

- A column for each of the different bits of data collected in the survey (e.g., age is in one column, education is in another).
- A single row at the top containing the headings that describe each column.
- A separate row for each respondent (i.e., all the data for the first respondent is shown the row under the headings, the row beneath this contains the data for the second respondent, and so on).

	A	B	C	D	E	F	G	H	I	J	K
1	ID	CARRIER	INCOME	MOVES	AGE	EDUCATION	EMPLOYMENT	USAGE	Q1a	Q1b	Q1c
2	1	Other	<7.5	0	35-44	High School	Full-time	9	Yes	Yes	Yes
3	2	AT&T	45-75	2	25-34	High School	Full-time	2	Yes	Yes	Yes
4	3	Other	NA	0	NA	NA	NA	6	Yes	Yes	Yes
5	4	Other	NA	2	65+	Did not finish school	Retired	7	Yes	Yes	Yes
6	5	Other	NA	0	65+	High School	At Home	0	Yes	Yes	Yes
7	6	Other	NA	0	65+	NA	NA	0	No	Yes	No
8	7	Other	25-35	0	45-54	Did not finish school	Full-time	3	Yes	Yes	Yes
9	8	Other	15-25	0	55-64	High School	Retired	1	Yes	Yes	Yes
10	9	Other	NA	0	55-64	Did not finish school	Full-time	0	Yes	Yes	Yes
11	10	Other	25-35	0	45-54	Did not finish school	At Home	2	Yes	Yes	No
12											

Correcting Metadata

The metadata of a data file is often incorrect and the first step in analyzing data is often to correct these errors. Where there are a substantial amount of errors in metadata it is often the case that the best solution is to try and get a better data file. This page lists the most common problems and their remedies.

Combining variables into multiple response questions

A very common problem with data files is that a question that was asked as a multiple response question appears as multiple single response questions. This is generally caused by the data collection software exporting the data incorrectly (unfortunately, most data collection programs make this mistake). Some programs, such as Q and Displayr, attempt to automatically correct this problem by looking for patterns in the data (e.g., if there are 10 variables, each of which has the same prefix of Awareness with and have the same code frame, then Q and Displayr automatically group the variables together). However, in most programs there is a need to manually group together variables as multiple response questions.

There are two common variants of this problem.

Multiple binary variables

Each category of the question appears as a separate summary table, with two categories (hence the term 'binary'). For example, the following shows a part of a MarketSight Summary Report, which

reveals that a question titled Unaided Awareness has been shown as multiple separate tables, rather than being grouped together. This problem has a number of different causes. One of them is that often the data file does not contain the relevant metadata. The other is that sometimes the information is in the file but the data analysis program does not interpret it properly (in this example, both Q and Displayr do group these variables together as a multiple response question).

All the major analysis programs have in-built tools designed to combine multiple variables together and thereby solve this problem.

Unaided awareness: AAPT/Cellular One	
Sample Size	725
No Count Column %	 668 92%
Yes Count Column %	 57 8%
Unaided awareness: New Tel	
Sample Size	725
No Count Column %	 708 98%
Yes Count Column %	 17 2%
Unaided awareness: One-tel	
Sample Size	725

See How to Combine Variables into Multiple Response Questions and Grids.

Multiple categorical variables

There are multiple tables (or variables) showing all of the different categories. Typically, the first table shows the first response that a respondent selected, the second shows the second response and so on.

All the major analysis programs have in-built tools designed to combine multiple variables together and thereby solve this problem.

Combining variables into grids

In just the way that multiple response questions are sometimes not correctly represented in a data file, requiring the variables to be combined (see the previous problems), often grid questions will also

initially appear split apart. Furthermore, many data analysis programs (e.g., SPSS, MarketSight and R) do not provide support for many types of grids.

Splitting questions

In some situations variables, are represented as grids or multiple response questions when it is more appropriate to instead represent them as separate variables. In particular, Q and Displayr automatically group together variables when they import data and from time-to-time they are over zealous and group together variables that should not be grouped together requiring that they need to be split apart.

See How to Split Questions Into Separate Variables.

Changing variable type

Most of the more modern survey analysis programs (e.g., R, Q, Displayr and MarketSight) use metadata to automate how they produce summary tables. For example, in the following output from MarketSight, the average is shown for IID - Interviewer Identification whereas percentages are shown for the other variables. This is because the metadata indicates that IID - Interviewer Identification is a numeric variable while Does respondent have a mobile phone? is a categorical variable. Thus, when the the metadata show the wrong Variable Type this causes the output to be inappropriate and the remedy is to change the variable type.

	Total
IID - Interviewer Identification	
Sample Size	710
Mean	1,409,497.99
Does respondent have a mobile phone?	
Sample Size	725
Yes Count Column %	 715 99%
No Count Column %	 6 1%
3 Count Column %	 1 0%
5 Count Column %	 3 0%
Work status	
Sample Size	700
Student Count Column %	 211 30%
Home maker Count Column %	 62 9%

See How to Change Variable Type and Question Type.

Changing question type

In addition to allowing the user to change from numeric to categorical and back, Q and Displayr also allow the user to change between different question types. Consider the table from Displayr below on the left. The metadata in this example has treated this table as a multiple response question, but the table on the right is automatically generated instead when the metadata is changed to indicate that the question is, in fact, a grid.

	%
Bureaucratic - AAPT/Cellular One	18% ↓
Bureaucratic - New Tel	14% ↓
Bureaucratic - One-tel	20% ↓
Bureaucratic - Optus	34% ↑
Bureaucratic - Orange (Hutchison)	17% ↓
Bureaucratic - Telstra (Mobile Net)	78% ↑
Bureaucratic - Virgin Mobile	14% ↓
Bureaucratic - Vodafone	21% ↓
Slow service - AAPT/Cellular One	23% ↓
Slow service - New Tel	22% ↓
Slow service - One-tel	38% ↑
Slow service - Optus	38% ↑
Slow service - Orange (Hutchison)	26%
Slow service - Telstra (Mobile Net)	45% ↑
Slow service - Virgin Mobile	20% ↓
Slow service - Vodafone	24% ↓
NET	98% ↑

%	Bureauc-ratic	Slow service	NET
AAPT/Cellular One	18%	23%	36%
New Tel	14% ↓	22% ↑	31%
One-tel	20% ↓	38% ↑	48%
Optus	34%	38%	57%
Orange (Hutchison)	17% ↓	26% ↑	37%
Telstra (Mobile Net)	78% ↑	45% ↓	87% ↑
Virgin Mobile	14%	20%	29%
Vodafone	21%	24%	38%
NET	95%	95%	98%

Question wording

Commonly data files will either refer to questions by their question number (for example, a question showing age data may be referred to as #q43, or some of the wording is shown but it is either messy or truncated (e.g., Please rate your satisf). In most programs it is a straightforward process for modifying these names.

See How to Change the Name of a Question or Variable.

Category wording

The following table, from SPSS, shows a quite common problem, where numbers appear instead of descriptions for some of the categories in the table. The key challenge that this usually presents is working out what the numbers really mean; there is usually no way of working this out directly from

the data (i.e., you need to ask whoever created the data file). Once this is known it is usually quite straightforward to fix the problem.

Occupation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-9	2	.3	.6	.6
	0	5	.7	1.5	2.1
	manager/administrator	33	4.6	9.8	11.9
	professional	85	11.7	25.3	37.2
	Associate professional	48	6.6	14.3	51.5
	Tradesperson and related	33	4.6	9.8	61.3
	clerical, sales and services	85	11.7	25.3	86.6
	production/transport worker	23	3.2	6.8	93.5
	Labourer and related worker	22	3.0	6.5	100.0
	Total	336	46.3	100.0	
Missing	System	389	53.7		
Total		725	100.0		

Missing Values

One of the trickiest aspects of cleaning and tidying a data file relates to missing values, which are also known as missing data. A value is said to be 'missing' where there is no valid data for a particular respondents for a variable. How missing values are treated when cleaning a data file can have a large impact on any conclusions from the analysis.

How programs treat missing values (re-basing)

The tables below have been created by exporting to PowerPoint from Q. Looking first at the table on the left, note that in the bottom row it reports that the NET (which in this case is a total) is 100% and that this corresponds to 718 respondents (**n**). Each of the percentages on this table has been

computed by dividing the number of people to select each option (**n**) by the total sample size. For example, the 5% for Manager/administrator is computed as 33 / 718. The first row on this table shows that 54% (389) of the respondents have missing data, which in this case was because they were students, retired, unemployed or home-makers.

The table on the right, by contrast, has excludes the respondents with missing data and, consequently, its total sample size, as shown in the bottom row of the table and also the **base n** in the footer of the table is 329 (i.e., 718 - 389). Note that in this table all the percentages are different. For example, the 33 Manager/administrators in the sample now correspondents to 10% of the sample.

The table on the right has been re-based; that is, the percentages have been computed with the missing data excluded from the calculations. This is done automatically by most survey analysis programs. However, to do this automatically the programs need to know which categories to treat as being missing. In the case of the table on the left, although the first category has a label which says it is missing data, there is not metadata which explicitly tells the program it is missing data and, consequently, it appears on the table and is analyzed in the same way as all the other categories. By contrast, with the example on the right the program has been told to treat the people with missing data as being actually missing and the percentages are updated accordingly.

	%	n
Missing data	54%	389
Manager/administrator	5%	33
Professional	12%	85
Associate professional	7%	48
Tradesperson and related	5%	33
Clerical, sales and services	12%	85
Production/transport worker	3%	23
Labourer and related worker	3%	22
Don't know/refused	0%	0
NET	100%	718

base n = 718; total n = 725; 7 missing

	%	n
Manager/administrator	10%	33
Professional	26%	85
Associate professional	15%	48
Tradesperson and related	10%	33
Clerical, sales and services	26%	85
Production/transport worker	7%	23
Labourer and related worker	7%	22
Don't know/refused	0%	0
NET	100%	329

base n = 329; total n = 725; 396 missing

It is important to appreciate that neither of the tables is necessarily wrong. Rather, they just have different interpretations. The table on the left shows that 5% of people are Managers/administrators, whereas the table on the right is interpreted as saying that 10% of employed people are Managers/administrators.

See How To Remove a Category and Re-Base a Table for information on how to re-base tables in different programs.

Common missing data situations

As the previous example illustrated, whether or not data should be treated as being missing is a decision that has to be made by the person analyzing the data, as the way that it needs to be treated depends upon context (e.g., in the case above, whether we wanted to compute proportions in the total sample or just among people that were employed).

Often the data will be automatically marked as being missing. In particular, with good quality data files any respondent who has not been asked a question will automatically be classified as having missing data on that question and any analyses will be automatically re-based with that respondent's data removed. Consequently, the main situations where the missing data metadata needs to be modified are:

- Where respondents have data that is marked as missing but who should not be treated as having missing data. Consider variables showing the number of children that people have in different age bands (e.g., the number of children aged under 2, from 2 to 5, etc.) Typically, where a person indicates they have no children they end up not being asked how many children they have in each age group and thus end up having missing values in the data file for the number of children in age group. However, in this instance we actually know that they have 0 children in each age category so the data needs to be changed from missing values to instead show that the data is not missing (see Recoding Variables).

- Where respondents have data that is not marked as missing but should be marked as missing. For example, in a study of voting intentions if 10% of the sample have said "Don't know" it often makes sense to treat these responses as missing data as otherwise the polling results indicate that that preference shares for the different political parties add up to less tan 100%.

See How To Show Missing Values on a Table.

Merging Categories and Creating NETs

The table on the left shows the proportion of people to select each age category in a survey. Often it is useful to merge categories together. The table on the middle shows merged categories. The table on the right shows NETs, which are essentially merged categories that are shown in addition to the

existing categories. Merging and the creation of NETs is widespread in survey analysis as it simplifies the data and the simpler the data the easier it is to interpret.

	%
Less than 18	0% ↓
18 to 24	13%
25 to 29	12%
30 to 34	10%
35 to 39	11%
40 to 44	11%
45 to 49	8%
50 to 54	12%
55 to 64	16% ↑
65 or more	7%
NET	100% ↑

	%
Under 30	25%
30 to 39	21%
40 to 49	19% ↓
50 or more	35% ↑
NET	100% ↑

	%
Less than 18	0% ↓
18 to 24	13%
25 to 29	12%
30 to 34	10%
35 to 39	11%
40 to 44	11%
45 to 49	8%
50 to 54	12%
55 to 64	16% ↑
65 or more	7%
Under 40	46% ↑
40 or more	54% ↑
NET	100% ↑

Some programs, such as Q and Displayr, allow you to modify the data so that whenever you create tables the merging and netting of categories is remembered. Other programs, such as SPSS Custom Tables, Survey Reporter and MarketSight, work by letting you create new combined categories at the time of creating crosstabs, while others require you to instead create new categories as new variables by recoding existing variables (e.g., SPSS, R).

Recoding Variables

Recoding involves substituting the values of a variable with values that are more useful. Consider the following summary table of age. Ordinarily in data analysis age is considered a categorical variable. But, from time-to-time it is useful to treat it as a numeric variable and compute its average. The average age is shown at the bottom of the table as 5.9 which is clearly not the true average age.

	%
Less than 18	0% ↓
18 to 24	13%
25 to 29	12%
30 to 34	10%
35 to 39	11%
40 to 44	11%
45 to 49	8%
50 to 54	12%
55 to 64	16% ↑
65 or more	7%
NET	100% ↑
Average	5.9 ↑

The reason for this curious average age calculation is that categorical variables, such as age, are usually stored in the data file with a 1 for the first category, a 2 for the second category and so on. This is shown below, where on the left we can see the age categories for 13 respondents while on the right we can see the values that are stored in the data file. When the average was computed it was computed using these values (e.g., somebody aged 25 to 29 was given a value of 3 when computing an average).

Labels	Q3. Age
1	25 to 29
2	45 to 49
3	25 to 29
4	25 to 29
5	55 to 64
6	55 to 64
7	50 to 54
8	35 to 39
9	65 or more
10	45 to 49
11	45 to 49
12	55 to 64
13	55 to 64

Values	Q3. Age
1	3.00
2	7.00
3	3.00
4	3.00
5	9.00
6	9.00
7	8.00
8	5.00
9	10.00
10	7.00
11	7.00
12	9.00
13	9.00

The left example below shows the original values in the data file. The values on the right show the values after recoding. This is an example of mid-point recoding, where the value assigned is the value in the middle of the range (e.g., 27 is in the middle of 25 to 29).

Label	Value
Less than 18	1
18 to 24	2
25 to 29	3
30 to 34	4
35 to 39	5
40 to 44	6
45 to 49	7
50 to 54	8
55 to 64	9
65 or more	10

Label	Value
Less than 18	9
18 to 24	21
25 to 29	27
30 to 34	32
35 to 39	37
40 to 44	42
45 to 49	47
50 to 54	52
55 to 64	60
65 or more	75

The recoded numeric variable is shown on the right below, and we can see from the table on the left that the average is now a much more sensible 42.5.

	%
Less than 18	0%
18 to 24	13%
25 to 29	12%
30 to 34	10%
35 to 39	11%
40 to 44	11%
45 to 49	8%
50 to 54	12%
55 to 64	16%
65 or more	7%
NET	100%
Average	42.5

Labels	Q3. Age
1	25 to 29
2	45 to 49
3	25 to 29
4	25 to 29
5	55 to 64
6	55 to 64
7	50 to 54
8	35 to 39
9	65 or more
10	45 to 49
11	45 to 49
12	55 to 64
13	55 to 64

Values	Q3. Age
1	27.00
2	47.00
3	27.00
4	27.00
5	60.00
6	60.00
7	52.00
8	37.00
9	75.00
10	47.00
11	47.00
12	60.00
13	60.00

Coding Text Variables

Coding refers to the process of creating a categorical variable

Text data

The following table shows the first ten responses to an open-ended question asking people why, in their opinion, people preferred different brands of colas.

People who drink Pepsi Max probably prefer sugar - free drinks
Big or large people drink Pepsi light and normal size people drink Pepsi Max
I prefer the flavour of Diet Coke to Pepsi and that is what I base my decision on
????????????????
Probably the diet or light colas appeal to those who want to reduce their sugar intake, which traditionally I suppose, is women. I found that once I stared to drink the light colas, I really disliked the taste of the other 'full strength' ones. If Coke or Pepsi are on special, I will buy the cheaper, because after the first drink, I adjust to the differences between the brands, and I like both.
Pepsi is full of sugar and Im a diabetic....
Pepsi gives the impression of being a "home brand" type drink and coke zero is used by people who want to impress
traditionalists versus rebels
health conscious
People who drink Coke Zero are health concious

Developing a code frame

Reading through the text data we can identify some themes. For example, we can classify the responses as falling into one or more of the following responses. Such a list of categories created from text data is known as a code frame. The number of each category is typically called the code when doing coding but it is the same as the value of a variable.

Value	Label
1	Sugar-free

2	Taste
3	Weight/Health
4	Price
5	Image
6	Other

The creation of a code frame is a highly subjective process. There's a good chance that if you read through the list of text data above you would have come up with a different code frame. For example, in the above code frame Sugar-free is distinct from Weight/Health, but an argument can be made that they are the same, with Sugar-free being a feature that ladders to Weight/Health. One approach to resolving this is to get a few different people to independently code the data and then resolve discrepancies by discussion, but such a rigorous approach is extremely rare outside of small academic studies.

Coding

The final stage of coding, which is itself called coding, involves assigning each response to one or more codes (one if it is to be interpreted as a single response question and more if it is interpreted as a multiple response question). Again, as with creation of the code frame this is a subjective process and greater accuracy can be obtained by getting multiple people to code the same data (but, such rigor is extremely rare in commercial studies).

People who drink Pepsi Max probably prefer sugar - free drinks	1
Big or large people drink Pepsi light and normal size people drink Pepsi Max	3
I prefer the flavour of Diet Coke to Pepsi and that is what I base my decision on	2
????????????????	Missing Value
Probably the diet or light colas appeal to those who want to reduce their sugar intake, which traditionally I suppose, is women. I found that once I stared to drink the light colas, I really disliked the taste of the other 'full	1, 2, 4

strength' ones. If Coke or Pepsi are on special, I will buy the cheaper, because after the first drink, I adjust to the differences between the brands, and I like both.	
Pepsi is full of sugar and Im a diabetic....	3
Pepsi gives the impression of being a "home brand" type drink and coke zero is used by people who want to impress	5
traditionalists versus rebels	5
health conscious	3
People who drink Coke Zero are health consitious	3

Creating New Variables

Often it is useful to construct a new variable. Sometimes, this will be a numeric variable. For example, if one question asks about the number of bottles of Coke consumed in a week and another asks about the number of bottles of Pepsi it may be useful to work out the number of bottles of cola consumed. Other times it will be useful to create a categorical variable. For example, it may be useful to combine together questions about age, marital status and children to create a new categorical variable indicating family life stage (e.g., young singles, young couples, etc.).

Copying, merging and recoding

Often a new variable needs to be derived from an existing variable. For example, it may be that there is an existing variable with categories of Strongly Disagree, Somewhat Disagree, Neither Agree nor Disagree, Somewhat Agree and Strongly Agree, and there is a desire to create a new variable with categories of Agree and Not Agree. In such a situation the straightforward approach is to copy the existing variable and then merge the categories in the copied variable. Where programs do not permit the user to merge categories, such as in SPSS, the same outcome can be achieved by recoding.

Automatic categorization

The more sophisticated programs, such as SPSS, R and Q have various in-built tools for automatically categorizing numeric variables (e.g., into quartiles).

Formulas

The most powerful, but also most complex, way of creating new variables is to use formulas. For example, if wanting to create a new numeric variable by adding two variables called q1 and q2 most programs will have a facility to create a new variable using a formula such as:

```
q1 + q2
```

Or, if creating a new categorical variable the formula may look something like:

```
if (age <= 35){  if (numberChildren == 0) { if (married) 2; //young couples
else 1; //young singles  }  else { if (married) 3; //young families else 4;//
young single families  }else{  if (numberChildren == 0) { if (married) 2; //older
couples else 1; //older singles  }  else { if (married) 3; //older families else
4;// older single families  }}
```

Multivariate analysis

Many multivariate analysis methods, such as Factor Analysis and Latent Class Analysis automatically output new variables that are summaries of the data.

Deleting Respondents

From time-to-time a data file will contain respondents whose data needs to be excluded from any and all analyses. Typically this is done by:

- Creating a new filter variable that indicates who needs to be deleted (or, who should not be deleted).
- Giving the analysis program an instruction to delete the respondents' data.

Checking Representativeness

Checking the representativeness of a study involves trying to work out if the data from the study is consistent with what is known about the population that the study seeks to understand.

Basic process for checking representativeness

1. Collect secondary data on your population (e.g., age, gender, place-of-residence, brands purchased).
2. Ensure that the questionnaire is written so that the survey data can be easily compared with the secondary data.
3. Compare the secondary data with the results of the survey.
4. Choose to either:
    - Weight the data.
    - Discard the data.
    - Use the data as it is.

Example

The table below shows the proportion of people in different age bands in a survey of the phone market. A quick examination of the data reveals that something is wrong. For example, only 3% of the sample are aged 30 to 34, whereas approximately 10 times this number are aged 20 to 24. This suggests that the sample may not be representative of the population at large.

Age	**Survey**
15 and under	0.1%
16-19 yrs	10.2%
20-24 yrs	32.0%
25-29 yrs	14.8%
30-34 yrs	3.5%

35-44 yrs	4.9%
45-54 yrs	27.7%
55-64 yrs	5.0%
65 and over	1.8%
Total	100%

In order to be confident that a sample is or is not consistent with what is known about a population we need some other data about the population. Ideally this will be from a source with known reliability, such as sales statistics or government studies of the market. However, it can also be obtained from other surveys where there was a good reason to believe in their representativeness. As an example, the following data shows the required age in the population for the survey.

Age	**Population #**	**Population %**
15 and under	10,670,000	3.3%
16-19 yrs	20,914,000	6.5%
20-24 yrs	25,619,000	8.0%
25-29 yrs	25,107,000	7.8%
30-34 yrs	24,605,000	7.7%
35-44 yrs	48,225,000	15.0%
45-54 yrs	47,261,000	14.7%
55-64 yrs	46,316,000	14.4%
65 and over	72,623,000	22.6%
Total	321,340,000	100%

Putting the sample results side-by-side with the known data for the market quickly reveals the scale of the problem with this particular study. Looking at the first row of numbers in the table below we can see that 0.1% of the sample were aged 15 and under, which compares to 3.3% in the population and thus the true proportion in the population is approximately 23.8 times that observed in the study (as all the numbers are rounded you will get a slightly different answer if you attempt to reproduce this calculation). Reading down the right-most column of the table we can see that there are no instances where the ratio of the percentages is at 1.0, which is approximately what is required for the study to be representative (small deviations are acceptable due to sampling error).

Age	Survey (A)	Population #	Population % (B)	(B)/(A)
15 and under	0.1%	10,670,000	3.3%	23.8
16-19 yrs	10.2%	20,914,000	6.5%	0.6
20-24 yrs	32.0%	25,619,000	8.0%	0.2
25-29 yrs	14.8%	25,107,000	7.8%	0.5
30-34 yrs	3.5%	24,605,000	7.7%	2.2
35-44 yrs	4.9%	48,225,000	15.0%	3.1
45-54 yrs	27.7%	47,261,000	14.7%	0.5
55-64 yrs	5.0%	46,316,000	14.4%	2.9
65 and over	1.8%	72,623,000	22.6%	12.4
Total	100%	321,340,000	100%	1.0

A similar process would then need to be performed for any other secondary data. Then, once all the comparisons have been done, a decision needs to be made, where the options are:

- Conclude that the data from the sample is consistent with what is known about the population. In this example such a conclusion is clearly not warranted.
- Weight the data, which involves adjusting how the data is analyzed so as to take into account the nature of differences between the sample and the population. This option only makes sense if any deviations between the survey's results and the population are considered to be

'sensible'. Typically, results as disparate as those shown above would indicate that the study was fundamentally flawed. However, there are scenarios when such discrepancies may be considered plausible. For example, it is commonly the case that women are more likely to respond to surveys than men and so if a survey exhibited an over-representation of women and all the discrepancies between the survey and secondary data could be explained by such a phenomena, then weighting would be a sensible remedy.

- Discard the data. That is, conclude that the discrepancies are so great as to render the survey untrustworthy. The discrepancies shown in the example above would suggest that such a conclusion would be justified, as if the survey is so inaccurate in terms of reflecting the population's age then it is reasonable to assume it will be similarly inaccurate for other questions in the survey where there is no way of checking. In practice, it is extremely rare to ever discard a survey, as most people take the view that it is better to have a bad survey than no survey.

Weighting

Weighting is used to adjust the results of a study to bring them more in line with what is known about a population. For example, if a sample contains 40% males and the population contains 49% males weighting can be used to correct the data to correct for this discrepancy.

Worked example

Consider the following data showing 10 people's favorite celebrity:

```
Brad Pitt Brad Pitt Brad Pitt Brad Pitt Brad Pitt Brad Pitt Brad Pitt Brad Pitt
Tiger Woods Tiger Woods
```

In this sample of ten, 80% of people have nominated Brad Pitt as their favorite celebrity. When we conduct market research studies, the whole point is to draw conclusions about the population, rather than just our sample, so we should go one step further and, rather than say "80% of the sample nominate Brad Pitt as their favorite celebrity" we should instead be conducting research so that we can say "Brad Pitt is the most popular celebrity, with 80% of people nominating as their favorite (we should also communicate the precision, as discussed in Determining The Sample Size.

Now, consider the impact of some additional information about the gender of our ten respondents:

```
Brad Pitt Female Brad Pitt Female Brad Pitt Female Brad Pitt Female Brad Pitt Female Brad Pitt Female Brad Pitt Female Brad Pitt Female Tiger Woods Male Tiger Woods Male
```

From this data we can see that the sample is unrepresentative in terms of gender, with eight of ten respondents being female 80%) and two being male (20%), compared to the true representation in the world of about 50% men and 50% women. Furthermore, gender seems to be the sole determinant of preference. As the sample is not representative in terms of gender and gender is correlated with our measure of favorite celebrity, it follows that any estimate of the favorite celebrity will only be valid if we take into account the over-representation of women in the sample.

We can improve our estimate by weighting. A weight is computed for every respondent in a sample, and it is computed by dividing the correct proportion by the observed proportion. The correct proportion of males in our population is 50% and the observed proportion is 20%, so the weight for each male is 50%/20%=2.5 and the weight for each female is 50%/80%=0.625. Thus, our data becomes:

```
Favourite Celebrity Weight Brad Pitt 0.625 Brad Pitt 0.625 Brad Pitt 0.625 Brad Pitt 0.625 Brad Pitt 0.625 Brad Pitt 0.625 Brad Pitt 0.625 Brad Pitt 0.625 Tiger Woods 2.5 Tiger Woods 2.5
```

We now compute our estimate of the proportion of people to have Brad Pitt as their favorite celebrity by summing up the weights of each of the respondents to prefer Brad Pitt and dividing this by the sum of all of the respondents' weights:

$$\frac{8 \times 0.625}{2 \times 2.5 + 8 \times 0.625} = \frac{5}{10} = 50\%$$

The approach described here for computing a weight is a relatively simple case, but the basic idea can be extended to deal with much more complicated cases (e.g., it is routine to simultaneously weight by geography, gender and age).

Weighting using standard analysis software

With the exception of Q, none of the general-purpose survey analysis programs contain useful tools for creating weights. Consequently, weights are usually created by either:

1. Have them computed using data processing software. Most companies that collect data offer the creation of weights as a service and it is commonplace to have weights included in the data file.

2. Compute them manually. That is, you need to:
    - Compute the observed proportion of people that have selected each answer.
    - Obtain the targets (i.e., the results believed to apply in the population; in the example above the targets were that 50% of the population was male and 50% was female.
    - Compute the weights by dividing targets by the observed proportions (note that this is only one approach to weighting and more complicated approaches to weighting, such as rim weighting cannot be done this way.

When weighting does and does not work

Consider the situation where a survey was well managed but due to some anticipated quirk of how the data was collected it over-represent males. If gender is known to relate to other variables of interest in the survey then it follows both that:

- The data needs to be weighted or otherwise any analyses will be misleading (as they will be biased by the misrepresentation of gender).
- The weighting will fix the data.

However, what if the gender problem is one of many unknown problems in the data. Imagine that the survey is also biased towards low income people, towards people that like to do surveys and towards people that do not get out of doors much. Fixing the gender problem by weighting will not fix any of these other problems. Furthermore, in general we will not know that we have these other problems, as we will have no data to check for representativeness.

The distinction between these two cases is crucial. Where the data is observed to be unrepresentative, weighting will only fix the data if there is good reason to believe that this is the only way in which the data will be representative.

Multiple Response Data Formats

Binary

Each category of the multiple response question is represented as a separate variable. Typically a `1` is used when an item has been selected and a `0` otherwise.

Max-Multi

Each variable can contain multiple categories, one for each of the options in the questionnaire. For example, if the question has 20 brands, but nobody selected more than five brands, it is possible to store the data as five variables, with the first containing one of the responses, the second another, and so on. There is no widely used term for this way of storing multiple response data; it is sometimes referred to as max-multi data.

Pros and cons of the different formats

The binary format is the standard and, in most situations, the preferable format. This is because it:

- Makes data analysis easier (as it is easier to compute new variables).
- Makes it possible to have missing values. For example, in a tracking study if a brand is in the questionnaire in some waves but not other waves then this can be reflected using a missing value in waves when it is not available. By contrast, where max-multi coding is used it is impossible to determine from the data whether an option was not selected, or, was not available.

The max-multi format's advantages are:

1. It uses up less memory on the computer (i.e., because it takes less variables). In the 1970s this meant that the max-multi was the standard format. Now, however, it tends to only be used sensibly in situations where there are questions with extremely large code frames (e.g., a list of 6,000 brands of car).
2. It is easier to create if manually entering data (i.e., it is much easier for a data entry person to type 2, 5 to indicate that the second and fifth categories were selected than it is for them to enter lots of 1s and 0s).

Variable Type

The type of a variable indicates how it should be interpreted when being analyzed. Sometimes the variable type is incorrectly set in a data file and most analysis programs permit the user to change the variable type to a more appropriate setting.

Numeric

A variable that contains data where it is meaningful to compute an average (e.g., height, weight, number of chocolates eaten in the past week). In SPSS this is referred to as a scale variable.

Categorical

A variable where each unique value indicates a particular category. For example, a 1 may indicate that somebody is male and a 2 may indicate that somebody is a female. Note that it is not meaningful to compute an average of a categorical variable. Such variables are sometimes referred to as being nominal.

Ordered Categorical

A categorical variable where each value indicates a particular category and the ordering of the values has a meaning. For example, a 1 may indicate being aged 18 to 34, a 2 may indicate an age of 35 to 54, and a 3 may represent 55 or more. Such variables are sometimes referred to as being ordinal variables.

Most analysis programs only provide limited support for ordered categorical variables, treating them as if they are (unordered) categorical variables. Consequently, it is commonplace to treat ordered categorical variables as being numeric.

Date

A numeric variable which contains a date. In more sophisticated programs, such as Q, special-purpose analysis methods exist to manipulate date variables (e.g., to automatically create moving averages, aggregate date periods, etc).

Money

A numeric variable that contains money. Generally, this is identical to a numeric variable in all ways except that the currency symbol is shown on tables.

Text

A variable containing text information. Most commonly, such a variable will either contain data from an open-ended question, or, will contain numeric or categorical data that has been stored with the wrong variable type.

Variable Name

A variable name is a name given to a variable in a data file. It is usually an abbreviation or a code that in some way represents the variable's meaning (e.g., `age` or `q42`). Variable names tend to be used when writing expressions to create new variables (e.g., `Q3 + Q4 * 2`). Historically, the names of variables were restricted to being eight characters or less and to have no punctuation (e.g., `Q23a`), but the only common restriction in more modern programs is that variable names may not contain spaces.

Basic Data Analysis

- Summary Tables
- Interpreting a Summary Report
- Crosstabs
- Significance Tests
- Sampling error
- Correlation
- Filtering
- Counted Values and Missing Values in Multiple Response Questions
- Examples

Basic Data Analysis

Once the data has been cleaned, tidied, and, if required, weighted, tables need to be created and interpreted and turned into reporting. This consists of the following steps, but they can be done in any order; summary tables, Crosstabs, significance testing, and filtering data.

Summary Tables

Creating a Summary Report

A summary report is a set of tables or charts which show summary statistics, such as averages and the proportion of the people in each category, for:

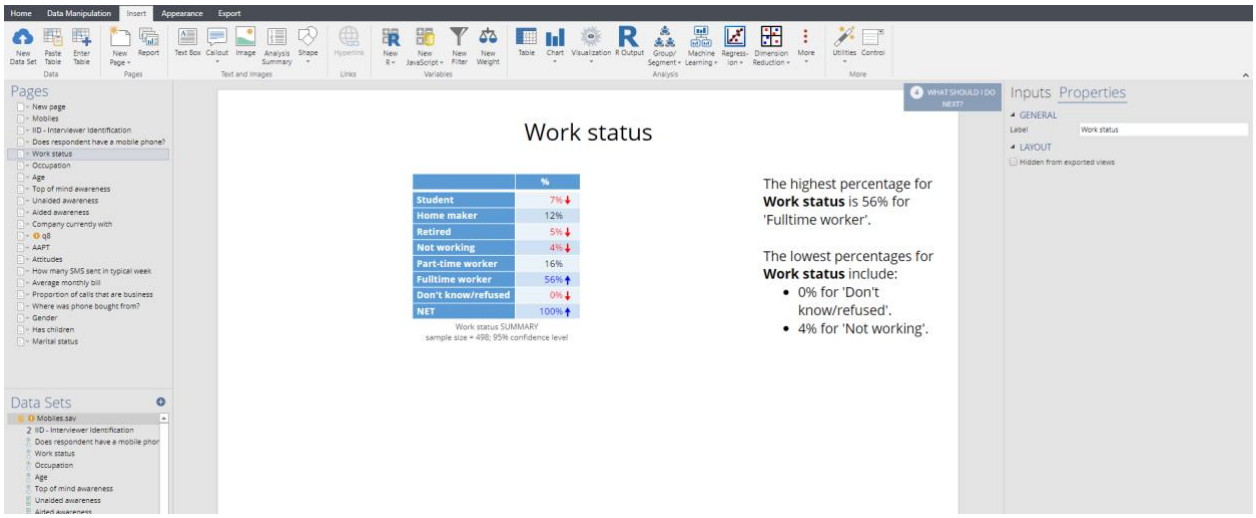
- All the questions in a survey.
- All the administrative records stored as variables in the data file (e.g., the time when the interview was commenced, the time the interview took to complete, the unique ID variable of each respondent).

Examples

The basic information shown in summary tables by most programs is basically the same, other than formatting. The following two examples are from programs that are about as dissimilar as you can get: the traditional summary table is the type commonly used by professional market researchers. The second example shows a style of summary developed for less-experienced researchers.

Displayr

The following screen is shown from Displayr. A separate page is created for each summary table (or, optionally, chart). All the tables are listed on the left of the screen. A text box highlights some of the key features of the table. Arrows and colors are used to highlight results that are significantly high or low.



MarketSight

The summary developed by MarketSight lists all the variables and questions, one after another, in a large table. The main statistics that are shown on the table - the percentages and the Sample Size are the same as those shown above. Additionally, the count is shown on every table automatically.

	Total
IID - Interviewer Identification	
Sample Size	710
Mean	1,409,497.99
Does respondent have a mobile phone?	
Sample Size	725
Yes Count Column %	 715 99%
No Count Column %	 6 1%
3 Count Column %	 1 0%
5 Count Column %	 3 0%
Work status	
Sample Size	700
Student Count Column %	 211 30%
Home maker Count Column %	 62 9%

Interpreting a Summary Report

Categorical and numeric variables

Generally, summary reports will show tables of percentages for categorical variables, such as age and gender, and tables showing averages for numeric variables. For example, in the summary report from MarketSight below we can see that the first table shows an average of a numeric variable and the second shows percentages and counts from a categorical variable.

	Total
IID - Interviewer Identification	
Sample Size	710
Mean	1,409,497.99
Does respondent have a mobile phone?	
Sample Size	725
Yes Count Column %	 715 99%
No Count Column %	 6 1%
3 Count Column %	 1 0%
5 Count Column %	 3 0%
Work status	
Sample Size	700
Student Count Column %	 211 30%
Home maker Count Column %	 62 9%

Switching between categorical and numeric variables

In most programs it is necessary to change the metadata to switch between the average and percentages. Exceptions to this are:

- In SPSS the user specifies whether to run a mean or frequency manually for each table.
- In Q and Data Cracker you can either change the metadata, or, if it is showing a percentage you can use **Statistics - Below** or **Statistics - Right** to add averages to the table of percentages.

Multiple response questions

With multiple response questions there are a couple of different ways of computing percentages:

- Percentage of respondents. The **%** column in the table below (which was computed using Q) shows the proportion of respondents to have selected a response (e.g., the 24% for AAPT is computed by dividing the 122 people to have selected this option by the 498 people that were shown the alternative (which, in this case, was the entire sample). Generally, it is this percentage that is used when reporting data from multiple response questions).
- Percentage of responses. The **% Responses** value of 6% is computed by dividing 122 by the total of all of the counts (i.e.,122/(122 + 46 + ... + 401)). This percentage is rarely used and is perhaps never actually useful, except as an input to data cleaning.

% % Responses n Base n	%	% Responses	n	Base n
AAPT	24% ↓	6%	122	498
New Tel	9% ↓	2%	46	498
One-tel	41% ↓	10%	203	498
Optus	91% ↑	21%	454	498
Orange	53%	13%	264	498
Telstra	91% ↑	21%	454	498
Virgin Mobile	34% ↓	8%	168	498
Vodafone	81% ↑	19%	401	498
NET	100% ↑	24%	498	498

Multiple response summary tables with messy data

When the data from a survey is 'neat', all the main data analysis programs used for analyzing surveys produce basically the same results. However, there are a couple of situations where programs can produce wildly different results.

Missing values

When multiple response data contains missing values the programs produce completely different results. Please see Counted Values and Missing Values in Multiple Response Questions for a discussion of how the programs differ and for instructions for making the programs produce more sensible results.

When not everybody is in a category

In a 'tidy' survey everybody is forced to have an answer in a multiple response question (i.e., people are not permitted to go onto the next question without selecting at least one of the alternatives). However, there are a few scenarios when not everybody will have a response:

1. When there are data integrity problems.
2. When people were not compelled to select an option.
3. When the multiple response question has been created by the user (e.g., if creating top 2 box scores).

In each of these situations different programs give different results. The two tables below are computed using data where everybody in the data has selected at least one category. And, as will occur with all of the standard programs, the results are the same. That is, the percentages on the table on the left, which has been computed using Q, are the same, bar rounding, with those on the right side of the second table, which was computed using SPSS. The only substantive difference between these tables relates to the bottom row, where Q shows a NET, which is the proportion of people to have selected one or more of the options, whereas SPSS shows the total.

%	%
AAPT	9% ↓
New Tel	2% ↓
One-tel	25% ↓
Optus	89% ↑
Orange	43% ↑
Telstra	88% ↑
Virgin Mobile	22% ↓
Vodafone	75% ↑
Other	6% ↓
Don't know	0% ↓
NET	100% ↑

		Responses		Percent of Cases
		N	Percent	
Unaided Awareness[a]	AAPT	44	2.5%	8.8%
	New Tel	9	0.5%	1.8%
	One-tel	124	6.9%	24.9%
	Optus	444	24.9%	89.2%
	Orange	214	12.0%	43.0%
	Telstra	436	24.4%	87.6%
	Virgin Mobile	109	6.1%	21.9%
	Vodafone	373	20.9%	74.9%
	Other	32	1.8%	6.4%
	Don't know	1	0.1%	0.2%
Total		1786	100.0%	358.6%

The two tables below are also computed using Q and SPSS. Further, they use the same data as used in the tables above, except that only the first four categories have been included in the analysis. Note that the Q analysis is almost the same. The percentages for each brand remain the same. The only difference relates to the NET, which is 100% for the table above, but 93% for the table below, which is because only 93% of the sample have selected one of the four brands shown. By contrast, the results for SPSS are all different. In fact, they are all about 8% higher on the table below compared to the table above. The reason for this is the way that it uses a somewhat strange formula. The SPSS percentages have been computed by dividing the number of people to have selected any option by the number of people to have selected one or more options. Looking at the AAPT data, in the table above SPSS shows 8.8% which is computed as 44 / 498, where 498 is the proportion of people to

have selected one or more option (i.e., the total sample). In the table below, however, 9.5% is shown which is 44 / 462, where 462 is the number of people to have selected one or more of the four brands used to construct the table.

%	%
AAPT	9% ↓
New Tel	2% ↓
One-tel	25% ↓
Optus	89% ↑
NET	93% ↑

		Responses		Percent of Cases
		N	Percent	
$un[a]	AAPT	44	7.1%	9.5%
	New Tel	9	1.4%	1.9%
	One-tel	124	20.0%	26.8%
	Optus	444	71.5%	96.1%
Total		621	100.0%	134.4%

It is important to appreciate that the discrepancy between the results is caused by having data where some people have not selected any of the categories and where the data does not suffer from this problem the different programs will give the same results. Additionally, the difference is one of those rare instances where one of the programs is producing numbers that are, in most situations, unhelpful (i.e., the results produced by the SPSS calculation in the second table are misleading, because most people would assume that they relate to the proportion of respondents to select the option and such an interpretation is incorrect. Unfortunately, the SPSS calculation is the 'standard' one and is used by most data analysis programs (which have generally been written under the assumption that people are compelled to choose at least one option).

The reason that the programs do it differently

As mentioned, in situations where the NET is 100% the two methods will get the same answer. The table-based method is the traditional approach. In a traditional survey the NET will always be 100%, because in a traditional survey run by a professional researcher there would always be a 'None of these' option and thus both methods get the same results. Thus, the traditional programs use the table-based method because it is faster to compute when there is no missing data. However, in situations where there is a chance that the data will be messy in some way the respondent-based method is preferable as it has the advantages that:

- The possibility of a problem is flagged by the NET not being 100%.
- The values that are estimated are sensible (i.e., it is much easier to explain that the percentage represents the proportion of respondents than it is to describe the percentage as representing the proportion amongst respondents that have selected at least one option).

Thus, as many of the traditional programs are developed under the assumption that the data is relatively clean they employ a method that is best in those situations, where as the more modern programs use the alternative method as it is safer in the modern world where the data is often messy.

How to switch between the different types of multiple response computations

In most programs it is possible to get the program to change the way that it computes the percentages on multiple response questions. In programs that use the respondent-based method the trick is to filter the table so that it only contains respondents that selected one or more options. In programs that use the table-based method the trick is to not tell the program that it is a multiple response question.

Using the summary report to guide data cleaning

At a minimum, the summary report should be reviewed to check that the results make sense, which essentially involves comparing results with things that are already known about the population being studied (this is discussed in detail in Checking Representativeness), and that they are plausible (e.g., if a respondent claims to have 99 mobile phones then this suggests that there is a problem.

More thorough data cleaning involves checking that the two-way relationships between variables are sensible. For example, if checking data on firm profitability, it is useful to eview the profitability per number of employees, as it may make sense for a firm to contribute 10 million dollars of profit to an industry, but it is less likely if the firm has 1 employee. This is done by creating Crosstabs. Typically, this is done as a part of the main data analysis rather than as a separate stage of data preparation.

The hard part of data cleaning is deciding what to do with “dirty” data. Consider as an example a data file that indicates that the person goes to the beach 99 times a month in summer. The options are to:

- Determine that the problem is that the metadata is incorrect. For example, it may be that a value of 99 does not represent the number of trips to the beach instead indicates that they person the person did said "don't know". See Correcting Metadata.
- Delete the incorrect value, replacing -99 with a special code indicating the data is invalid. This results in missing values and then there is often a need to use special analysis tools that can address the missing data. See Missing Values.
- Change the value (e.g., replacing 99 with 9). See Recoding Variables.
- Change the value to multiple values and assign probabilities to the different values. Although this can be the most appropriate thing to do, it is extraordinarily unusual for something like this to occur in a real-world commercial study and as such this approach, which is known as multiple imputation, is discussed no further.

- Delete the entire record of data that is dirty, which involves making the assumption that this one error indicates all their data is wrong. See Deleting Respondents.

In order to work out which of these is appropriate we need to understand the cause of the poor data (e.g., key punching errors, corrupted data, respondent error), as if we clean the data without understanding the cause of the problems, we run the high risk that other data that we have not spotted as being dirty is inaccurate and that the “clean” data does not accurately represent the market.

Crosstabs

A table showing the relationship between two questions in a survey is called a crosstab.

Reading crosstabs

The following table is a crosstab of age by whether or not somebody has a listed phone number.

		Unlisted phone number		
Age		No	Yes	**NET**
18-34	% within column	24%	49%	**29%**
	n	185	90	**275**
35-44	% within column	20%	26%	**21%**
	n	153	48	**201**
45-54	% within column	17%	10%	**16%**
	n	133	19	**152**
55-64	% within column	17%	11%	**16%**
	n	130	21	**151**
65+	% within column	23%	3%	**19%**
	n	178	6	**184**
NET	% within column	**100%**	**100%**	**100%**
	n	**779**	**184**	**963**

This table shows the number of observations with each combination of the two questions in each cell of the table. The numbers of observations are often referred to as the counts. We can see, for example, that 185 people are aged 18 to 34 and do not have an unlisted phone number.

Column percentages

Column percentages are shown on the table above. These percentages are computed by dividing the counts for an individual cell by the total number of counts for the column. A column percent shows the proportion of people in each row from among those in the column. For example, 24% of all people without an unlisted phone number are aged 18 to 34 in the sample (i.e., 185 / 779 = 24%) and thus we can say that based on this sample we estimate that 24% of people with an unlisted phone number are aged 18 to 24.

Row percentages

Row percentages are computed by dividing the count for a cell by the total sample size for that row. A row percent shows the proportion of people in a column category from among those in the row. For example, as 185 people are aged 18 to 34 in the No column and there are a total 275 people aged 18 to 34 the row percentage is 67% (i.e., 185 / 275) and thus we can say that based on this sample we estimate that 67% of people aged 18 to 34 have an unlisted phone number.

Working out whether the table shows row or column percentages

Some crosstabs do not clearly label whether percentages are row and column percentages (e.g., the example below). When reading a table, the easiest way to check if it is showing row or column percentages is to check to see which direction the numbers add up to 100%. In the table above, the percentages add up to 100% in each column and, furthermore, this is indicated on the table by the NET, and thus it shows column percentages.

Checking to see if the percentages add up to 100% only works where the categories in the rows (or columns) are mutually exclusive. Where the data is from a multiple response question it is more difficult, as the percentages will add up to more than 100% (as people can be in more than one category). An example is shown in the table below, which shows two different types of column percentages:

- The percentage of people to have selected each option (% Valid Cases).
- The percentage of options selected (% Total Mentions).

(See Counted Values and Missing Values in Multiple Response Questions for a more detailed discussion of how to interpret such tables).

In the crosstab below the percentages do not add up to 100% in either direction and there is nothing in the way the table is labelled to make it clear whether the table is showing row or column percentages.

	Age								
	Total	16-19 yrs	20-24 yrs	25-29 yrs	30-34 yrs	35-44 yrs	45-54 yrs	55-64 yrs	65 and over
Unaided awareness									
Sample Size	498	36	47	51	48	87	78	75	76
Valid Cases	498	36	47	51	48	87	78	75	76
Total Mentions	1,786	127	188	205	189	321	259	243	254
AAPT									
Mentions	44	3	3	6	7	7	6	7	5
% Valid Cases	9%	8%	6%	12%	15%	8%	8%	9%	7%
% Total Mentions	2%	2%	2%	3%	4%	2%	2%	3%	2%
New Tel									
Mentions	9	0	2	4	0	2	0	0	1
% Valid Cases	2%	0%	4%	8%	0%	2%	0%	0%	1%
% Total Mentions	1%	0%	1%	2%	0%	1%	0%	0%	0%
One-tel									
Mentions	124	6	15	**18**	13	21	**15**	16	20
% Valid Cases	25%	17%	32%	**35%**	27%	24%	**19%**	21%	26%
% Total Mentions	7%	5%	8%	**9%**	7%	7%	**6%**	7%	8%
Optus									
Mentions	444	34	42	46	43	73	73	63	70
% Valid Cases	89%	94%	89%	90%	90%	84%	94%	84%	92%
% Total Mentions	25%	27%	22%	22%	23%	23%	28%	26%	28%
Orange									
Mentions	214	11	18	25	22	41	33	36	28
% Valid Cases	43%	31%	38%	49%	46%	47%	42%	48%	37%
% Total Mentions	12%	9%	10%	12%	12%	13%	13%	15%	11%
Telstra									
Mentions	436	**25**	44	42	44	**79**	64	**65**	73
% Valid Cases	88%	**69%**	94%	82%	92%	**91%**	82%	**87%**	96%
% Total Mentions	24%	**20%**	23%	20%	23%	**25%**	25%	**27%**	29%
Virgin Mobile									
Mentions	109	**16**	**20**	**13**	**12**	**20**	**13**	**7**	**8**
% Valid Cases	22%	**44%**	**43%**	**25%**	**25%**	**23%**	**17%**	**9%**	**11%**
% Total Mentions	6%	**13%**	**11%**	**6%**	**6%**	**6%**	**5%**	**3%**	**3%**
Vodafone									
Mentions	373	**31**	**42**	**46**	**43**	**72**	**49**	**44**	**46**
% Valid Cases	75%	**86%**	**89%**	**90%**	**90%**	**83%**	**63%**	**59%**	**61%**
% Total Mentions	21%	**24%**	**22%**	**22%**	**23%**	**22%**	**19%**	**18%**	**18%**
Other									
Mentions	32	1	2	5	4	6	6	5	3
% Valid Cases	6%	3%	4%	10%	8%	7%	8%	7%	4%
% Total Mentions	2%	1%	1%	2%	2%	2%	2%	2%	1%
Don't know									
Mentions	1	0	0	0	1	0	0	0	0
% Valid Cases	0%	0%	0%	0%	2%	0%	0%	0%	0%
% Total Mentions	0%	0%	0%	0%	1%	0%	0%	0%	0%

In most cases when a percentage is shown on a crosstab it is a column percentage. This table shows column percentages. Where the trick of adding up the percentages does not work, as in this example,

there are a few ways we can deduce whether a particular set of numbers is row or column percentages.

- The position of the sample sizes on a table. By convention, if the sample sizes appear at the top of the table then column percentages are being shown and if the sample sizes appear in a column then the row percentages are shown. In the example above the sample sizes are shown at the top, suggesting that the two percentages shown are different variants of column percentages.
- The position of the % signs on a table. By convention, if % symbol only appears at the top of each column in a table then column percentages are being shown and if the % symbol appears at the beginning of each row then row percentages are shown.
- The degree of variation in the totals of percentages. For example, in the table below we can see that the percentages vary quite a lot within each column, but within each row they are reasonably similar, which indicates that the table shows column percentages (similarly, if the variability was greater in the rows this would indicate that row percentages were shown).

Other statistics

Most commonly crosstabs show and percentages. However, where the variables are not categorical, then other statistics such as averages, meadians and correlations are shown in the cells of a crosstab.

Significance Tests

A significance test is a way of working out if a particular difference is likely to be meaningful or be a fluke.

Sampling error

Imagine doing a study of 200 consumers and finding that 41% said that Coca-Cola was their favorite soft drink. Now, imagine you did another study and found that in the next study 40% of people preferred Coca-Cola. And, imagine you did a third study and found that 43% of people preferred Coca-Cola. What can you infer from the differences between these studies? There are three explanations:

Explanation 1: The world changed in some way between each of these studies and the proportion of people preferring Coca-Cola dropped a little and then increased (i.e., moved from 41% to 40% and then up to 43%).

Explanation 2: The difference between the two studies is just random noise. More specifically, as each study only sampled 200 people it is to be expected that we should get small differences between the results of these. Or, to use the jargon, there is sampling error.[note 1]

Explanation 3: A mix of explanations 1 and 2.

Significance tests

A significance test is a rule of thumb that is used to help to determine whether a difference between two numbers is likely to reflect a meaningful difference in the world at large (i.e., explanation 1 above), or, is merely a fluke caused by sampling error (i.e., explanation 2).

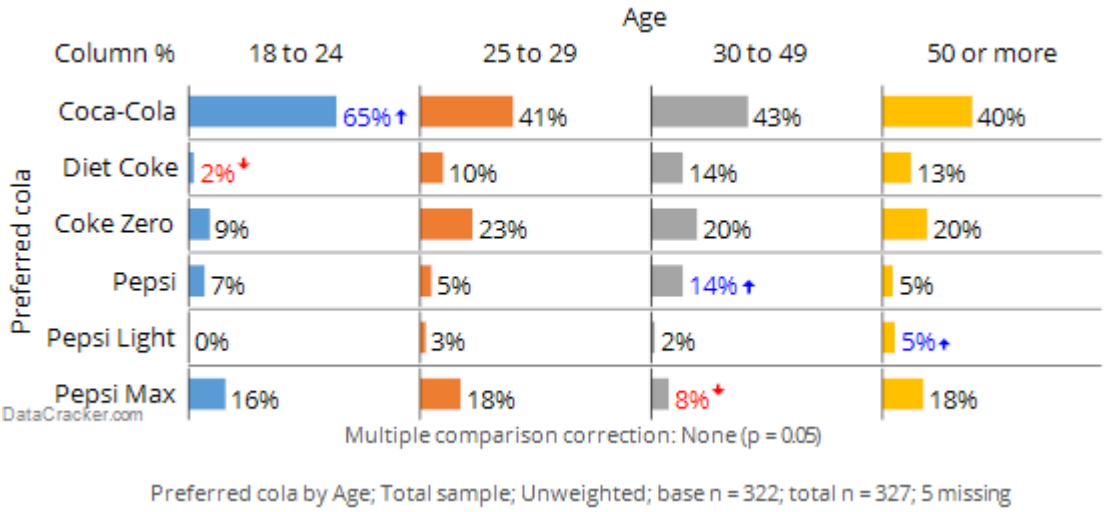
There are many thousands of different significance tests with exotic names like Wilk's lambda, Fisher's Exact Test and so on. However, when analyzing survey data there is generally no need to go into such specific detail about which test to use and when as most significance tests that are applied when analyzing real-world surveys are either exception tests or column comparisons.

Exceptions tests

Consider the following chart from Displayr. Reading across the Coca-Cola row we can see that:

- 65% of people aged 18 to 24 prefer Coca-Cola.
- 41% of people aged 25 to 29 prefer Coca-Cola.
- 43% of people aged 30 to 49 prefer Coca-Cola.
- 40% of people aged 50 or more prefer Coca-Cola.

That we get different results in each of the age groups is to be expected. The process of selecting people to participate in a survey means that by chance alone we expect that we will get slightly different results in the different age groups even if it was the case that there really is no difference between the age groups in terms of preference for Coca-Cola (i.e,. due to sampling error). However, the level of preference for 18 to 24 year olds is substantially higher. In the chart below the font color and the arrow indicate that this result is significantly high. That is, because the result has been marked as being statistically significant, the implication is that the much higher result observed for the 18 to 24s is not merely a fluke and signifies that in the population at large it is true that 18 to 24 year olds have a higher level of preference for Coca-Cola.



Preferred cola by Age; Total sample; Unweighted; base n = 322; total n = 327; 5 missing

Looking elsewhere on the table we can see that: Diet Coke preference seems to be low among people aged 18 to 24, Pepsi scores relatively well among the 30 to 49 year olds and so on. Each of these results are examples of exception tests, which are statistical tests that identify results that are, in some way, exceptions to the norm.[note 2]

Now look carefully at the row for Coke Zero. The score for the 18 to 24 year olds is less than half that of the other age groups. However, it is not marked as being significant and thus the conclusion is that the relatively low score for 18 to 24 year olds may be a fluke and does not reflect a true difference in the population at large. The word 'may' has been italicized to emphasize a key point: there is no way of known for sure whether the low score among the younger people in the survey reflects a difference in the population at large or is just a weird result that occurs in this particular sample. Thus, all significance tests are just guides they rarely prove anything and we always need to apply some commonsense when interpreting them.

Column comparisons

The table below shows exactly the same data from the same survey as shown above. However, whereas the chart above showed results as exceptions, this one instead shows a more complicated type of significance test called column comparisons. Each of the columns is represented by a letter, shown at the bottom of the page. Some of the cells of the table contain letters and these indicate that the result in the cell is significantly higher the results in the columns that are listed. For example, looking at the Coca-Cola row, the appearance of b c d indicates that the preference for Coca-Cola of 65% among the 18 to 24 year olds is significantly higher than the preference scores of the 25 to 29 year olds (b), the 30 to 49 year olds (c) and the people aged 50 or more (d). That the letter are in lowercase tells us that the difference is not super-strong (in the exception shown above, the length of the arrows communicates the degree of statistical significance).

Column % Column Comparisons	18 to 24	25 to 29	30 to 49	50 or more	NET
Coca-Cola	65% b c d	41%	43%	40%	44% -
Diet Coke	2%	10%	14% a	13%	11% -
Coke Zero	9%	23%	20%	20%	19% -
Pepsi	7%	5%	14% d	5%	9% -
Pepsi Light	0%	3%	2%	5%	3% -
Pepsi Max	16%	18%	8%	18% c	14% -
NET	100% -	100% -	100% -	100% -	100% -
Column Names	A	B	C	D	E

Multiple comparison correction: None (p = 0.05); Column comparison symbols: a, b, c... (p <= 0.05), A, B, C... (p <= 0.001); No test symbol: -; Not significant symbol:

Preferred cola by Age; Total sample; Unweighted; base n = 322; total n = 327; 5 missing

Note that while many of the conclusions that we can get from this table are similar to those from the chart above, there are some differences. For example, in the chart above we drew the conclusion that the Diet Coke preference was significantly lower among the 18 to 24 year olds than among the population at large. However, the column comparisons tell us only that the 18 to 24s have a lower score than the 30 to 49s (i.e., we know this because the a for the 30 to 49s tells us that they have a stronger preference than the 18 to 24s who are represented as column a).

Why do the two ways of doing the tests get different results? There are some technical explanations.[note 3] But all the really amount to is this: the different approaches make slightly different technical results and, consequently, they get slightly different results. An analogy that is useful is to think about different ways of reporting news: we can get the same story reported on TV, in a newspaper and in a blog and each way will end up emphasising slightly different aspects of the truth.

The determinants of statistical significance

There are many different factors that influence whether a particular difference is reported as being statistically significant or not, including:

1. The size of differences being compared (i.e., the bigger the difference the more likely it will be significant). This is exactly the same idea that is discussed on in the page on Determining The Sample Size.
2. The sample size. Differences observed in larger sample sizes are more likely to be statistically significant.
3. The specific confidence level of the testing.
4. The number of technical assumptions that are made in the test (e.g., assumptions of normality). In the main, the fewer assumptions that are made the lower the chance that a result is concluded as being statistically significant.
5. If and how the data has been weighted. The greater the effect of the weighting the less likely that results will be statistically signicant.
6. The number of tables that are viewed and the size of the tables. The greater the amount of analyses that are viewed, the greater the chance of fluky results.
7. How the data has been collected (in particular, what approach to sampling was adopted).
8. The technical proficiency of the person that has written the software conducting the test. In particular, most formulas presented in introductory statistical courses only take into account the first three of the issues listed above and most commercial programs deal with the weighting incorrectly. The general ambiguity of statistical testing in terms of it not being able to give definitive conclusions combined with the large number of technical errors that are made in practical applications of significant tests again lead to the same conclusion presented earlier: statistical tests are nothing more and nothing less than a useful way of identifying interesting results that may reflect how the world works but also may just be weird flukes.

A more technical explanation

The above description is not a simplification. It is an accurate depiction of what statistical testing does and is. A more theoretical discussion of how to interpret statistical tests on tables is at Statistical Tests on Tables at SurveyAnalysis.org.

Notes

1. Jump up↑ Or, to be more precise, sampling error is the difference between what we observe in a random sample and what we would have obtained had we interviewed in the population.
2. Jump up↑ The term exception test is not a standard term. The closest there is to a standard term for such a test is studentized residuals in contingency tables, but even this is a pretty obscure term.
3. Jump up↑ In particular, the exceptions test has more statistical power due to the pooling of the sample, the columns comparisons are not transitive and there are smaller sample sizes for column comparisons than for exception tests.

Correlation

'Correlation' refers to the degree to which two things have a statistical relationship. For example, if people that have higher incomes consume more wine, we say that wine consumption is correlated with income.

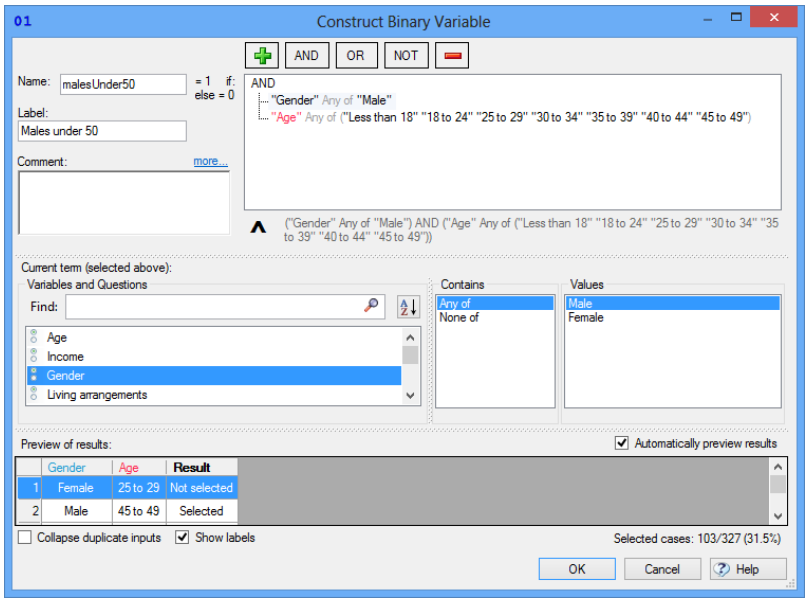
Filtering

Filtering data involves restricting the analysis to a particular sub-group of the sample. For example, if you view the data of people in your sample that are male and aged under 50 you have filtered the data.

Creating filters

Filters are created using rules regarding which respondents should be included and which should be excluded from the analysis. While there are some nuances, generally filters are created by various AND and OR rules. For example, your rule may be to include people that are `aged under 50 AND are males`, or, the rule may be `aged under 50 OR are males`. There is no consistence between the different data analysis programs in terms of how filters are created. In SPSS, for example, filters have to be created by typing an expression. For example, a filter of males under 50 would be entered as `q2 <= 7 & q3 == 1`, where `q2` and `q3` are Variable Names and 7 and 1are specific values that represent age and gender categories respectively.

By contrast, Q instead uses the same basic logic, but presented in a 'tree' type format (on the left), whereas Displayr uses a less-flexible but easier-to-use grid of checkboxes.



Choose the data that describes the cases in the filter:

Data: Age

Second variable: Gender

Choose the categories to include in the filter:

	Gender	
Age	Male	Female
18 to 24	☐	☐
25 to 29	☐	☐
30 to 34	☐	☐
35 to 39	☐	☐
40 to 44	☐	☐
45 to 49	☐	☐
50 to 54	☐	☐
55 to 64	☐	☐
65 or more	☐	☐

Selected: 0/327 (.00%)

Filter variables

Almost all programs treat the creation of a filter as being equivalent to creating a new variable, where the variable contains two categories, one representing the people in the filter and one representing the people not in the filter group. Typically, these are added to the data file allowing them to be re-used.

Applying filters

Once a filter has been created it can usually be re-used by selecting it from a list of saved filters. The only prominent exception to this is SPSS, in which you need to create a new filter but can do so by using the older filter (e.g., if the previously-created was called var001 then the expression for the new filter if re-using it would be var001. Another difference between SPSS and most programs is that in SPSS a filter is either on or off, whereas in other programs the filter is specifically applied to separate analyses.

Counted Values and Missing Values in Multiple *Response Questions*

Counted values

In a single response question it is usually obvious that the correct way to compute the proportions is to compute the number of people that selected a category and divide this by the total number of people that selected at least one category. At an intuitive level it makes sense that percentages of multiple response data would be computed in the same way. However, the way that data is stored prevents it from being quite so simple. Usually, multiple response data is stored so that there is one variable for each brand. However, it is not always clear how to analyze this particular variable.

In some data files the code frame will be set up as:

```
0 Not selected 1 Selected
```

In such a situation it is usually pretty obvious that the correct way to compute the proportions is to work out the proportion of people with a 1 in their data (it gets more complicated if there is missing

data; this is discussed in the next section). Or, phrasing it in a different way, the correct way of computing the proportions is to count the higher value (i.e., the 1).

Similarly, if there is no metadata and the variable only contains 0s and 1s it is still obvious that the 1s should be counted.

However, where it gets complicated is when there are values other than 0 and 1 in the data. For example, sometimes the code frame will be:

```
1 Yes 2 No
```

To a human being it is obvious at the Yes responses should be counted. However, in one sense, this is the opposite to the previous examples, as now we are counting the lowest of the observed values rather than the highest.

Due to the potential ambiguity, the way that most programs works is that they either force the user to specify a specific value (e.g., SPSS requires the user to specify the **Counted value**), or, they give the user the ability to inspect and modify the setting. For example, in Q the user specifies whether the analysis should or should not **Count this value** and in Displayr the user has the option to **Select Categories**.

Multiple response data where the variables contain multiple values

Sometimes the variables contain more than two values, so it is not at all obvious which of the values should be counted. There are two very difference instances of this.

Case A: Max-Multi data

If the data is in max-multi format then different options need to be selected at the time of creating the multiple response set.

In SPSS, for example, the data needs to be selected as **Categories** when defining the multiple response sets, whereas in Q there is a special question type of **Pick Any - Compact** designed for this type of data.

Case B: Recoding grid questions

Often it is useful to treat some types of grid questions as if they are multiple response questions. Most commonly, with a question that gets people to rate agreement using five points (e.g., Strongly disagree; Somewhat disagree; Neither agree nor disagree; Somewhat Agree; Strongly Agree it is common to turn this into a top 2 box scores (i.e., the NET of Somewhat Agree and Strongly Agree. There are numerous ways of doing this. However, the simplest is to treat the data as being multiple response and count multiple values. For example, count the 4 and 5 values (assuming they correspond to Somewhat Agree and Strongly Agree). This can be done in most programs by recoding the existing variables so that they have only two values (E.g., 0 and 1) and then treating the data as if from a multiple response question. In the case of Q and Displayr they both permit the specification of multiple counted values (i.e., there is no need to recode the variables in these programs).

Missing values

The following table shows the data for the first 10 of 498 respondents from the Mobiles Example. Note that, for example, the first respondent has only provided data for brands 1, 2 and 7 (i.e., has missing values for all of the others). This data is from a question where people were presented with a list of brands and asked which they had shown before (i.e., it is an Aided Awareness question). Where respondents have missing values (shown as a .) this is because they had indicated in an earlier Unaided Awareness question that they were aware of the brands.

Q6_1	Q6_2	Q6_3	Q6_4	Q6_5	Q6_6	Q6_7	Q6_8
Yes	No	.	.	.	.	Yes	.
Yes	No	Yes	.	Yes	.	Yes	Yes
Yes	Yes	.	.	.	.	Yes	.
No	No	Yes	.	Yes	.	No	Yes
No	No	Yes	.	.	.	.	.
Yes	Yes	Yes	.	.	Yes	Yes	.
Yes	No	Yes	.	.	.	.	.
Yes	Yes	.	.	Yes	.	Yes	.
Yes	No	No	.	No	.	No	.
Yes	Yes	Yes	.	Yes	.	Yes	.

The only way to compute a valid summary table of this multiple response question is if the data has been set up so that the analysis program knows that the correct interpretation is that a person is aware of a given brand if either they have said Yes, or, they have missing data. By default no analysis program will work this out. For example, the resulting summary tables in SPSS and Q are shown below. Both are incorrect in this instance.

SPSS multiple response summary table with missing data

$aided Frequencies

		Responses		Percent of Cases
		N	Percent	
Aided Awareness[a]	AAPT	78	21.7%	73.6%
	New Tel	37	10.3%	34.9%
	One-tel	79	22.0%	74.5%
	Optus	10	2.8%	9.4%
	Orange	50	13.9%	47.2%
	Telstra	18	5.0%	17.0%
	Virgin Mobile	59	16.4%	55.7%
	Vodafone	28	7.8%	26.4%
Total		359	100.0%	338.7%

a. Dichotomy group tabulated at value 1.

Q multiple response summary table with missing values

% n Base n	%	n	Base n
AAPT	17%	78	454
New Tel	8% ↓	37	489
One-tel	21%	79	374
Optus	19%	10	54
Orange	18%	50	284
Telstra	29%	18	62
Virgin Mobile	15%	59	389
Vodafone	22%	28	125
NET	NaN%	0	0

base n = 0 to 489

To better understand the data and how to compute valid percentages it is helpful to look at the following table, which indicates the number of respondents to have data of `Yes`, `No` or `Missing` data for each option. Looking at the SPSS table above, the 21.5% shown for Responses `AAPT` has been computed by dividing the 78 people that said `Yes` for `AAPT` by the total number of Yes records for all the brands. Less obviously, the 71.6% shown for `Percent of cases` has been computed by dividing 78 by 106, where 106 is the number of people to have a `Yes` response in the data for at least one of the brands (this number is not shown on the table and cannot be deduced from the table). In the presence of missing data neither of these statistics has a useful meaning (i.e., they are not estimates that relate to the population of phone users).

The table computed by Q (above and to the right) shows 17.6% for AAPT. This has been computed by dividing the 78 by the total number of people to have said either Yesor No for AAPT (i.e., 17.6% = 78/(376 + 78)). This percentage does have a real meaning which is useful in some contexts. The interpreation is that 17.6% of people asked whether they were aware of AAPT said they were aware. However, in this specific example, where we know that everybody with missing data was aware of AAPT, Q's default calculation is also unhelpful.

	Missing data	No	Yes	Total
AAPT	44	376	78	498
New Tel	9	452	37	498
One-tel	124	295	79	498
Optus	444	44	10	498
Orange	214	234	50	498
Telstra	436	44	18	498
Virgin Mobile	109	330	59	498
Vodafone	373	97	28	498
Vodafone	373	97	28	498
Total	1753	1872	359	3984

With this type of data the correct calculation is to compute the aided awareness as the proportion of people to have said Yes or having missing data and divide this by the total number of people in the study. In the case of AAPT, for example, the correct proportion is 24.5% (i.e., (44 + 78) / 498). The following table shows the correct proportions for all of the brands:

Brand	**Aided Awareness**
AAPT	24%
New Tel	9%
One-tel	41%
Optus	91%
Orange	53%
Telstra	91%

Virgin Mobile	34%
Vodafone	81%

Using software to compute the proportions correctly

The standard way to fix the data is to:

1. Recode the variables so that the missing values are recoded as having a value of 1.
2. If it is not already in the data, create a none of these alternative (this is necessary because SPSS and some of the older analysis packages require that each respondent has at least one `Yes` response in order for the percentages to be correctly calculated).

The standard method can be done in Q and Displayr as well, but both of these programs have an easier way of fixing this problem.

Computing the correct proportions in Displayr

1. Select a variable set which corresponds to a multiple-response question under **Data Sets**.
2. On the right, select **Properties > DATA VALUES > Missing Data**, select **Include in Analyses** for each of the categories and press **OK**.
3. On the right, select **Properties > DATA VALUES > Select Categories** and ensure that `Yes` and `Missing` data are selected and press **OK**.

Computing the correct proportions in Q

1. In the **Tables** tab, right-click on one of the categories of the question and select **Values**.
2. Fill in the dialog box as shown below.

	Label	Missing Data	Count This Value
1	Missing data	☐	☑
2	No	☐	☐
3	Yes	☐	☑

Examples

Mobiles

This study is from the Australian mobile (cell) phone market.

The questionnaire: Mobiles Questionnaire.pdf

SPSS data file: Mobiles.sav

Colas

This study is from the cola market.

The questionnaire: Colas Questionnaire.pdf

SPSS data file: Colas.sav

Advanced Data Analysis

- Predictive modeling
- Notes
- Segmentation
- Perceptual mapping
- Identifying correlated variables

Advanced Data Analysis

Predictive modeling

Predictive modeling involves working out how one set of variables predicts another variable. There are two 'flavors' of this type of analysis that are commonplace: driver analysis and targeting.

Driver analysis

Driver analysis seeks to work out the relative role that different drivers of preference play in a market. For example, if there is one question that measures overall satisfaction and another question which measures satisfaction with different areas of a company (e.g., customer service, price, quality), then driver analysis can be used to quantify the relative impact of each of the overall areas on overall preference. Driver analysis is usually conducted using Regression. Choice Modeling is a type of driver analysis.

Regression

Regression is a statistical tool for quantifying a model. The key output of regression is a formula, such as:

```
$Sales = $121 + 4.1 × $TV Advertising Expenditure + 3.2 × $Online Advertising Expenditure
```

With such a formula it is possible to:

- Make *predictions*. For example, using the formula above, a firm that spends $1,000,000 on TV advertising and nothing on advertising expenditure is predicted to have sales of $121 + 4.1 × $1,000,000 = $4,100,121.
- Draw conclusions about *relative* effectiveness. For example, the equation above shows that TV advertising is more effective than online advertising.

Most regression models are characterized by having one *dependent variable* and one or more *independent variables*.

In the example above the dependent variable is sales. Common dependent variables in survey analysis applications of regression include:

- Overall satisfaction with a product or a service.
- Likelihood to recommend a product or service.
- Net Promoter Score.
- Likelihood to use a product or service again.
- Product quality.
- Frequency of buying or using a product or service.

The *independent variables* will typically be a list of variables that are believed to determine the value of the dependent variable. In the example above, the independent variables are $TV Advertising Expenditure$Online Advertising Expenditure</tt>. In most applications of regression to survey analysis, the independent variables are either:

- Demographic variables. For example, if wishing to identify high value customers, the dependent variable may be amount of money spent and the independent variables would be demographics.
- Measurements of performance in different areas. For example, if the dependent variable measures satisfaction with an airline the independent variables could be things such as satisfaction with the food, satisfaction with the cabin crew, satisfaction with the in-flight entertainment, and so on. Such a regression model is known as a *driver analysis*.
- Measurements of effort in different areas. For example, expenditures on different types of advertising, such as in the above example.

Targeting

Targeting, which seeks to identify the distinguishing characteristics of a group of people of particular interest (e.g., if wanting to understand the demographic profile of the heaviest users of a service). There are many dozens of techniques in widespread use for targeting. However, most of these techniques are very complicated and easy to misinterpret. The only one of these tools that can readily be used by non-statisticians is Predictive Trees.

Predictive Trees

A *predictive tree* is an analysis that looks like an upside down tree. Although it looks quite complicated this 'tree' is just a graphical representation of a table. The term *predictive tree* is not a standard term; the same basic idea has lots of different brand names, including: *Rpart* (R), *Tree* and *AnswerTree*

(SPSS) and *CHAID* (Statistical Innovations), *CART*, regression trees, classification trees and *decision trees*.

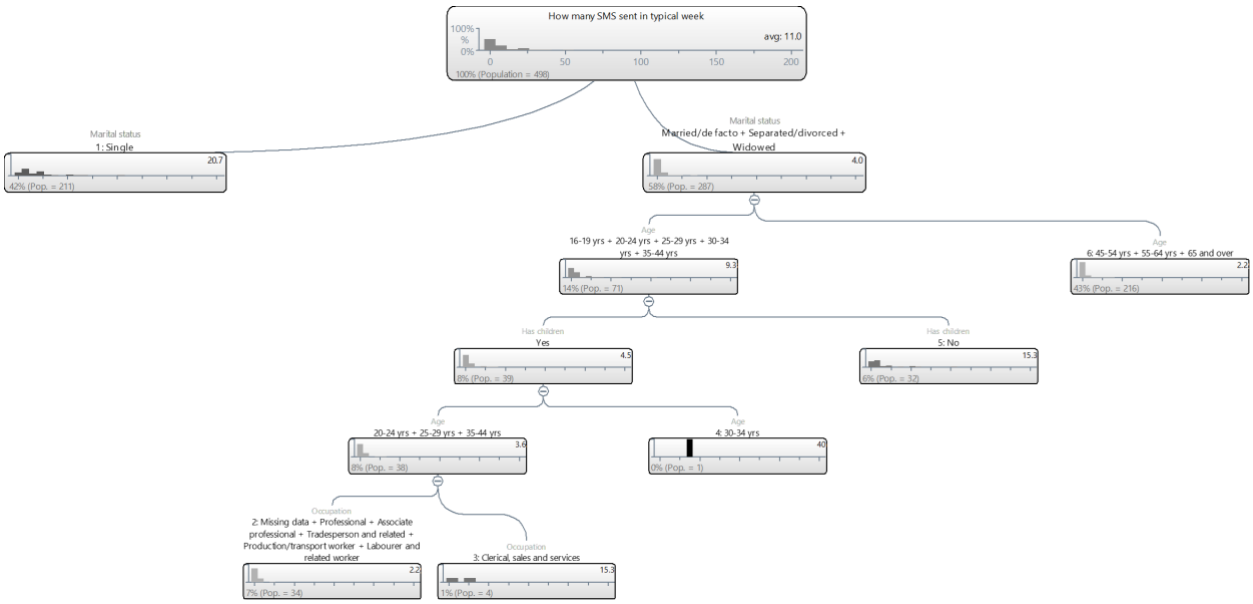
Example

The tree below predicts the number of SMS of people in the Mobiles example study. The top box on the tree shows the data for the entire sample (i.e., 100%). It shows an average of 11.0 text messages per person and the histogram (column chart) shows that the most common answer is around 0. From this tree we can see that the first *predictor variable* is Marital status: people that are Single sent an average of 20.7 SMSs, compared to 4.0 among the non-single people.

As the Single category is not split further this tells us that once we know that somebody is single then we can make a prediction without any further information, where the prediction is that a single person will make 20.7 SMS. From the histogram we can see that there is still considerable variation among these single people.

Looking now at the people that are not single (i.e., Married/de facto + Separated/Divorced + Widowed), this group is split by age and we can see that people that are Married/de facto + Separated/Divorced + Widowed and are aged under 45 have an average number of SMS per month of 9.3, compared to 2.2 among those that are aged 45 or more and are Married/de facto + Separated/Divorced + Widowed.

Looking further down the tree we can see that this younger group is split further by whether or not they have children, age again and occupation.



Viewing the tree as a table

The tree above can also be expressed as an (admittedly ugly) table:

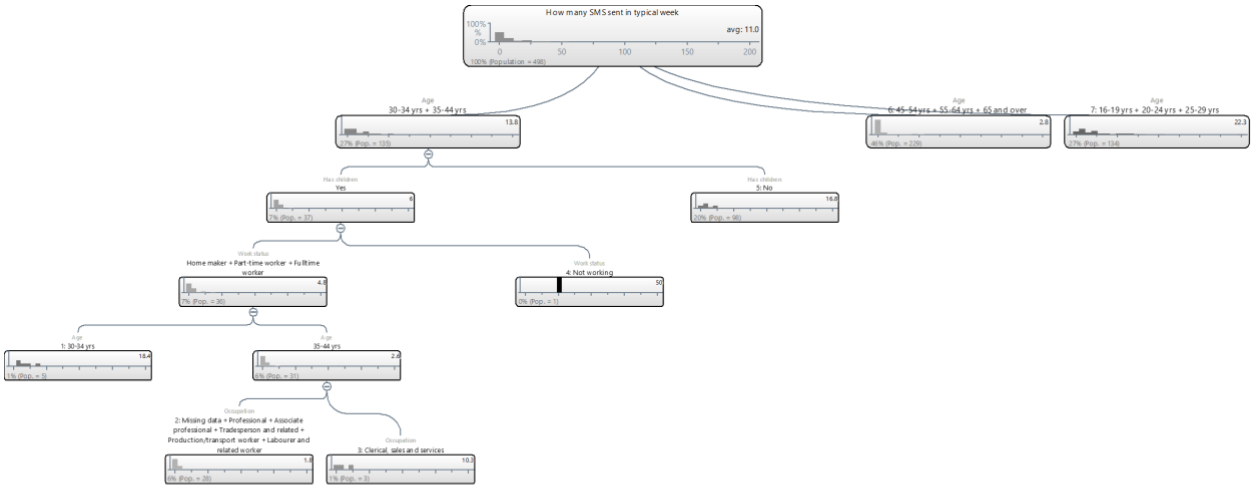
Average Column n	Marital status Married/de facto + Separated/divorced + Widowed Age 16-19 yrs + 20-24 yrs + 25-29 yrs + 30-34 yrs + 35-44 yrs Has children Yes Age 20-24 yrs + 25-29 yrs + 35-44 yrs Occupation 1: Single	2: Missing data + Professional + Associate professional + Tradesperson and related + Production/transport worker + Labourer and related worker	3: Clerical, sales and services	4: 30-34 yrs	5: No	6: 45-54 yrs + 55-64 yrs + 65 and over	NET
How many SMS sent in typical week	20.7↑ 211	2.2↓ 34	15.3 4	40.0 1	15.3 32	2.2↓ 216	11.0 498

The strengths and weaknesses of predictive trees

The key strength of predictive trees is ease-of-use. Provided that you apply a bit of commonsense and take the time to learn how to use the software that you are using, it is hard to go particularly wrong with a tree. Predictive trees are the closest thing that there is to an idiot-proof advanced analysis technique. By contrast, regression, which can be used for similar problems, is frequently used incorrectly by very experienced researchers.

The major limitations of predictive trees are that:

- They are at their best with very large sample sizes. Predictive trees with less than a few hundred observations are often not useful. This is because each time the tree splits the sample size also splits, so with small sample sizes the trees end up having only one or two variables.
- Predictive trees cannot be used to make conclusions about which variables are strong and which are weak predictors. For example, looking at the Single group a natural, but incorrect, interpretation is that age is irrelevant because it is not used to split this group further. However, such an interpretation is not correct because it ignores that being single is correlated with being young. To appreciate this, we can exclude marital status and re-grow the tree. The revised tree is shown below. Note that now Age is shown to be the first predictor of number of SMS per week and it seems to be close to be similar in its predictive accuracy to marital status, with predictions ranging from 2.8 to 22.3, compared to from 4.0 to 20.7 for marital status.[note 1]



Software

Most advanced statistical techniques are reasonably standard, with different programs providing basically the same outputs. This is not the case with predictive trees. Most software companies create their trees in different ways and the distinctions between the trees created by different programs can be large.

All of the common predictive tree programs are technically described as being *recursive divisive* algorithms, which is a fancy way of saying that they:

1. Find the variable that best predicts what is being predicted.
2. Split the sample into groups that are relatively similar in terms of this variable (e.g., in the tree at the top of this page, the groups were defined based on marital status).
3. For each of the new groups, repeat steps 1 and 2, continually splitting each group and further splitting until no more splitting is practical. (In computer science and logic the idea of continually reapplying the same process is called *recursion*).

The key differences between the different bits of software are:

- What they can predict. Most programs will only predict a single categorical or a single numeric variable. The major exception to this is Q, which can simultaneously predict multiple numeric, categorical and other exotic data types (e.g., Choice Modeling data).

- How many groups they form each time they do a split. In the examples above, the program (in this case *Q*), tries to optimally work out the number of groups at each *level* of the tree. A

similar approach is used by SPSS's *Tree*, *AnswerTree*, *CHAID* and *Displayr*. The alternative approach is to always split into two groups; this approach is used in CART.

- What rules they use to determine when to stop further splitting groups in the tree. Each of the programs has various ways of doing this, including:
    - Requiring that additional splits improve the predictive accuracy of the tree (different programs implement in this in different ways; *CHAID* uses significance tests, *Displayr* and *Q* use information criteria and CART usually uses *cross-validation*).
    - Specifying a minimum sample size after which a group cannot be further split.
    - Specifying a maximum number of *levels* of the tree.
- The number of options available. *Displayr*, for example, only lets the user choose between a small and a bigger tree, SPSS's *tree* let's the user choose between various tree algorithms (e.g., different variants of CHAID and CART), R's rpart can be customized in pretty much any way that is desired if the user is skilled enough and *Q*lets the user change the predictor variables available within each group.

Notes

1. Jump up↑ Indeed, it may look that age is a better predictor, as it has a larger range. However, in this case occupation is preferred because it has a fewer number of categories with broadly similar predictive accuracy; this is an example of the principle of *parsimony* being taken into account when creating a model.

Segmentation

Segmentation involves finding groups of people that have given similar responses in a survey with the goal being that similar strategies can be developed for each of the *segments* of similar people. For example, if a study has been conducted looking at the type of leisure activities that people undertake then segmentation will identify groups of people that undertake similar types of leisure activities. There are three main ways that people segment: *judgment*, *cluster analysis* and *latent class analysis*.

Judgment

When judgment is used to create segment it usually involves selecting a single variable, or a combination of a small number of variables, that are known to relate to many of the key variables in the survey. For example, if the crosstabs revealed that age was correlated with many of the other questions in the survey than it may make sense to segment using age.

Cluster analysis

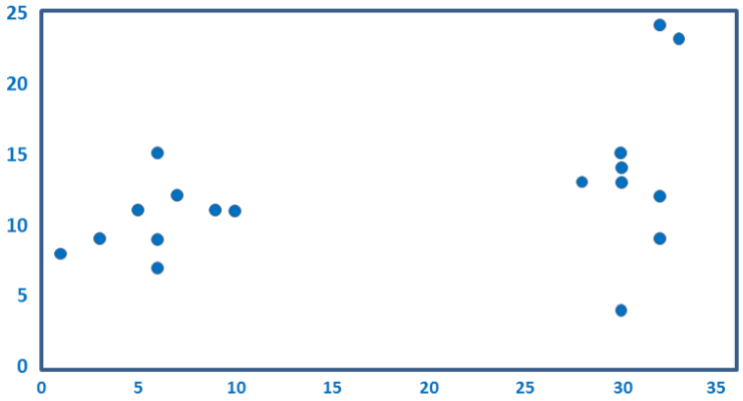
The traditional approach to conducting segmentation has been to use Cluster Analysis. Cluster analysis assumes that:

- There is no missing data (i.e., each respondent has provided data on all the variables.
- All the variables are numeric.
- All the variable have the same range (e.g., the same highest and lowest values).

It is often the case that one or more of these assumptions is not met. There are various techniques that can be used to try and overcome these assumptions.

The basics

Cluster analysis finds groups of *similar* respondents, where respondents are considered to be similar if there are relatively small differences between their average ratings. As an example, look at the plot below. It is a scatterplot showing data for 18 respondents on two numeric variables. You can hopefully see that the dots fall into two groups. If you used cluster analysis to analyze this data, provided you didn't do something wrong, it would also identify the same two groups that you can see. These groups are then called *clusters*.



Typically we have many more than two variables and this makes it difficult for us to visualize the data and determine the clusters. It is in such situations that we need to use cluster analysis. Consider the data below, which shows the responses of 20 respondents to 7 variables (this is from SPSS). Can you see a pattern? Perhaps you can find one if you work hard. However, these 20 respondents are from a total sample of 498 respondents and few people could read a table showing all of their data and identify any clusters, which is why cluster analysis (or, better yet, Latent Class Analysis) is used to find clusters in the data. SurveyAnalysis.org has a detailed discussion of how cluster analysis works.

	q10a	q10b	q10c	q10d	q10e	q10f	q10g
1	Neither	Agree a little	Strongly di...	Strongly di...	Disagree a...	Disagree a...	Agree a little
2	Agree a little	Strongly a...	Neither	Disagree a...	Agree a little	Strongly a...	Neither
3	Neither	Agree a little	Agree a little	Disagree a...	Disagree a...	Neither	Agree a little
4	Disagree a...	Disagree a...	Agree a little	Disagree a...	Disagree a...	Strongly di...	Agree a little
5	Neither	Neither	Strongly di...	Disagree a...	Disagree a...	Neither	Neither
6	Agree a little	Strongly a...	Agree a little	Disagree a...	Strongly a...	Strongly a...	Disagree a...
7	Agree a little	Neither	Disagree a...	Agree a little	Agree a little	Agree a little	Agree a little
8	Agree a little	Neither	Agree a little	Agree a little	Neither	Agree a little	Agree a little
9	Strongly a...	Neither	Disagree a...	Disagree a...	Agree a little	Strongly di...	Strongly a...
10	Strongly a...	Strongly a...	Strongly a...	Strongly a...	Agree a little	Strongly a...	Agree a little
11	Neither	Strongly a...	Agree a little	Strongly a...	Neither	Strongly a...	Strongly a...
12	Neither	Agree a little	Agree a little	Agree a little	Neither	Disagree a...	Disagree a...
13	Neither	Agree a little	Neither	Agree a little	Agree a little	Disagree a...	Disagree a...
14	Strongly di...	Agree a little	Strongly a...	Strongly di...	Disagree a...	Disagree a...	Agree a little
15	Neither	Disagree a...	Strongly di...	Agree a little	Strongly a...	Agree a little	Neither
16	Agree a little	Strongly a...	Disagree a...	Strongly di...	Disagree a...	Strongly a...	Neither
17	Neither	Strongly a...	Neither	Disagree a...	Strongly di...	Neither	Agree a little
18	Agree a little	Strongly a...	Agree a little	Neither	Strongly di...	Neither	Strongly di...
19	Disagree a...	Agree a little	Agree a little	Disagree a...	Strongly a...	Strongly a...	Agree a little
20	Strongly a...	Strongly a...	Strongly a...	Agree a little	Neither	Neither	Neither

Example

This example analyzes the data that was previewed at the end of the previous section. You can download the SPSS data file and questionnaire if you wish to replicate the example. The data contains seven variables measuring attitudes towards mobile phones, where respondents were asked to give their degree of agreement/disagreement with the following statements:

- Technology is fascinating
- I am often surprised by the size of bills
- I find it difficult to determine best deal
- I spent a lot of time shopping for best deal
- I closely monitor the time I spend on the phone
- Cost is a factor when deciding where to SMS or phone
- I try to keep calls short and to the point

The scale used was:

`Strongly agree Agree a little Neither Disagree a little Strongly disagree DON'T KNOW`

Consequently, prior to analyzing the data the data needed to be prepared in two ways:

1. People that said `DON'T KNOW` were assigned missing value codes in the corresponding variables.
2. The data was recoded so that a value of 1 was assigned to `Strongly disagree`, 2 to `Disagree a little` and so on up to 5. Note that these precise values are arbitrary as there is no good reason why these categories should be assigned values from 1 to 5 with a space of 1 between each scale point. While arbitrary, this is, nevertheless, standard practice.

SPSS's **Two Step Cluster** analysis routine, which is the best of the cluster analysis techniques that is available in SPSS,[note 1] recommends the following five cluster solution.

The top row of the table shows the sizes of the clusters. We can see that approximately 25% of the sample is in the first cluster, 22% in the second and so on.

The variables are then listed underneath each cluster in the order of their importance in determining cluster membership. Looking at cluster 1 we can see that the variable shown at the top is `Difficult to determine best deal`. As its average value is 1.92 and, as a 2 represents `Disagree a little` we can conclude that people in cluster 1 find it relatively easy to identify the best deal. Looking elsewhere in the table we can see that all the other clusters found it hard to find the best deal (i.e., as

their averages are much higher than 1.92). Similarly, if we look at cluster 2 we can see that the most important determinant of membership of this cluster was level of agreement with the statement `Closely monitors time on phone.`

Size	24.9% (119)	22.0% (105)	20.8% (99)	18.0% (86)	14.3% (68)
Inputs	Difficult to determine best deal 1.92	Closely monitors time on phone 4.42	Spent a lot of time shopping for best deal 1.90	Surprised by bill size 4.55	Calls kept short and to the point 1.82
	Closely monitors time on phone 3.60	Calls kept short and to the point 4.68	Technology fascinating 2.14	Difficult to determine best deal 4.29	Closely monitors time on phone 1.74
	Technology fascinating 3.54	Difficult to determine best deal 4.28	Difficult to determine best deal 4.12	Spent a lot of time shopping for best deal 3.86	Cost is a factor when deciding to SMS or phone 1.94
	Calls kept short and to the point 3.99	Spent a lot of time shopping for best deal 3.71	Calls kept short and to the point 4.18	Closely monitors time on phone 2.49	Spent a lot of time shopping for best deal 2.37
	Cost is a factor when deciding to SMS or phone 3.34	Technology fascinating 3.88	Surprised by bill size 3.13	Calls kept short and to the point 3.47	Difficult to determine best deal 3.06
	Surprised by bill size 3.42	Cost is a factor when deciding to SMS or phone 3.83	Closely monitors time on phone 2.73	Technology fascinating 3.41	Technology fascinating 3.41
	Spent a lot of time shopping for best deal 2.75	Surprised by bill size 3.74	Cost is a factor when deciding to SMS or phone 2.67	Cost is a factor when deciding to SMS or phone 3.03	Surprised by bill size 3.62

In a real-world study the basic process from this point would be to:

- Create a summary of the unique aspects of each of the clusters.
- Run crosstabs of the clusters against other interesting tables to see if there are any further interesting relationships.
- Examine alternative cluster analysis solutions. Generally it is useful to explore solutions with fewer clusters than those that are automatically suggested. In this case, a five cluster solution was automatically identified so it would be advisable to also review solutions with four, three and two clusters. (Note that there is no truly scientific method for determining the number of clusters and that all automated methods for selecting the number of clusters really do is identify the maximum number that is likely to be sensible.)
- Come up with evocative names for each of the clusters (e.g., "The big talkers").

Latent class analysis

Latent class analysis is essentially an improved version of Cluster Analysis and it is used for the same types of things as is cluster analysis. In the main, in survey analysis, this involves findings segments. Latent class analysis improves on cluster analysis in two important ways:

- It is able to handle many different types of data (e.g., rankings, ratings, Choice Modeling).
- It automatically addresses Missing Values.

Example

The output below has been created by *Q* (and is almost identical to that created by *Displayr*). Seven variables from the Mobiles study have been used in the analysis: six are numeric variables measuring attitudes (see Cluster Analysis for a discussion of this data), and the seventh is a categorical variable measuring the average monthly bill.*Q* has automatically identified two segments. The first segment contains 60% of the market and they are people with relatively lower average bills whose attitudes reveal that they are relatively cost focused. The second segment predominantly contains people with higher monthly spends who are less concerned with price but are more likely to be surprised by their bill size.

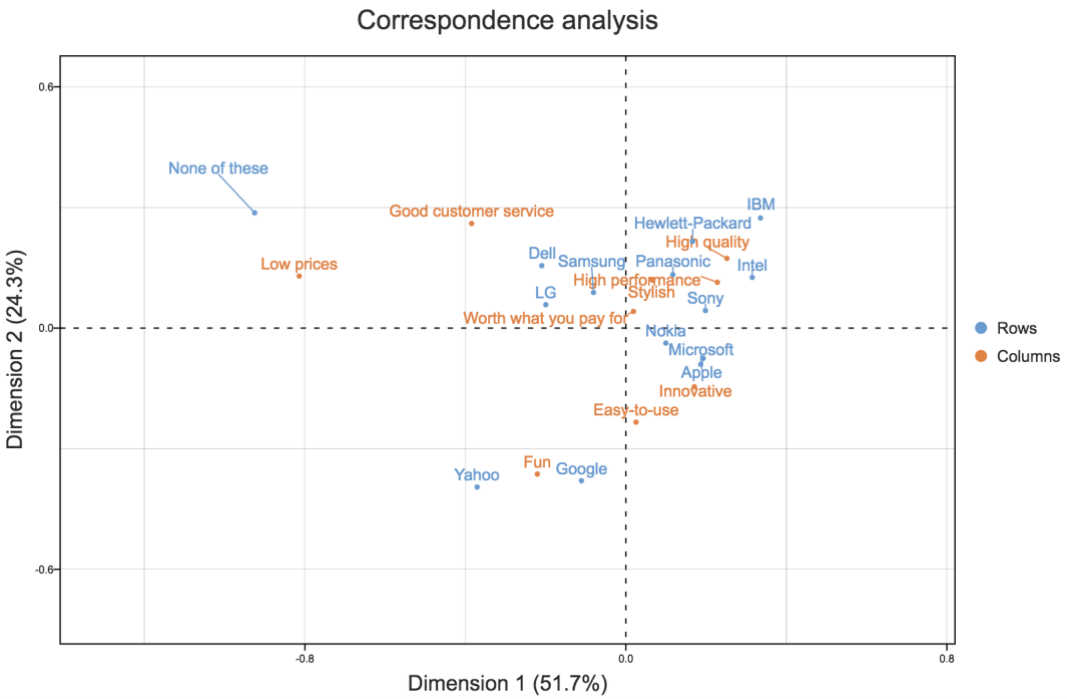
Selection of the number of segments

A trade-off needs to be made when selecting the number of segments. The more segments one has, the greater the extent to which the analysis reflects the diversity observed in the data. This suggests that having a large number of segments is desirable. But, the more segments one has, the greater the risk that the diversity that is identified is meaningless diversity, just reflecting the properties of the specific data used in the analysis rather than the true diversity in the world at large. This suggests it is preferable to have fewer segments.

Various heuristics have been developed that attempt to find sweet spot in this trade-off. The most widely used is the *Bayesian Information Criterion (BIC)*, which computes a value for each possible number of segments and recommends that the best segmentation is the one with the lowest *BIC*. It should be kept in mind that heuristics such as this are just rules of thumb, and it is not appropriate to assume that the number of segments identified by the *BIC* is in any sense optimal. While caution should be exercised if choosing a greater number of segments than suggested by the *BIC*, there is often good reason to select a smaller number of segments, as this is often more practical in terms of implementing the segmentation.

Perceptual mapping

Perceptual maps are charts that show the the relationships between different categories in a survey. For example, the *map* below shows how different brands of cola are perceived. Most perceptual maps used in survey analysis are created using Correspondence Analysis. The example below was created in Displayr



Identifying correlated variables

Sometimes it is useful to understand how if there are correlations between large numbers of variables. For example, this can be useful to:

- Understand how attitudes and/or behaviors are interrelated.
- Identify redundant questions in a questionnaire if there is a need to simplify it.
- Identify redundant concepts in form of a concept test.
- Summarize data.
- *Transform* data prior to the application of other multivariate techniques (e.g., Cluster Analysis and Regression).

The identification of correlated variables is generally done using Principal Components Analysis, but most of the time when researchers conduct principal components analysis they refer to it as *factor analysis* (which is technically a different technique, but the differences are trivial).

Principal components analysis identifies variables that are highly correlated with each other.

Principal Components Analysis

Principal components analysis identifies variables that are highly correlated with each other.

Example

This example analyzes seven variables measuring the extent of agreement with the following statements:

- Technology is fascinating
- I am often surprised by the size of bills
- I find it difficult to determine best deal

- I spent a lot of time shopping for best deal
- I closely monitor the time I spend on the phone
- Cost is a factor when deciding where to SMS or phone
- I try to keep calls kept short and to the point

Please refer to Cluster Analysis for more detail on this data and how it has been processed prior to being used in analysis. The key output of principal components analysis is the *rotated component matrix*, such as the one shown below (computed using SPSS). The seven variables that have been analyzed can be reduced to three variables. The seven original variables are shown in the rows and the three new variables are represented by the columns and are referred to as *components*. The numbers in the table are correlations but when conducting principal components analysis they are typically referred to as *loadings*.

Rotated Component Matrix[a]

	Component		
	1	2	3
Closely monitors time on phone	.81	.11	-.07
Calls kept short and to the point	.69	-.20	.18
Cost is a factor when deciding to SMS or phone	.67	.10	-.07
Technology fascinating	.06	.78	-.26
Surprised by bill size	-.15	.66	.38
Spent a lot of time shopping for best deal	.36	.46	.31
Difficult to determine best deal	.01	.01	.90

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

This output is interpreted as follows:

- The first component is highly correlated with Closely monitors time on phone and reasonably strongly correlated with Calls kept short and to the point and Cost is a factor when deciding to SMS or phone. Thus, this first component seems to relate to reflect a cost sensitivity dimension (that is, it suggests that a key distinction between people is that they vary in terms of how concerned they are about price).
- The second dimension is a bit more difficult to interpret. There are relatively high loadings for fascination with technology and being surprised by bill size. Why these two variables would be correlated is difficult to explain and is perhaps more of a weird fluke than a profound insight (not that because these two variables are both correlated with the second component this implies that they are also correlated with each other).
- The third component essentially only reflects agreement with the idea that it is difficult to determine the best deal.

- The variable Spent a lot of time shopping or the best deal is not highly correlated with any of the components and thus is largely ignored.

The real conclusion to draw from this analysis is that the principal components analysis has failed to identify much that is interesting. Perhaps a more interesting solution could be found by investigating more components (e.g., four or five), but it is unlikely that this would be so useful as having, say, a five component solution instead of the original seven variables is not much in the way of a simplification of the data.

Software

Almost all applications of principal components analysis in survey analysis employ a *varimax rotation*.

Product	How to do it
Q	2. In the **Variables and Questions** tab select all the variables that you wish to use, right-click and select **Set Question** and specify a **Question Type** of **Number - Multi**. 3. On the **Tables'** tab select **Create > Traditional Multivariate Analysis > Principal Components Analysis**, select the question and press **OK** .
SPSS	4. **Analyze > Dimension Reduction > Factor**. 5. Select the variables you wish to analyze. 6. **Rotations > Varimax > Continue > OK**.

Rotated Component Matrix[a]

	Component		
	1	2	3
Closely monitors time on phone	.81	.11	-.07
Calls kept short and to the point	.69	-.20	.18
Cost is a factor when deciding to SMS or phone	.67	.10	-.07
Technology fascinating	.06	.78	-.26
Surprised by bill size	-.15	.66	.38
Spent a lot of time shopping for best deal	.36	.46	.31
Difficult to determine best deal	.01	.01	.90

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Reporting

There is no standard approach to reporting market research. Partly this is because different studies have different requirements. It is also because people like to work in basic ways. However, an approach to structuring reporting that can work in most situations is:

1. Conduct all the analyses listed in the Analysis Plan and any additional analyses.
2. Use Data Reduction to distill the essence of each of the analyses.
3. Creating charts.
4. Structuring and Sharing Findings

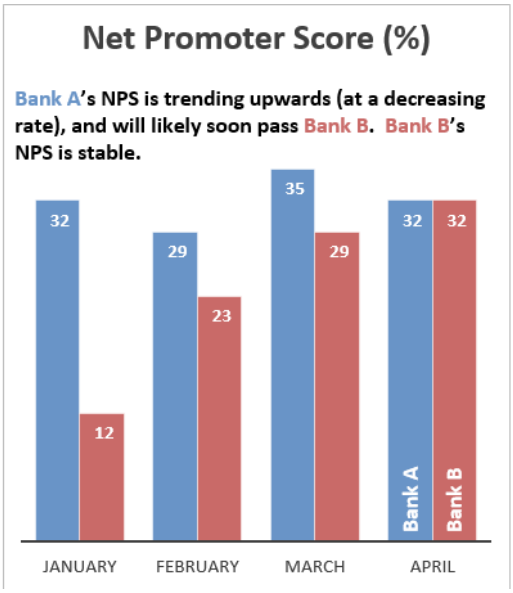
Reporting

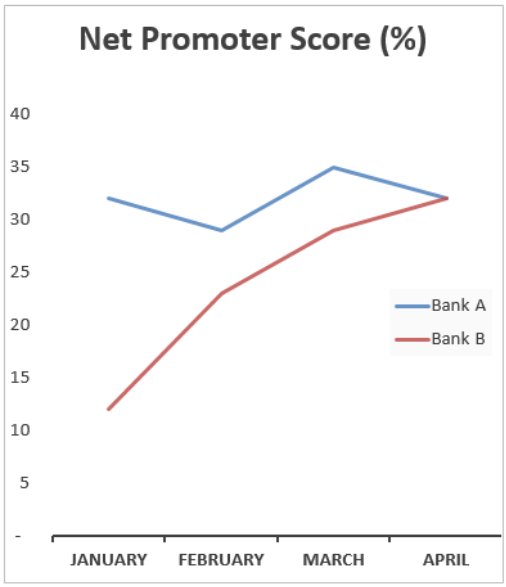
Charts

Our brains are much more efficient at processing images than at processing numbers. By presenting tables as charts we can make it much easier for people to interpret findings. Plotting of Summary Tables using pie charts and column charts is straightforward. Unfortunately, with more complex tables, it is not the case that any chart is better than a table. Our brains are only good at interpreting images when the images are structured in particular ways.

Use columns for comparisons and lines for trends

The two charts below show the same data. When we look at the plot on the left our brain automatically tries to compare the height of the columns. If we look hard enough and read the comment we can see a trend. However, when we look at the plot on the right the trend is so clear it does not need to be described. The general conclusion is to use columns when wanting to emphasize and compare individual results and lines when wanting to emphasize and compare trends.



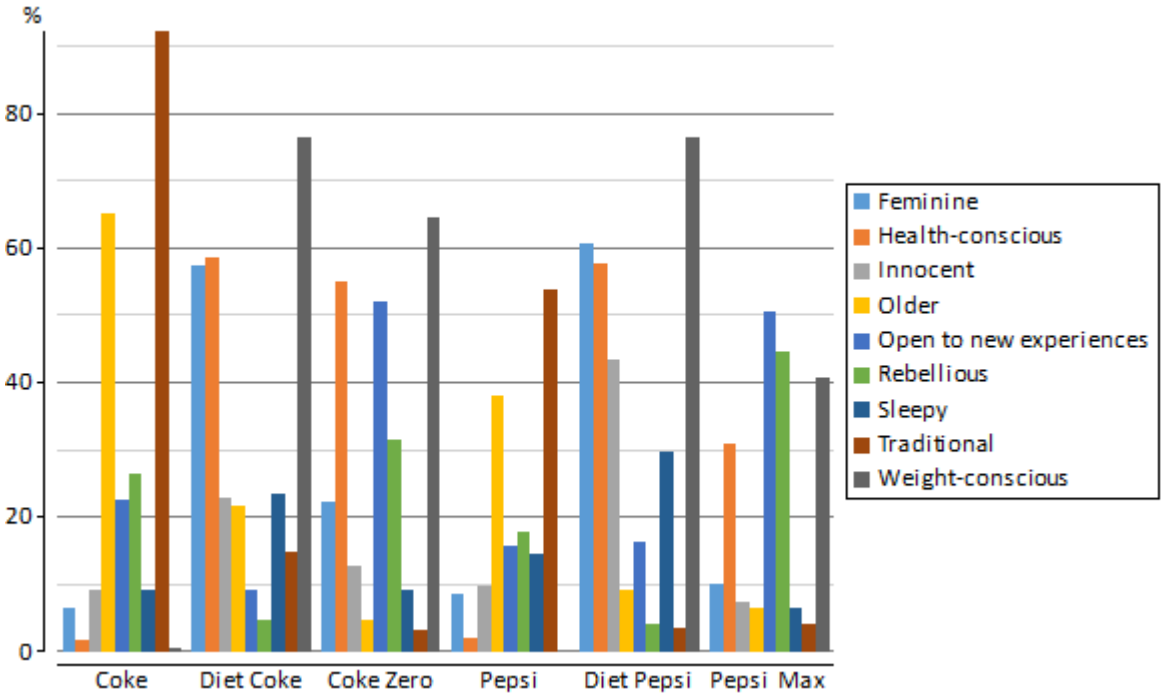


Minimize difficult comparisons by using grid layouts

The table below is large and complicated. It is not straightforward to examine it for conclusions.

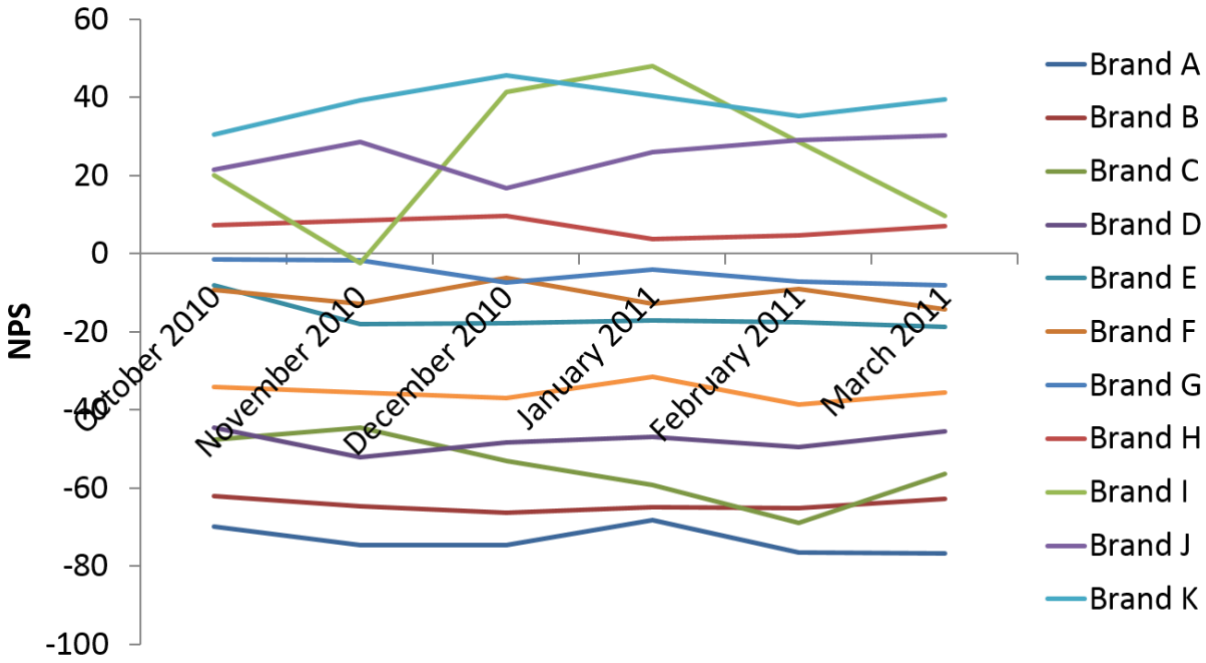
%	Feminine	Health-conscious	Innocent	Older	Open to new experiences	Rebellious	Sleepy	Traditional	Weight-conscious	NET
Coke	6%	2%	9%	65%	23%	26%	9%	92%	1%	98%
Diet Coke	57%	59%	23%	22%	9%	5%	24%	15%	76%	92%
Coke Zero	22%	55%	13%	5%	52%	31%	9%	3%	65%	91%
Pepsi	9%	2%	10%	38%	16%	18%	14%	54%	0%	79%
Diet Pepsi	61%	58%	43%	9%	16%	4%	30%	3%	76%	95%
Pepsi Max	10%	31%	7%	6%	50%	45%	6%	4%	41%	87%
NET	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A 'standard' way to plot such data is as a column chart, as shown below. While prettier than the table, it is actually less useful. Consider trying to work out which brand is most strongly associated with being Sleepy. On the table we can quickly read down the appropriate column. In the chart below it is much harder to work out.

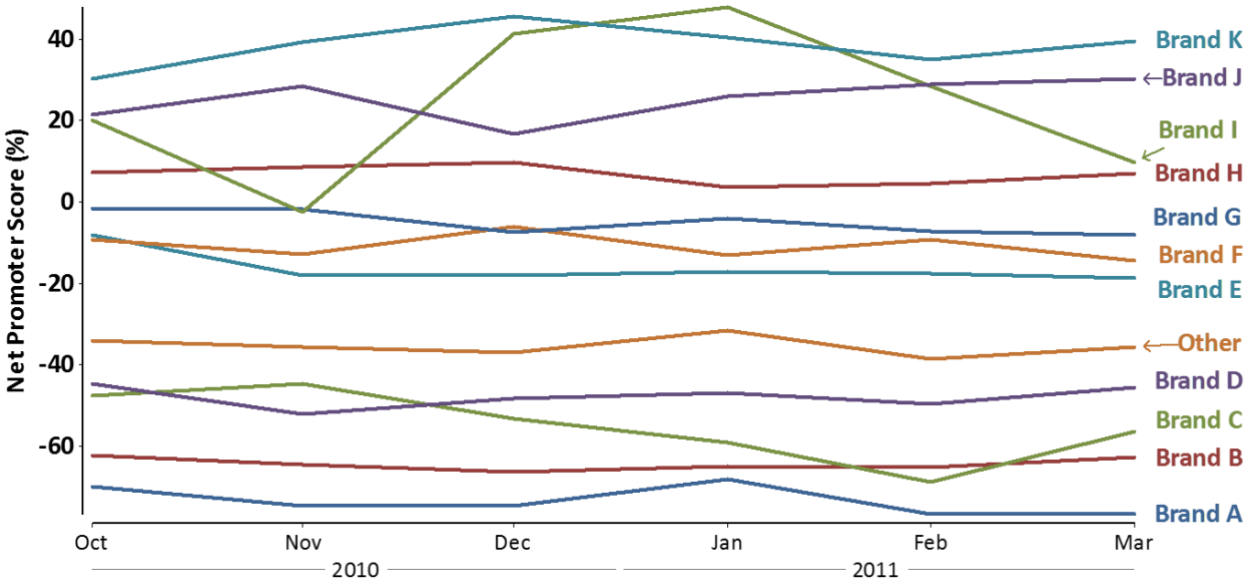


The solution to this charting problem, and to many charting problems, is to create a structure that is more grid-like, as shown below. The reason that this chart is better is that it replaces numbers with bars, without sacrificing the ability to read by rows or columns. Thus, we can quickly scan this chart to identify that, for example, Diet Pepsi scores the highest on Sleepy.

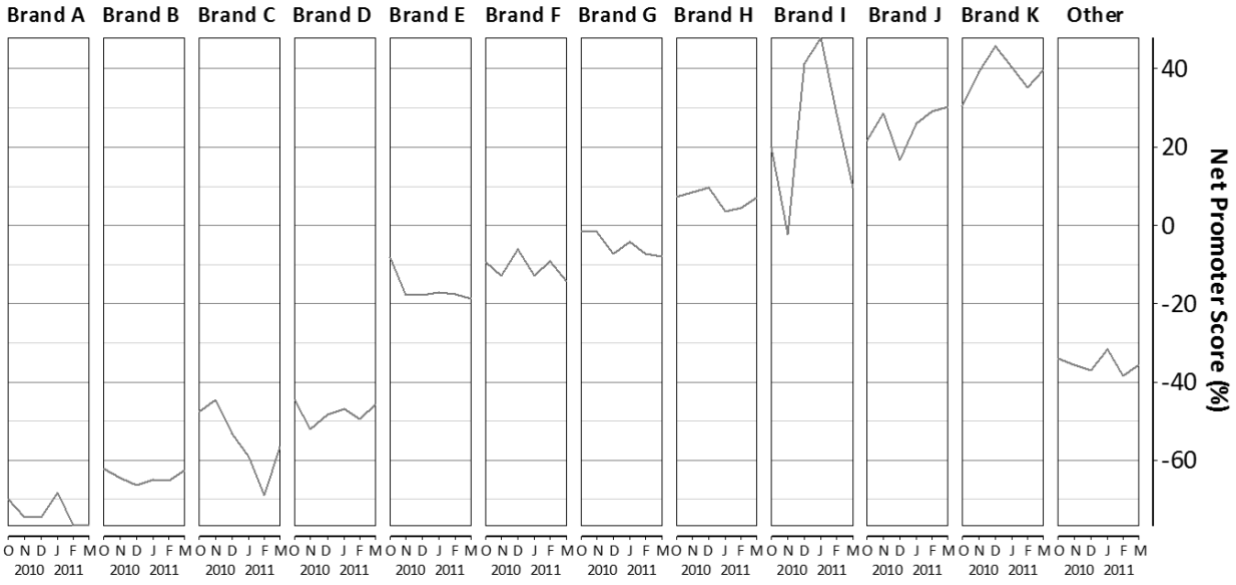
This same basic principle is useful for time series data as well. Consider the default line chart created by Excel, shown below. It is very different to disentangle the meaning in this data.



We can get a small improvement by replacing the legend with individual labels on each line (this makes the comparisons easier because the reader does not have to continually work out which line to use).



However, the chart below, which again adopts a grid like layout, and also orders the results, is substantially easier to interpret.

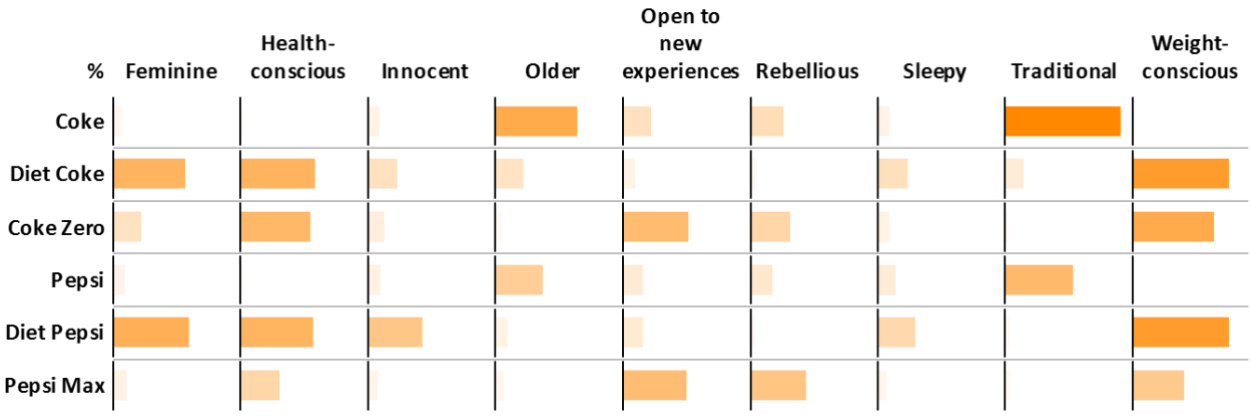


Use color and redundancy to emphasize

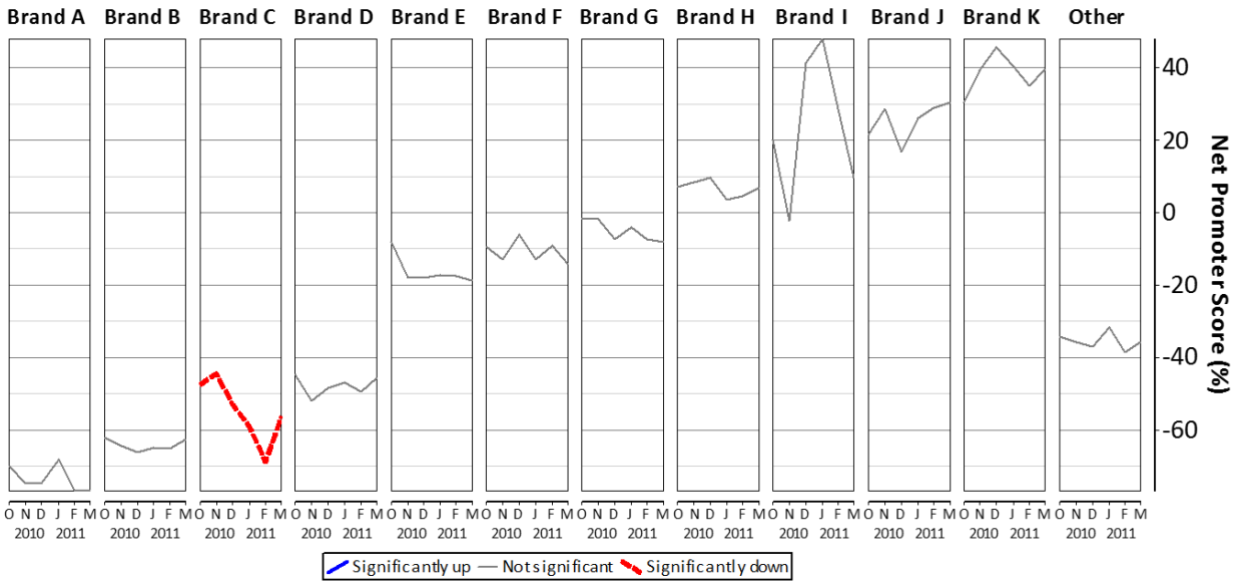
In a traditional chart is used to indicate different series. For example, in this chart blue is used for feminine and health orange for health-consciousness. This use of color is unhelpful. Our brains infer the colors as indicating that the columns are 'different' in some way, which makes it harder for us to

accurately compare the columns and thus color in this context completely undermines the purpose of the chart.

A better use of color is in the chart below. Color is still being used but it is being used in a different way. The intensity of the color is proportional to the length of the bars. This redundant use of information helps our brains (see Ordering for a further improvement of this chart).



The next chart improves on the earlier time series chart. Note that color is used here to emphasize interesting results. In this case, it is used as a way of communicating the results of statistical testing.



Software

Most statistical packages can do the types of plots shown on this page (e.g., *SPSS, R, SAS*). They can also be created with a bit of work in *Excel* (e.g., by creating multiple charts and lining them up, or, by buying add-ins). The simpler plots on this page were created in Excel. The 'grid'-style plots were created in *Q*, which is the only specialist survey-analysis program that implements these styles of plots.

Data Reduction

The easy part of data analysis is creating lots of crosstabs. The hard part is trying to figure out what it all means. Our brains are greatly limited in terms of the quantity of information that they can process. The more information that our brains have to process, the worse the job that they do. Consequently, a secret to interpreting data is to reduce the quantity of data. The academic term for this is *data reduction*.[1] This page illustrates some of the more elementary aspects of data reduction by working through an example.

Deleting

The crosstab below shows how preferences for different brands of cola differ by age, gender and income. It is quite typical of the type used by commercial consultants when analyzing surveys. It is a very large crosstab. Indeed, it is so large that it is impractical for anybody to read everything on this crosstab and thus key findings hidden in the crosstab may easily be missed.

		BANNER																							
		Age								Gender			Income												
Prefered cola		18 to 24	25 to 29	30 to 39	40 to 49	50 to 54	55 to 64	65 or more	NET	Male	Female	NET	Less than $15,000	$15,001 to $30,000	$30,001 to $45,000	$45,001 to $60,000	$60,001 to $90,000	$90,001 to $120,000	$120,001 to $150,000	$150,001 to $200,000	$200,001 or more	Don't know + I'd rather not say	NET	NET
Coca-Cola	%within column	65.1%	41.0%	55.1%	28.3%	46.2%	36.0%	36.4%	44.4%	45.9%	43.0%	44.4%	55.6%	43.8%	56.7%	40.7%	47.0%	40.0%	31.8%	47.4%	57.1%	38.9%	44.4%	**44.4%**
	%within row	19.6%	11.2%	26.6%	11.9%	12.6%	12.6%	5.6%	100.0%	50.3%	49.7%	100.0%	3.5%	9.8%	11.9%	15.4%	21.7%	11.2%	4.9%	6.3%	5.6%	9.8%	100.0%	**100.0%**
	n	28	16	38	17	18	18	8	143	72	71	143	5	14	17	22	31	16	7	9	8	14	143	**143**
	Column Comparisons	b D f g			d f				-			-											-	-
Diet Coke	%within column	2.3%	10.3%	13.0%	15.0%	7.7%	12.0%	22.7%	11.5%	8.3%	14.5%	11.5%	.0%	18.8%	3.3%	13.0%	7.6%	17.5%	9.1%	10.5%	7.1%	16.7%	11.5%	**11.5%**
	%within row	2.7%	10.8%	24.3%	24.3%	8.1%	16.2%	13.5%	100.0%	35.1%	64.9%	100.0%	.0%	16.2%	2.7%	18.9%	13.5%	18.9%	5.4%	5.4%	2.7%	16.2%	100.0%	**100.0%**
	n	1	4	9	9	3	6	5	37	13	24	37	0	6	1	7	5	7	2	2	1	6	37	**37**
	Column Comparisons				a			a	-			-											-	-
Coke Zero	%within column	9.3%	23.1%	18.8%	21.7%	28.2%	16.0%	13.6%	18.9%	15.9%	21.8%	18.9%	33.3%	18.8%	16.7%	18.5%	18.2%	25.0%	18.2%	21.1%	7.1%	16.7%	18.9%	**18.9%**
	%within row	6.6%	14.8%	21.3%	21.3%	18.0%	13.1%	4.9%	100.0%	41.0%	59.0%	100.0%	4.9%	9.8%	8.2%	16.4%	19.7%	16.4%	6.6%	6.6%	1.6%	9.8%	100.0%	**100.0%**
	n	4	9	13	13	11	8	3	61	25	36	61	3	6	5	10	12	10	4	4	1	6	61	**61**
	Column Comparisons					a			-			-												-	-
Pepsi	%within column	7.0%	5.1%	7.2%	21.7%	2.6%	6.0%	4.5%	8.7%	11.5%	6.1%	8.7%	11.1%	3.1%	10.0%	9.3%	9.1%	7.5%	4.5%	15.8%	7.1%	11.1%	8.7%	**8.7%**
	%within row	10.7%	7.1%	17.9%	46.4%	3.6%	10.7%	3.6%	100.0%	64.3%	35.7%	100.0%	3.6%	3.6%	10.7%	17.9%	21.4%	10.7%	3.6%	10.7%	3.6%	14.3%	100.0%	**100.0%**
	n	3	2	5	13	1	3	1	28	18	10	28	1	1	3	5	6	3	1	3	1	4	28	**28**
	Column Comparisons				a b c e f				-			-												-	-
Pepsi Light	%within column	.0%	2.6%	.0%	3.3%	2.6%	6.0%	9.1%	2.8%	1.3%	4.2%	2.8%	.0%	3.1%	3.3%	3.7%	3.0%	.0%	4.5%	.0%	.0%	5.6%	2.8%	**2.8%**
	%within row	.0%	11.1%	.0%	22.2%	11.1%	33.3%	22.2%	100.0%	22.2%	77.8%	100.0%	.0%	11.1%	11.1%	22.2%	22.2%	.0%	11.1%	.0%	.0%	22.2%	100.0%	**100.0%**
	n	0	1	0	2	1	3	2	9	2	7	9	0	1	1	2	2	0	1	0	0	2	9	**9**
	Column Comparisons						c	a c	-			-												-	-
Pepsi Max	%within column	16.3%	17.9%	5.8%	10.0%	12.8%	24.0%	13.6%	13.7%	17.2%	10.3%	13.7%	.0%	12.5%	10.0%	14.8%	15.2%	10.0%	31.8%	5.3%	21.4%	11.1%	13.7%	**13.7%**
	%within row	15.9%	15.9%	9.1%	13.6%	11.4%	27.3%	6.8%	100.0%	61.4%	38.6%	100.0%	.0%	9.1%	6.8%	18.2%	22.7%	9.1%	15.9%	2.3%	6.8%	9.1%	100.0%	**100.0%**
	n	7	7	4	6	5	12	3	44	27	17	44	0	4	3	8	10	4	7	1	3	4	44	**44**
	Column Comparisons		c				c d		-			-							c f h					-	-
NET	%within column	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**
	%within row	**13.4%**	**12.1%**	**21.4%**	**18.6%**	**12.1%**	**15.5%**	**6.8%**	**100.0%**	**48.8%**	**51.2%**	**100.0%**	**2.8%**	**9.9%**	**9.3%**	**16.8%**	**20.5%**	**12.4%**	**6.8%**	**5.9%**	**4.3%**	**11.2%**	**100.0%**	**100.0%**
	n	**43**	**39**	**69**	**60**	**39**	**50**	**22**	**322**	**157**	**165**	**322**	**9**	**32**	**30**	**54**	**66**	**40**	**22**	**19**	**14**	**36**	**322**	**322**
	Column Comparisons	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Column Names		A	B	C	D	E	F	G	H	A	B	C	A	B	C	D	E	F	G	H	I	J	K	**A**

We can greatly improve this crosstab by deleting things. The reason that this works is that the more things on the page the more our brain gets distracted and thus the more we delete, the clearer the story in the data will become. The key things that we can delete from this table are:

- All the statistics except for the column percentages. That is, we should delete counts (labelled **N**), row percentages (**% within row**). Of course, in some situations these additional statistics may prove to be useful, but more often-than-not they are not useful and thus the default should be to remove them.
- The decimal places. Due to sampling error it is very rare that differences in decimal places are relevant and thus they should not be shown.
- All of the income and gender data. The column comparisons show no significant differences for gender and thus the differences between male and female preferences for different brands of cola are likely due to sampling error and are thus not interesting. The case for deleting income is a little more complex. The only significant difference relates to the preference for Pepsi Max being relatively high among the people with incomes of $120,001 to $150,000/. This pattern seems counter intuitive: why should this small brand appeal to such a narrow income-based demographic. Consequently, it likely reflects one

of those flukes that comes up from time-to-time in surveys and thus, in the absence of any corroborating evidence, should be ignored.

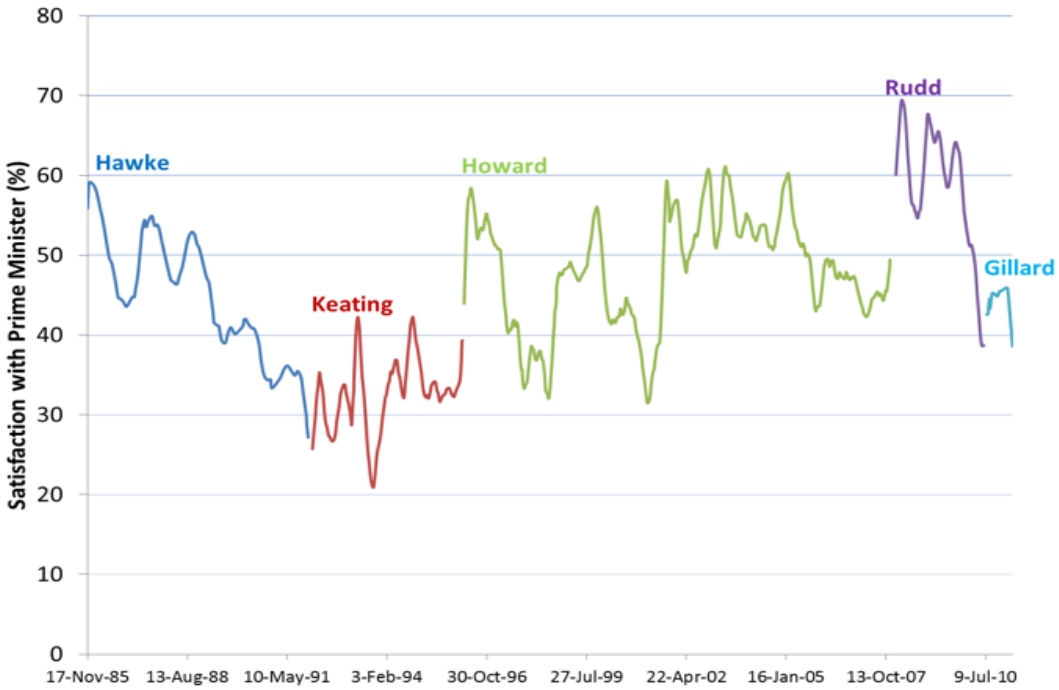
The much simpler crosstab that results is shown below.

		Age							
Preferred cola		18 to 24	25 to 29	30 to 39	40 to 49	50 to 54	55 to 64	65 or more	**NET**
Coca-Cola	% within column	65%	41%	55%	28%	46%	36%	36%	**44%**
	Column Comparisons	b D f g		d f					-
Diet Coke	% within column	2%	10%	13%	15%	8%	12%	23%	**11%**
	Column Comparisons				a			a	-
Coke Zero	% within column	9%	23%	19%	22%	28%	16%	14%	**19%**
	Column Comparisons					a			-
Pepsi	% within column	7%	5%	7%	22%	3%	6%	5%	**9%**
	Column Comparisons				a b c e f				-
Pepsi Light	% within column	0%	3%	0%	3%	3%	6%	9%	**3%**
	Column Comparisons						c	a c	-
Pepsi Max	% within column	16%	18%	6%	10%	13%	24%	14%	**14%**
	Column Comparisons		c				c d		-
NET	% within column	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Column Comparisons	-	-	-	-	-	-	-	-
Column Names		A	B	C	D	E	F	G	**H**

The same basic idea can be applied to charts as well. The example below is from a PowerPoint presentation showing the popularity over time of some of Australia's Prime Ministers. The plot beneath makes the pattern in the data much more clear by:

- Deleting the background colors.
- *Smoothing* the lines (i.e., deleting the small random deviations).
- Deleting the *chart junk* (i.e., the callouts with the Prime Minters' names.

"Are you satisfied or dissatisfied with the way … is doing her job as Prime Minister?"
Paul Keating becoming Prime Minister 20 Dec 1991
John Howard becoming Prime Minister 11 Mar 1996
Kevin Rudd becoming Prime Minister 3 Dec 2007
Julia Gillard becoming Prime Minister 24 Jun 2010
17-Nov-85
13-Aug-88
3-Feb-94
30-Oct-96
27-Jul-99
22-Apr-02
16-Jan-05
13-Oct-07
9-Jul-10
AMSRS NATIONAL CONFERENCE 2011


Satisfaction with Prime Minister (%)
80
70
60
50
40
30
20
10
0
Hawke
Keating
Howard
Rudd
Gillard
17-Nov-85
13-Aug-88
10-May-91
3-Feb-94
30-Oct-96
27-Jul-99
22-Apr-02
16-Jan-05
13-Oct-07
9-Jul-10

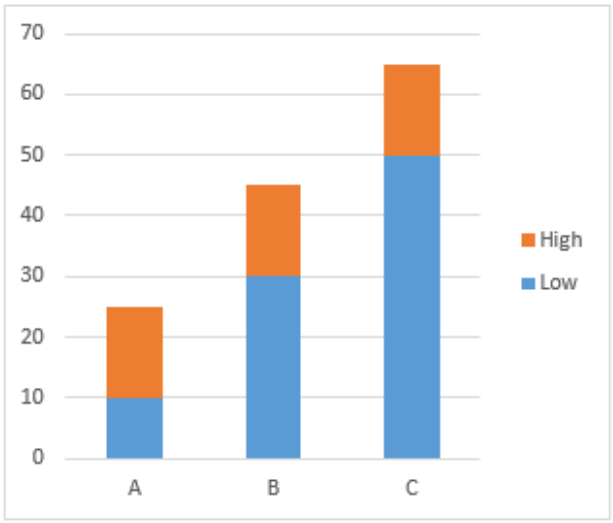
Ordering

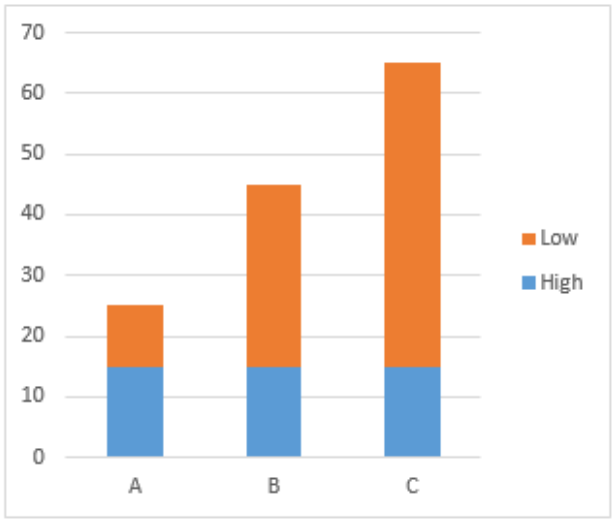
The two tables below show the same data. The one on the left is ordered alphabetically. The one on the right has been sorted by values. For most purposes, the table on the right is more useful as it allows people to quickly work out the relative performance of the options (which is usually the main interest when interpreting any table).

	%
AAPT	9%
New Tel	2%
One-tel	25%
Optus	89%
Orange	43%
Telstra	88%
Virgin Mobile	22%
Vodafone	75%
Other	6%
Don't know	0%
NET	100%

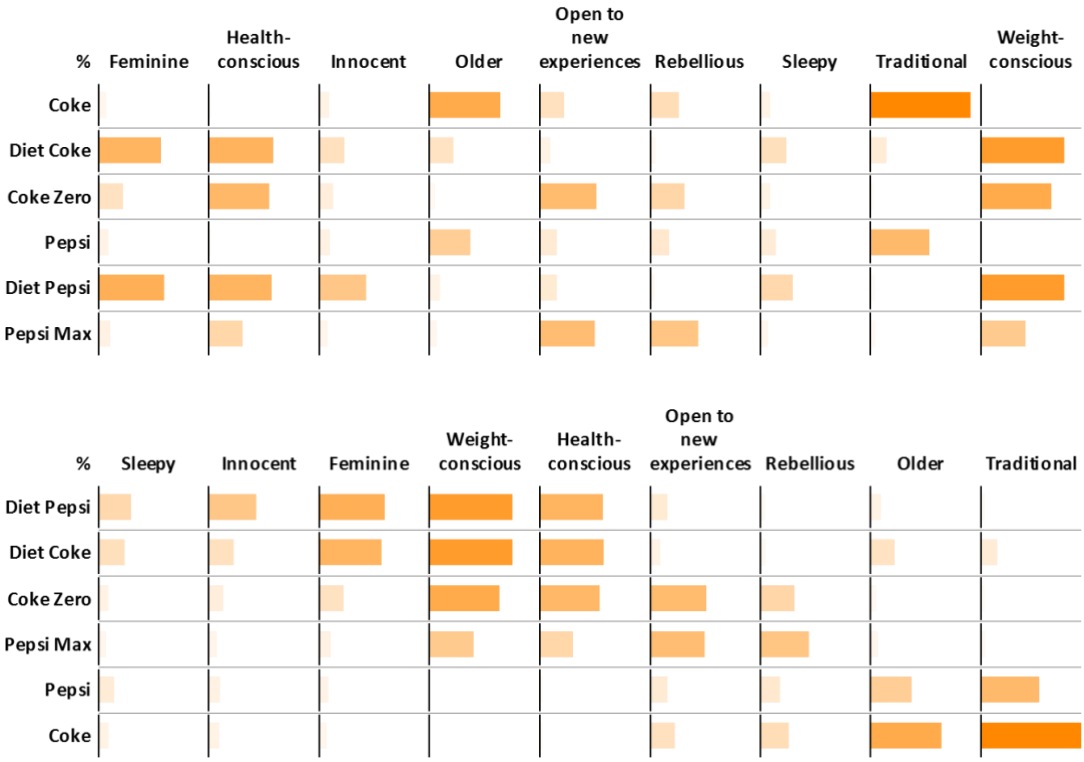
	%
Optus	89%
Telstra	88%
Vodafone	75%
Orange	43%
One-tel	25%
Virgin Mobile	22%
AAPT	9%
Other	6%
New Tel	2%
Don't know	0%
NET	100%

The idea of ordering information is more general than that of sorting. Consider the following two charts. The one on the left is substantially better in terms of communicating the pattern.[2]





When dealing with large tables the most useful way of ordering the tables is to use *diagonalization*. The basic idea of diagonalization is that the rows and columns of tables are re-ordered such that a near-diagonal pattern appears in the table. This allows our brains to recognize a pattern, which in turn makes it much easier to process the data. Note how in the second of the plots below it is really easy to see which brands are similar to which other brands and why.



Summarizing

Summarizing data involves replacing a larger set of data with a smaller set of data. Ordering: data is ordered in a way to make the story more obvious. Grouping: combing bits of data together. Summarizing

The second stage in trying to work out what a survey means is *data reduction*. As the name implies, this involves reducing the amount of data so that conclusions become easier to find.

References

1. Jump up↑ Collins, M. (1992). "The data reduction approach to survey analysis." Journal of the Market Research Society 34(2): 149-162. Ehrenberg, A. S. C. (1975). Data reduction: analyzing and interpreting statistical data. New York, John Wiley.
2. Jump up↑ Ehrenberg, A. S. C. (1975). Data reduction: analyzing and interpreting statistical data. New York, John Wiley.

Q Tutorial

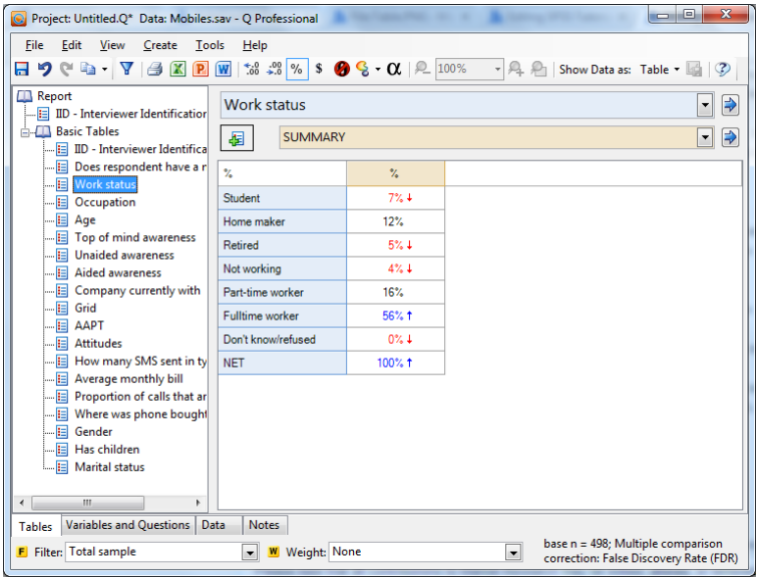
- Importing the data into the software and creating a summary report
- Summary tables of grid questions
- Creating a crosstab
- Filtering the data
- Creating NETs
- Rebasing tables
- Charts
- Changing metadata
- Computing minimums, medians and maximums
- Recoding numeric data
- Simple predictive model
- Creating market segments
- Reporting and sharing results

Q Tutorial

This tutorial provides a basic overview of some of the core functionality of *Q*.

Importing the data into the software and creating a summary report

1. Download the example SPSS data file and questionnaire (e.g., to your downloads folder or to your desktop).
2. Open Q (they provide one month free trials).
3. Click on **File > Import New Data File (New Project)** and press **Yes**.
4. Press **Yes** if asked any questions (these questions are asking if you wish to have the data file automatically tidied; this is almost always a good idea). You will now see a table showing the average value of the surveyanalysis::ID Variable.
5. Select **Create > Basic Tables**.
6. Select all the tables on the left and move them across to the **Key questions** box.
7. Press **OK**.
8. Click on Work Status (on the left). This is the first of the tables containing interesting table. Your screen should now look like the screen below.



Summary tables of grid questions

Grid questions are automatically identified when the data is read into Q. To see an example click on the table (shown on the left) called Grid (by default, Q names grid questions as grid, because the raw data files generally do not contain names for such questions).

%	Friendly	Low prices	Fashionable	Reliable	Good coverage	The best phones	Conveniently located stores	Leaders in mobile phone technology	NET
AAPT	19% ↑	19% ↑	13%	9%	5% ↓	11%	5% ↓	12%	49%
New Tel	11% +	10%	11%	6%	2% ↓	9%	3% ↓	9%	35% ↓
One-tel	14% +	18% ↑	12%	6% ↓	3% ↓	10%	3% ↓	9%	44%
Optus	44%	37% ↓	33% ↓	47% +	35%	44%	62% ↑	45%	90% +
Orange (Hutchison)	30% +	39% ↑	39% ↑	11% ↓	7% ↓	21%	14% ↓	21%	75%
Telstra (Mobile Net)	29% ↓	15% ↓	21% ↓	64% ↑	74% ↑	41%	66% ↑	61% ↑	94% ↑
Virgin Mobile	32% ↑	34% ↑	39% ↑	8% ↓	4% ↓	14%	9% ↓	14% ↓	67%
Vodafone	42%	34% ↓	49%	42%	42% +	41%	51% +	43%	89%
NET	95%	100% ↑	97%	99% ↑	99% ↑	84% ↓	96%	95%	100%

Creating a crosstab

In this example we will create a crosstab of Unaided awareness by age:

1. Press , which takes a copy of the current table. We will then modify this copy (this is the basic workflow of Q; tables are created as modifications of existing tables).
2. Click on the blue menu/box which currently says Grid and select Unaided awareness.
3. Type age into light brownish menu which currently displays SUMMARY and select Age. Note that interesting results are highlighted (these are the results of automatic statistics tests).
4. To show the number of observations in each cell of the table, right-click on the table and select **Statistics - cells > n**.

Filtering the data

Q has a number of different tools for creating the filters. To filter a table to only show the data of males in full-time work:

1. Press to copy the current table.
2. Follow the steps in #Creating a crosstab and create a crosstab of Gender by Work status.
3. Select the cell that corresponds to males in full-time work. (More generally, select the applicable cells.)
4. Right-click and select **Create filter'**.
5. Enter a name of Working Men and press **OK**. You have now created a filter, but we need to apply it.
6. Click back on the table showing unaided awareness by age.
7. Click on the **Filter** drop-down menu on the bottom left of the screen and select Working Men.

To modify or make a more complicated filter press the yellow **F** at the bottom left of the screen, right-click and select **Edit Variable**.

Creating NETs

To produce a of the proportion of people that said they Strongly agree or Agree a little with the statement Technology is fascinating:

1. Click on the Attitudes table (left of the screen).
2. Click on Strongly Agree and drag it on top of Agree a little.

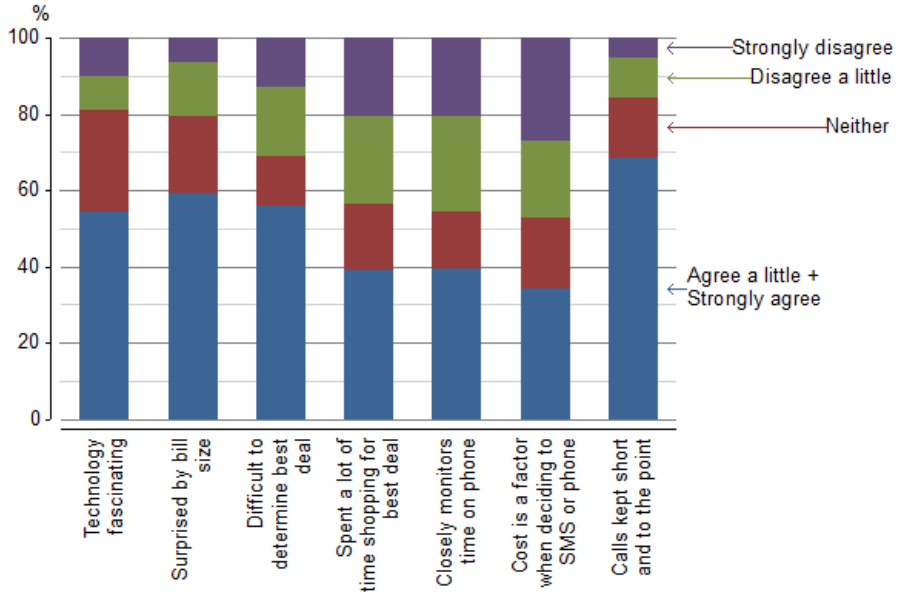
Rebasing tables

Re-basing involves recomputing the percentages on the table with some categories removed. In the above example the total sample size is 498 which includes two people that said DON'T KNOW. A more useful calculation of NET AGREE would remove these two respondents. This is done in Q as follows:

1. Right-click on DON'T KNOW.
2. Press **Remove'**.

Charts

A crosstab can be shown as a chart by clicking *"Show Data as'* (top-middle of the screen) and selecting the desired chart type.



Changing metadata

This example shows how to change data from appearing as categorical to instead appearing as numeric:

1. Click on the How many SMS sent in typical week table.
2. Press the blue arrow to the right of where it says How many SMS sent in typical week.

Data Manipulation > Structure > Average.

Computing minimums, medians and maximums

1. Select a table that is showing an average (e.g., see the previous section).
2. **Home > Statistics Cells > Median**.
3. **Home > Statistics Cells > Minimum**.
4. **Home > Statistics Cells > Maximum**.

Recoding numeric data

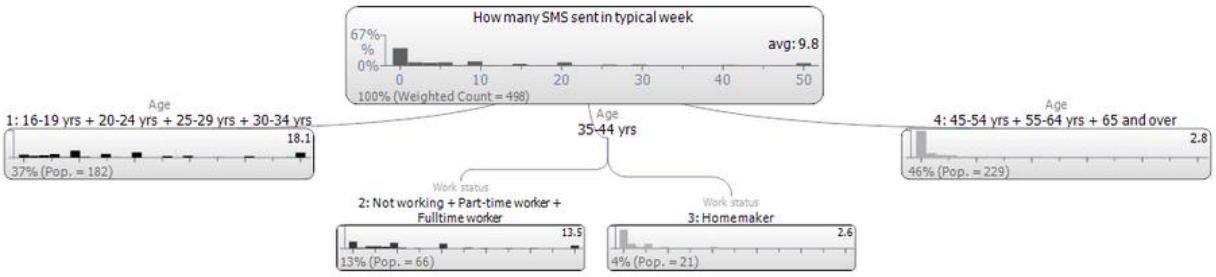
In the above example the range goes from 0 to 200 calls per week with a median of 3. This suggests that only a very small number of respondents have given high ratings and there may be advantage in recoding some of the higher values to lower values (so that they become less influential as *outliers*. We can do this by;

1. Select a table containing the data we wish to recode.
2. **Data Manipulation > Data Values > Recode Values**.
3. Replace all the values greater than 50 with the value of 50.
4. Press OK.

Simple predictive model

This example predicts the number of SMS (text messages) by demographics, assuming that you have recoded the data as numeric (see the previous section):

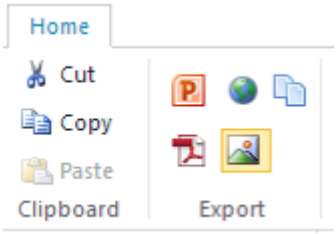
1. **Insert > Predictive Tree**
2. Select How many SMS sent in typical week.
3. Select:
   - Work status
   - Occupation
   - Age
4. Choose **Show more patterns**.
5. Press **Create Predictive Tree**. You should get a tree that looks like this:



Creating market segments

In this section of the tutorial we will create some market segments using Latent Class Analysis:

- Create a new page (**Insert > Pages > New Page > Table/Chart**
- **Insert > Analysis > Groups/Segments**

Reporting and sharing results

The report can be exported to PowerPoint, as PDFs, as images and as online reports by clicking the appropriate icons from the ribbon.