Levels of measurement in psychological research:

Psychology is a science. As such it generally involves objective measurement of the phenomena being studied, whatever these might be. However, not all measurements are the same. There are important distinctions between different kinds of measurements that you need to be aware of, because how you measure things affects what kinds of statistical test you can use on your data.

1. "Numbers" which are really names - the "nominal" scale of measurement:

Sometimes all you can do is place people into categories and record the frequency with which each category occurs. In this situation, you might use *numbers* as *names* for the categories. The examples I used in one of my lectures were the numbers on footballers' jerseys, and house numbers. These are not "real" numbers, and you cannot do any arithmetic with them other than count how many instances of each category occur.

This can sometimes be confusing, especially when using SPSS, which requires you to use numbers in order to code participants on various attributes. For example, in order to tell SPSS about the gender of your participants, you might use "1" to stand for "male" and "2" to stand for "female", like this:

subject name	gender
1	1
2	1
3	1
4	2
5	2
6	2

You can certainly *try* to perform various mathematical operations on these data, because SPSS will unintelligently go along with your demands. However the results will be quite meaningless. Here, the mean of the subject names is 3.5 and the mean gender is 1.5. Neither of these makes any sense, because it is impossible to combine names or genders in this way. If I go to "variable view" in SPSS, and use the "value label" option, I can make SPSS show names as words instead of code numbers. The absurdity of trying to do arithmetic on these values is now even more obvious: what's the average of three "males" and three "females", or the average of "Bob", "Bill", "Eric", etc.?

.

subject name	gender
Bob	male
Bill	male
Eric	male
Cynthia	female
Ethel	female
Doris	female

An example of a study using nominal data:

Here's a very crude study on the effects of noise on performance. We give people a single problem to solve. Half of them tackle it under very noisy conditions, and half tackle it in silence. Our measure is simply whether each person got the problem right or wrong. We could code these data using "1" for "right" and "2" for "wrong".

All you can do with nominal data like these is use the Chi-Square test to see if there are any significant differences in the frequencies with which the various categories occur. Here, we would merely count up how many people passed or failed each task, and then look to see if these frequencies differed from what we would expect to have obtained by chance.

Note that it makes no sense to calculate means and standard deviations for frequency data; a frequency is a frequency, pure and simple. As a result, any graphs would simply show the frequency with which each category occurred, with no error bars.

2. Measurements using proper numbers - the "ordinal", "interval" and "ratio" scales:

Deciding whether or not you have nominal (frequency) data is usually fairly straightforward. Think about the data provided by each participant: if all you know is that the participant falls into one of a number of categories, then you have data on a nominal level of measurement. If you have one or more scores from each participant, then it is clear that you do not have data on a nominal scale. However what you then need to do is to decide whether your data are measured on an ordinal, interval or ratio scale. This is sometimes tricky to decide. It comes down to two issues:

(a) are there **equal intervals** between the various points on your measuring scale?
(b) does the measuring scale have a **true zero point**, as opposed to an arbitrary one?

If data are measured on an **ordinal** scale, then (as the name implies!) they can be placed in some kind of *order*. Examples of ordinal scales might be: "small", "medium", "big"; "very tired", "quite tired", "awake", "very awake"; and "very happy", "happy", "neutral", "unhappy", "very unhappy". However, the points on an ordinal scale are not necessarily equally spaced. You can't do anything other than arrange the values in order of magnitude (amount of whatever it is you are trying to measure, such as size, alertness or mood, in the case of the scales just mentioned) is pretty much all that you can do with them.

The classic example of an ordinal scale is sporting performance. If you are told who comes first, second and third in a horse race, you know that the horse who came first was faster than the horse who came second, who in turn was faster than the horse who came third. Thus you can rank the horses in order of "speed". However, if this is all your data consist of, you don't know anything more about the horses' performance: it might be that the first horse beat the second one by a few seconds, and that the second horse beat the third one by a minute. Or it might be that the first horse beat the other two by minutes, and the second and third horses had very similar times. An ordinal scale of "first", "second" and "third" contains no information about the distances between these points on the scale.

In contrast to ordinal scales, if data are measured on an **interval** or **ratio** scale, the distances between the various points on the scale are equivalent across the whole range of measurements. The distinction between interval and ratio scales is rather subtle: a ratio scale has a true zero point, whereas the interval scale does not. If there is a zero value on an interval scale, it is merely an arbitrary point on the scale that is regarded as "zero" by definition.

The classic illustration of interval and ratio scales is temperature. Both the Centigrade and Fahrenheit temperature scales are interval scales. In both cases there are zero points, but these do not represent a true absence of temperature - they are merely arbitrary points on the scale. On both of these scales, it is quite possible to have temperatures below zero. In contrast, the Kelvin temperature scale is a ratio scale: zero degrees on this scale is defined as a complete absence of heat. Ratio scales have a true zero point, marking a total absence of the attribute being measured. The existence of this zero point means that you can make additional statements about the relationships between different points on a ratio scale.

To illustrate this, consider the example of temperature again. On all three scales, the points on the scale are equally spaced wherever you happen to be on the scale. Therefore on all three scales, it is possible to say that the temperature has increased by one degree, or decreased by two degrees, etc. A degree of temperature is a constant amount, and so a change from 21 to 22 degrees is the same amount of change in temperature as a change from 3 to 4 degrees. However, with the two interval scales (Fahrenheit and Centigrade), the absence of a true zero point makes it impossible to make ratio statements such as "it is twice as hot today as it was yesterday". You *can* make statements like this with a ratio scale, because on a ratio scale the zero point is a true zero (a total absence of the property being measured) and not just another point on the scale.

Centigrade	100	90	80	70	60	50	40	30	20	10	0	-10	-20	-30	-40	-50	-60 ..	-273
Farenheit	212	194	176	158	140	122	104	86	68	50	32	14	-4	-22	-40	-58	-76 ..	-460
Kelvin	373	363	353	343	333	323	313	303	293	283	273	263	253	243	233	223	213 ..	0

Examples of ratio scales in psychology are things such as reaction time, and individual scores such as "number of items correctly recalled" or "number of errors". With these kinds of measures, it is valid to make statements about ratios, such as "Fred was twice as fast as Dorothy", or "Fred made half as many mistakes as Cynthia". The statement "Fred got no items correct" is also valid, because there is a true zero on a "number correct" scale, representing a complete absence of correct responses.

Examples of interval scales include most IQ tests. There is no true zero point on an IQ test, so although I can say that "my IQ is 70 points higher than yours", I cannot say that "I have an IQ of 140 and you have an IQ of 70, so therefore I am twice as intelligent as you".

In practice, you don't need to worry too much about the difference between interval and ratio scales, because that won't affect your choice of statistical test. A simple way to choose between them is to think of whether a score of zero on your scale represents a complete absence of the thing being measured. if it does, you have a ratio scale; if not, you have an interval scale.

You *do* need to be able to appreciate the difference between these scales and an ordinal scale, because parametric statistical tests require data to be measurements on either an interval or ratio scale (i.e. they should not be used on ordinal data).

An example of an experiment producing ratio data:

Imagine we perform an experiment comparing two groups in terms of their memory for the details in a passage of text. Suppose that our measurement consists of recording the number of details from the passage correctly recalled by each participant. This is most definitely a *ratio* scale of measurement, with equal intervals and a true zero point (a participant could, in theory, remember no items at all from the passage; and it is meaningful to make statements like "participants in one group recalled twice as many items as participants in the other group").

What kinds of statistical tests can you perform on ordinal, interval and ratio data?

Statistical tests can be divided into two kinds: parametric tests (which make certain assumptions about the nature of the data on which you are performing the test) and nonparametric tests (which don't make those assumptions). (Don't worry, this topic will be much clearer once it's covered in depth in the "Research Skills 2" course).

The three assumptions that need to be met in order for you to perform a parametric test are:

(a) the data should be roughly normally distributed;

(b) the data should show homogeneity of variance (the spread of scores in the different conditions of the study should be roughly similar);

(c) the data should be measured on an interval or ratio scale.

Strictly speaking, parametric tests should only be used on data that satisfy these three assumptions. However an important part of the university experience is learning to be tolerant of ambiguity, and so you should be aware that this is a grey area in practice. Not all researchers and statisticians think that it's a problem to use parametric tests on ranked data, so you may well come across published research that uses parametric tests on ordinal data such as personality measures, attitude scale data, Likert scale scores, etc.

Type of data:	**Permissible descriptive statistics:**	**Permissible inferential statistics:**
Nominal	Counts (frequencies). Statements like "more people chose coffee than tea as their preferred drink".	Chi-Square
Ordinal	Median, mode (mean, though arguable). Statements like "people liked coffee more than tea". (But we don't know by how much).	Nonparametric tests (e.g. Wilcoxon, Mann-Whitney, Friedman, Kruskal-Wallis)
Interval	Median, mode, mean. Statements like "on the 'Beverage Appreciation Scale', people gave coffee a higher score than tea".	Parametric tests (e.g. t-tests, ANOVA)
Ratio	Median, mode, mean. Statements like "people drank twice as many cups of coffee than they did tea". (This is a ratio statement).	same as interval

The case of Likert Scales:

A popular measuring tool in psychology is the Likert Scale. This usually consists of a statement plus a rating scale that goes in apparently equal increments from an extreme negative to an extreme positive opinion. For example:

"Cats are evil little monsters".

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

A participant who gives a rating of "strongly agree" clearly feels more deeply that cats are evil than does someone who gives a rating of "disagree". Sometimes the verbal labels are replaced by numbers, such as 1-5, where 1 would be "strongly disagree" and 5 would be "strongly agree".

What kind of data are produced by Likert Scales? It's clearly not a ratio scale, as there is no true zero point, so at least we can exclude that option. However, are these ordinal or interval data?

Some people would argue that Likert Scales produce interval data, others that they are really ordinal data[1]. The central issue is whether or not the increments on the scale are truly equally spaced. At first sight they appear to be, especially if numbers are used to represent the different points on the scale, so that the scale runs from 1-5, 1-9, -3 to +3 or whatever. However, if you think about the psychological property that you are trying to measure with this scale, it's clear that in fact it is an ordinal scale.

If I give a rating of 5 and you give a rating of 5, we know that we both strongly agree with the statement, but we have no way of knowing for certain whether we really do have similar depths of antipathy to cats. We might both be using the same verbal label to represent different levels of feeling. Replacing the verbal label with a number from 1 to 5 makes these data look like an interval scale because the *numbers* are equally spaced - but we cannot know whether the *psychological property underlying the responses* is also equally spaced. How can we be certain that "amount of depth of feeling about cats" falls on such a linear scale?

We have no way at all of knowing whether the differences between the different points on the scale are truly equivalent, as they must be in order for it to be regarded as an interval scale. Is the difference in cat-hatred between me and someone who gives a rating of "agree" really the same as the difference in cat-hatred between that person and someone who gives a rating of "neither agree nor disagree"? We cannot do anything more than place people in order of magnitude of cat loathing, on the basis of their responses to this item. Therefore this is an ordinal scale.

What statistics can I do with Likert Scale data?

If you accept the argument above, then Likert Scale data are not suitable for parametric tests which require the data to be measured on an interval or ratio scale. However, as mentioned earlier, this is a grey area: in practice, researchers often do perform parametric tests on them.

[1] These include the people involved in this course!

What about descriptive statistics? Does it make sense to summarise these data with means and standard deviations? The answer is a qualified yes! Any means and standard deviations obtained from rating data (whether from the "Risky Shift" data or from a Likert Scale) are perfectly valid as *descriptions of participants' behaviour*, i.e. how participants responded when faced with a question and asked to pick a response. So it is fine to say something like "the mean rating chosen was 4.6, with a standard deviation of 1.2". This tells us that "typical" performance was to pick a rating of "agree" or similar, although there was some spread around this choice.

However, what this actually means in the context of the *underlying psychological construct* of "attitudes to cat morality" is a different question. Our data tell us that most people think cats are rather evil, but as we cannot know for certain that everyone who gave a particular rating really did feel exactly the same way about cats, we would have to be cautious in interpreting these data. For example, if we had three groups (cat lovers, people who were quite indifferent to cats, and cat haters) and they gave us different mean ratings (say "1", "2" and "5" respectively) we could say that the mean ratings for the groups differ, and that the three groups differ in their attitudes to cats. We would not be able to say much more than that.

In short, the use of Likert Scales raises an important point: in psychology, you need to distinguish between measuring people's behaviour, and interpreting what those measurements actually represent in psychological terms. This problem isn't unique to Likert Scales. For example, suppose you measure the reaction times of young people and old people and find a difference between them. The difference itself is real, but what gives rise to it may be much less clear-cut. It might stem from cognitive decline in the elderly participants, or the use of different strategies between the two groups, or some combination of the two. When you perform a study, always think carefully about what it is that you are *really* measuring.

Thanks to Linda Tip and Sarah Laurence for their contributions to the arguments in this handout.