Quantitative Methods
Module 4: Measurement

4.01 Measurement: Operationalization

Before we turn to the topic of measurement I'll briefly clarify the terms variable and operationalization. Earlier I referred to variables as operationalized constructs, but the term **variable** can also refer to a representation of a construct that is still somewhat abstract. We use the term **operationalization** if we want to explicitly indicate we're talking about a specific, concrete method to measure or manipulate a construct.

Say I'm interested in the construct 'political power'. I can represent this construct with the variable 'influence in parliament'. This representation is more specific, but we could still come up with very different procedures to operationalize or measure the variable 'influence in parliament'.

For example, I could count the number of bills that a Member of Parliament, or Congress, got passed, or I could count the number of years someone has been in parliament, or I could ask political analysts to rate members of parliament in terms of their influence.

So operationalization means selection or creation of a specific procedure to measure or manipulate the construct of interest. An operationalization makes it possible to assign people an actual score on the variable of interest.

Suppose I want to operationalize the construct 'love of animals'. I can observe a person interacting with a cat and count how often the person pets or strokes the cat. I could also decide on a different operationalization or **operational definition** by creating a questionnaire with statements like 'I usually avoid other people's pets' and 'I love animals'.

What about independent variables that are manipulated? Well, suppose I want to know if exposure to animals increases love of animals. I can operationalize the variable 'exposure to animals' by creating two levels of exposure. I could randomly assign people to take care of a cat from a shelter for a month or assign them to a control condition with not cat.

Another operationalization would be to take one half of a school class to a petting zoo and the other half to an abstract art museum. Or I could assign participants to watch an animal documentary or a train documentary. As you can see, the possibilities for operationalization are endless. Of course some operationalizations are better than others.

An operationalization doesn't necessarily capture or represent the construct in its entirety. As our constructs get 'fuzzier' or more complex, there's a greater chance that we measure or manipulate only a part of the construct. For example, if we measure love of animals with a self-report questionnaire, we measure feelings and attitudes, which might give a more positive image. If we measure love of animals by placing camera's in people's homes and observing behavior we might find lower scores.

We might find that, compared to their self-reported love of animals, people show a lot less love when their cat wakes them up at five AM or blocks the TV when they're watching they're favorite show.
It's important to keep in mind what aspect of the construct the operationalization actually measures or manipulates, especially once we use the data to draw conclusions about the entire construct in our hypothesis. Our conclusion may apply only to a limited aspect of the construct.

4.02 Measurement: Measurement structure

Once a construct has been operationalized we're ready to start measuring. Unfortunately, measurement is much less straight forward in the social sciences than in the natural sciences. Therefore it is extremely important to know to what we mean when we say we're *measuring* depression or political persuasiveness. We should know what we're capturing with the numbers that result from measurement, but more importantly, we should know what information we're *not* capturing.

So what is measurement? Ok, here we go: Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers. Take the property body length. I can determine qualitative relations on this property between three 'objects' - in this example: two persons and a cat- by standing them side-by-side and observing and comparing how far their heads stick out.

The first thing I can tell is that they're all of different length. I could represent this inequality relation by using different labels or numbers to represent the inequalities. Of course it would be weird to use the numbers 2, 10 and 14 like this, because there is another type of relation that we can immediately see, which is *not* reflected in the assigned numbers.
I'm talking about the order relation between A, B and C. Person B is the tallest, so he should receive the highest number, 14, and C is shortest and so should receive the lowest number, 2. We use the ordering of the numbers to represent the ordering of people - and cat- in terms of body length.

And we don't need to stop there. We can also determine if the difference in length between person A and person B is the same or larger than the difference between person A and C. We hold a piece of cardboard over A's head an cut the cardboard where it reaches the top of B's head. Then we hold the piece of

cardboard over C's head and compare to A. Suppose the cardboard reaches exactly to A's head, then the differences in length are the same.
We can represent this relation of 'equal differences' by using numbers that differ by the same amount. For example, the difference between the numbers for B and A, 14 and 10 is four, so we could change the number assigned to C from a 2 to a 6.
The equal differences in numbers between B and A, 14-10=4, and A and C 10-6=4, now accurately reflect the equal differences in body length between A and B and A and C. If the difference between A and C had been larger than the difference between A and B, then the difference in the corresponding numbers should have reflected this.

There's one more type of relation we can observe for the property body length. We can compare ratios of body length. We could take the piece of cardboard we cut out earlier, cut some extra pieces of exactly the same length and see how many cardboard units C, A and B are tall.

Ok, now suppose it takes two cardboard units to reach the top of C's head, four to reach the top of B's head and three to reach the top of A's head. This means B is twice as tall as C.
We could reflect this relation by changing the numbers again to 9, 12 and 6. We're still using different numbers for different lengths, the ordering of the numbers corresponds to the ordering of body lengths, the differences between A and B and A and C are the same (twelve minus nine is three and nine minus six is also three) and now the number for person B is twice as large as the number for C.

So you can see that measurement is the representation of empirical relations, between objects or persons on a certain property, by using the numerical relations between numbers. We can differentiate lengths, order them and compare differences and ratios of body length. We determined these empirical relations by looking and using some cardboard. Of course this method is pretty laborious if we want to assess body length for a great number of people.

Assigning numbers in a way that captures these empirical relations, for example by using a tape measure, makes our life a lot easier; especially if we want to compare or aggregate over many people. And of course assigning numbers to represent a property allows us to use statistics to help describe and draw conclusions about the property we are interested in.

4.03 Measurement: Measurement levels

We saw that measurement is the representation of relations between people on a certain property by using corresponding relations between numbers. Consider body length. For this property we can distinguish different body lengths, we can order them, compare differences in body length and even compare ratios. We can use the numerical relations of numbers to represent all these relations.

For body length all four possible relations - inequality, order, differences and ratios - can meaningfully interpreted. Unfortunately this is not the case for psychological and social properties. For most properties in the social sciences we can only determine some of these relations. The term **measurement level** is used to indicate what type of relation can be meaningfully interpreted.

If the only relation you can determine is that of inequality, distinguishing between values, then we call this a **nominal** variable. An instrument that can only differentiate between values is said to have a **nominal measurement level**. Examples are nationality, sex or pet preference.
A German has a different nationality than a Brit, women are of a different sex than men, dog people have a different preference for pets than cat people or hamster people. One value doesn't represent a greater degree of the property than any other value, they're just different. A German doesn't have more nationality than a Brit, women don't have more sex, well, let's say more gender than men and being a dog person doesn't mean you have a stronger animal preference than a cat or hamster person. There's no order to these properties.

Ordinal variables allow for differentiation *and* ordering of values. Suppose I want to measure math ability and use the number of correct answers on a math test with ten questions. The higher someone's math ability, the more answers they get right. We can order people's test scores to reflect their order in math ability, but differences or ratios of scores don't reflect differences or ratios of math ability.

We have no way of showing that the difference between a score of four and five is the same in terms of different math ability as the difference between a score of seven and eight. Sure, the difference in right answers is the same but how can we show this corresponds to an equal difference in math ability? We can't. And the same goes for ratios: someone with a score of ten doesn't have twice the mental math ability of someone with a score of five.

It actually remains to be seen if the test scores are measured at the ordinal level, that is, if they accurately reflect the order in math ability. What if someone with a score of one spent all their time on the hardest question and got it right, where someone else focused on the easier questions and got a score of three?

Only if the questions are equally difficult, can we use the test scores to accurately reflect the ordering of students on math ability. In that case the math test is said to measure at the **ordinal level**.

For **interval** variables it's possible to distinguish and order values, but also to interpret *differences* between values. Temperature is a good example. Suppose I'm heating up four pans filled with water on a stove and I measure temperature with a thermometer in degrees Fahrenheit. A pan of water reading 90 degrees Fahrenheit is hotter than one that reads 80; we can verify this by sticking our hand in. And the same goes for two pans reading 40 and 50 degrees.

We can also verify that when we heat up the 80 degree water to 90 degrees the expansion of a liquid, like the quicksilver in a thermometer, is the same as the expansion when we heat up water at 40 degrees to 50 degrees. So the difference between 80 and 90 and 40 and 50 is the same.

We can't say however, that water at 80 degrees Fahrenheit is twice as hot as water at 40 degrees. This is because the zero point for temperature is arbitrarily defined. The value 0 doesn't correspond to the absence of temperature, it corresponds to the temperature required to freeze brine, or salt water. The Celsius scale defines zero as the temperature at which fresh water freezes.
If we consider the same temperatures as before but now in degrees Celsius we see that thirty-two point two minus twenty-six point six is five point six, just like ten minus four point four is five point six. But twenty-six point six is nowhere near twice four point four. This is because the scales use different zero points.

Unlike interval variables, **ratio** variables have a non-arbitrary zero point that's the same for any scale you might choose. Of course length is an obvious example. The absence of length, 0 length, is the same whether you measure in inches or in centimeters. Variables measured at the interval or ratio level are very rare in the social sciences.

On final remark: The structure of a property doesn't have to be fully captured by a measurement instrument. Take age, a ratio property. I could measure age by asking respondents to indicate their age in years, thereby preserving the ratio level.
I could also ask them whether they are under twenty, twenty to thirty-nine, forty to fifty-nine or sixty or older, assigning the scores one, two, three and four. By creating age categories we no longer know exactly how old someone is. We can say that people in a higher category are older, but not by how much. By categorizing the variable we've lost the ratio and interval level information.

4.04 Measurement: Variable types

We identified the measurement levels nominal, ordinal, interval and ratio. In this video I'll look at how you can interpret variables with different measurement levels. I'll also discuss other ways to classify variables according to their measurement characteristics.

Categorical variables distinguish either unordered or ordered categories. A special type of categorical variable is a **binary**, or **dichotomous** variable. This type of variable has exactly two categories, such as male or female, smoker or non-smoker, furry or hairless. The categories can be natural, like the male/female dichotomy or created by the researcher such as under twenty years of age and twenty or older. Categorical variables with more than two categories are sometimes called **polytomous**.

For categorical variables differences between values are uninterpretable. Of course nominal and ordinal variables are both categorical. So how can

we interpret numerical results from categorical variables? Well, suppose I measure animal preference in a group of my friends by assigning the numbers one, two and three respectively to friends who prefer dogs, cats, or hamsters.
It doesn't make sense to look at the mean animal preference of, say, 1.2 and say that my friends have a low preference for animals. I could have assigned the numbers in the reversed order, resulting in a high mean of 2.8. For a nominal variable like animal preference it only makes sense to look at frequencies, how many people there are in each category.

What about a math test with ten questions that measures math ability at the ordinal level? Suppose I administer the test to my friends and find a mean score of 6.2. Is this mean interpretable? Well not if the scores only reflect ordering. Because I could reassign the person with the highest score the number 15. The ordering is still the same so the relations are preserved. Of course if I did this, the average test score be suddenly much higher.
The value is arbitrary and not informative of real differences between people on the property of interest. So if you have an ordinal variable you should stick to frequencies and statistics like the mode and the median.

Categorical variables can be contrasted with **quantitative** variables. Quantitative variables allow us to determine not only that people differ on a certain property, but also to what extent they differ. Interval and ratio variables are quantitative.

For quantitative variables like temperature, weight and length it does make sense to calculate a mean and, for example, to compare means of groups. This is because the mean is influenced by the distance between numbers. For quantitative variables the distance between numbers actually corresponds to the distance between people on the property of interest.

For example, if I measure the weight of friends who own a cat and the weight of my friends who own a dog, I can compare their mean weight to see if cat people are heavier because they don't get the extra exercise from walking their pet.

A final distinction that you should be able to make is between discrete and continuous variables. For continuous variables it's always possible, in theory anyway, to find a value between any other two values. Consider body weight: If one person weighs 65 kilograms and another 66 kilograms, we can easily image finding someone who weighs 65.5 or 65.72 or 65.268334. As long as our measurement instrument is precise enough, any value between the minimum and maximum on the scale should be possible.

Discrete variables on the other hand, can only take on a limited set of values. Nominal and ordinal variables are by their nature discrete. But quantitative variables can also be discrete. Take the number of pets someone has owned. This is a ratio variable, because differences can be compared: the difference between two and four pets is the same as between one and three pets, and because ratio's can be compared: someone with four pets owns twice as many pets as someone who with two pets.

The set of possible values is very limited however. We start at 0 and then the values, 1, 2, 3, 4. But 1.3 pets or 4.7 pets are not valid values. So here we have an example of a discrete ratio variable.

The distinction between continuous and discrete variables is less relevant, because it's not associated with how you can interpret the measurement results unlike the distinction between categorical and quantitative variables.

4.05 Measurement: Measurement validity

Until now I've discussed operationalization and measurement levels without asking "Are we measuring accurately, do our measurements reflect the construct we are interested in?" in other words: is our instrument valid? The validity of an instrument or manipulation method is commonly referred to as **measurement** or **construct validity**.

How do we assess construct validity? Well, suppose I've created a questionnaire that aims to measure fondness of cats. A higher score indicates someone is more of a cat person. How do we determine if this score actually reflects the property 'fondness of cats'?
Well, we could determine its **face validity**. Face validity refers to how well the instrument represents the property according to the assessment of experts. An expert opinion can be useful in the development phase, but of course experts can always be wrong. A slightly better alternative is to show the instrument has **predictive validity** or **criterion validity**, by demonstrating the instrument can predict a relevant property or criterion. Of course the ability to predict something doesn't mean the scores used for prediction accurately reflect the intended construct.

Suppose I create an instrument to measure motivation that can predict job satisfaction. That doesn't mean the instrument measures motivation, it could reflect another construct entirely, say general positive attitude. So **criterion validity** can support the claim that we are measuring *something* consistently, but it has limited value in demonstrating that this is indeed the intended construct.
What would be ideal is if we already had a valid instrument for the property of interest. We could then administer both instruments and see whether the scores on the new scale agreed with the already validated scale. Unfortunately there aren't many gold standard instruments for social and psychological constructs.

Another solution would be if we could directly check our measurements. Consider body length: We can use a tape measure and then check if the person whose head sticks out furthest gets the highest measurement result. This purely qualitative way to assess a property is cumbersome, but it allows us to directly check the validity of a tape measure or a bathroom scale.
For social and psychological constructs the situation is very different. We don't have an undisputed, direct way to determine whether one person is more

intelligent or fonder of cats than another. So is there another way to assess construct validity? Well we can go about it indirectly, by seeing whether the scores relate to similar and different variables in a way that we expect. We refer to this as **convergent** and **discriminant validity**.

For example, I would expect scores on my cat fondness scale to show agreement, or converge, with scores on an observational measure of cat fondness, where people spend ten minutes in a room with a cat and we count the number of times the person looks at or pets the cat.
I would expect less agreement between my cat fondness questionnaire and a questionnaire on fondness of wild animals. It wouldn't be strange to find some association, but I would expect it to be smaller. Finally, I would expect no association with a variable that is supposedly unrelated to cat fondness, like fondness of pizza.

A systematic method to assess convergent and discriminant validity is called a **multi-trait multi method matrix** approach. In this approach we use different instruments, for example different questionnaires or observation and self-report instruments to measure two traits.

Let's take cat fondness and pizza fondness as an example. We would expect a very high association between cat fondness, measured observationally and measured through self-report. And the same goes for pizza fondness; we would expect the observational and self-report instruments of pizza fondness to show strong convergence.

We would expect a very small to zero association between cat fondness and pizza fondness both measured using self-report. A small association is possible because some people tend to give socially desirable or generally positive answers to questionnaires. The same zero to very small association can also be expected between cat and pizza fondness measured by observation. Finally, we would expect no association between different constructs measured with different methods.

If the relations show all the expected patterns then we have indirectly supported the construct validity of these four instruments. Of course this is a laborious process, because a lack of convergent or discriminant validity could be due to any one of the instruments. This would require a new study that combines the instruments with others in new ways to find out where the problem exactly lies. Hopefully you can appreciate how challenging it is to assess the construct validity of social and psychological constructs.

4.06 Measurement: Measurement reliability

A measurement instrument should be valid, but also reliable. **Measurement reliability** refers to the instrument's **consistency** or **stability** or **precision**. A reliable instrument will result in highly similar scores if we repeatedly measure a stable property in the same person.
My bathroom scale, for example, isn't *perfectly* reliable. If I step on it three times in a row it will usually show two or three different readings. But as long as the readings differ by one or two hundred grams, the scale's reliability is good enough for me.

So how do we determine the reliability of instruments that measure social and psychological constructs? Well in some cases it's as easy as administering the instrument twice to a group of participants and determining how strongly the results from the first and second measurement agree. This is called **test-retest reliability**. We can use this method if we're measuring things like weight or reaction times, but as soon as a person's memory of their answer is involved, things become more complicated.

Suppose I have a questionnaire measuring fondness of cats. It consists of five questions. If I ask a group of my friends to fill out this questionnaire once and then again fifteen minutes later, they will probably still remember the answers they gave the first time.
Now if I find a high consistency in the scores on the first and second measurement, is this because the instrument is reliable or because my friends have pretty good memories and would like to seem consistent in their attitudes?

One way to solve this problem is to look at the consistency, not between different times, but between different parts of the instrument at one time. This is referred to as **internal consistency**. We compare responses on the first three and the last two questions. Of course you can only do this if the instrument consists of several questions that are supposed to be comparable and measure the same construct.
If this is the case then you can determine the **split-halves reliability** by randomly splitting the test in half and assessing the association between the first and second half. There are also statistics that are equivalent to the average of all possible ways to split of the test.

If measurement consists of observation instead of self-report, you can have the observer rate the same behavior twice and assess the association between the two moments. This is referred to as **intra-observer consistency** or reliability.

Of course the memory of the observer can inflate the association. Since it shouldn't matter who makes the observations you could also assess the reliability of observation by having two different people observe and rate the behavior and look at the association between the two raters' scores. We call this **inter-observer consistency** or **inter-rater reliability**.

Ok so we've seen different ways to establish how reliable, or precise an instrument is. But what is it that makes an instrument less reliable? If the

instrument perfectly reflects someone's **'true score'** or true value on the property of interest, then the measurement result, or **'observed score**' should be the same every time.

But what if we systematically measure an additional construct? Take my cat fondness scale, what if these questions also tap into the construct 'general positive attitude'? This could result in a systematically higher score for people with a positive attitude. We call this **systematic error**. This means our instrument is less valid, but not less reliable.

As long as the observed score is determined only by the **'true score'** on cat fondness and the **'systematic error'** caused by the second construct, positive attitude, then we would still get the same observed score every time we measure the same person.

Reliability is influenced by **random error**, error that's entirely due to chance. If the observed score is in part determined by random fluctuations, then we get different values each time we measure the same person.
If a scale is entirely unreliable, if there is no association between observed scores at different measurement moments, then we're basically measuring random error or noise. Put another way, this means that at least some reliability is required before an instrument can be valid. The reverse does not hold. A perfectly reliable instrument can be entirely invalid. This happens when it perfectly measures a different construct than it was supposed to measure.

Let's consider the possibilities in more detail. Of course the worst-case scenario is when an instrument has low reliability and low validity: a lot of random and systematic error. Even if the true score contributes a little to the observed score, it will be almost impossible to distinguish this contribution.

An instrument can also have low reliability and high validity: a lot of random error but very little systematic error. We are measuring the right property, just very imprecisely. An instrument can also have high reliability and low validity: a small amount of random error but a lot of systematic error. We're measuring the wrong property very precisely.

Best-case scenario is high reliability and of course high validity: a small amount of random error and systematic error. The observed score is mainly determined by the true score. We are measuring the right construct with great precision.

Of course the trick is to separate all these error components from the true score, even if there is a fair amount of systematic and random error. Psychometricians and sociometricians aim to do this by using statistical modeling to partial out the random and systematic error.

4.07 Measurement: Survey, questionnaire, test

Surveys, **questionnaires** and **tests** are very popular measurement instruments in the social sciences. **'Survey'** is a general term that can refer to a list of questions asking about biographical information, opinions, attitudes, traits, behavior, basically anything. Surveys generally cover a variety of topics.

The term **'questionnaire'** is used when the focus is on one construct, or a related set of constructs, usually psychological traits, emotional states or attitudes. The term **'test'** is used when the aim is to measure an **ability**, such as general intelligence or math proficiency.

Surveys, questionnaires and tests should always include a clear instruction. For example, for a math test, it's important to know whether you should choose the right answer or the best answer and to how many decimals numerical answers should be rounded. The instruction can also provide a cover story. This will prevent participants from trying to guess the purpose of the study, possibly distorting their responses.

Surveys can be administered by an interviewer who visits respondents, goes through the questions and records the answers. This is very expensive though, so a more common method is to use self-report. This means people read, and respond to the questions by themselves. The survey can be completed using paper-and-pencil but of course the use of online applications is becoming much more common. Online administration is easier for the respondent: no need to visit the post office, you can complete the survey in your own time and help buttons and pop-up comments can provide extra instruction if necessary.

Online administration offers *researchers* added control: control over the order in which questions are viewed, checks to ensure all required answers are actually filled in and identification of strange response patterns - like the same answer to every question. A disadvantage of online administration is the low response rate. People easily delete an email with a link to an online questionnaire. It's much harder to turn away an interviewer at your door!

Surveys, test and questionnaires all consist of a series of questions. We refer to the questions as **items**, because sometimes they consist of statements or even single words that respondents can agree or disagree with. The question, statement or word that a participant has to respond to is called the **stem**. The stem is usually accompanied by a set of discrete **response options** or a continuous range to choose from.

A psychological attitude, trait or state is almost always measured with items that describe feelings, thoughts or behavior that represent the relevant property. Usually, several items are used to measure the same construct. By using more than one item, random errors will generally 'cancel out'. Using several items also allows us to assess reliability by checking the internal consistency of the items.

Suppose I want to measure fondness of cats with the following five items:

1. Petting a cat is enjoyable.
2. I hate it when a cat jumps on my lap.
3. When near a cat I'm afraid to get scratched.
4. I frequently look at cat videos on the Internet.
5. Twenty years from now I can see myself having a cat.

People can choose from three answer options: Disagree, neutral or agree, scored 1, 2 and 3.

Items that are supposed to measure the same construct, or the same aspect of a construct, form a **scale**. When the items that form a scale are added together we get a sum score that indicates a person's value on the property.

Of course in our example agreement with some items indicates high cat fondness while agreement with others indicates low cat fondness. The items that are negatively worded, items 2 and 3, need to be **recoded**. Disagreement should be coded as 3 and agreement as 1! After recoding, adding the item scores results in a possible scale score between five and fifteen. A higher sum score means someone is more of a cat person.

Questionnaires frequently measure different aspects or dimensions of a psychological property by using subscales. Different sets of items tap into different aspects or maybe even different but related constructs altogether.

For example, if I'm interested in measuring your academic motivation I could distinguish between intrinsic motivation, extrinsic motivation and fear of failure. There are statistical methods that assess whether these dimensions are in fact distinguishable based on the pattern of responses provided by the respondents of course.

4.08 Measurement: Scales and response options

The most commonly used type of scale is a **Likert scale** or **summative** scale. Such a scale consists of comparable statements that are supposed to measure the same property. Respondents can indicate to what extent they agree with or endorse each statement.

Likert items should be monotone, meaning that respondents are consistently more likely to agree with the item if they possess the property to a greater degree. This is necessary because the scores on the items will be added.[1] Let's take my cat fondness

[1] Items are:
1. Petting a cat is enjoyable.
2. I hate it when a cat jumps on my lap.
3. When near a cat I'm afraid to get scratched.
4. I frequently look at cat videos on the Internet.
5. Twenty years from now I can see myself having a cat.

questionnaire as an example. Items one, four and five will show stronger agreement from people who are fond of cats; items two and three will show stronger disagreement. After items two and three are recoded, all items are monotone: A higher score indicates higher cat fondness.

An example of a *non*-monotone item is "Cats are a necessary evil to get rid of mice". People who love cats will disagree, but extreme cat haters will also disagree, they will feel cats are just evil and not necessary at all. A high score could indicate high or low cat fondness.

There are other types of scales, such as differential scales that allow for non-monotone items and cumulative scales where the items themselves should show consistent ordering, each item expressing the property more strongly than the previous one. These scales are not used very often though.

Likert items generally have three to seven discrete response options, indicating strength of agreement. But a **visual analog** or **graphic rating scale** can also be used. This is a graphic representation, usually a simple line segment with two extremes at each end, like "disagree" and "agree". A respondent simply marks the line to indicate their position.
Some items will be "harder" to endorse than others because they represent more extreme cat fondness. Psychometricians and sociometricians use statistical techniques such as **item response theory** to assess an item's "difficulty".

OK, so how are Likert scale questionnaires constructed? Well the construction of a questionnaire starts with a clear **description** of the property to be measured. Often different **dimensions** can be distinguished, which are measured with a subset of items that form a **subscale**.

Identifying dimensions requires an in-depth analysis of the construct. It helps to consider different types of situations or areas in which the property can be expressed. For example: academic motivation can be intrinsic - you love the subject matter - but it can also be extrinsic - you expect to get a better job if you graduate.

It also helps to consider the different ways in which a construct can be expressed. For example, academic motivation can be expressed in attitudes and behavior; aggression can be expressed verbally or physically.
Once the relevant dimensions are identified, items are generated for each unique combination of dimensions. Ideally, each item describes a specific situation and a specific expression of the property. Respondents can easily misinterpret vague or too general items.

In general, all items, not just Likert items, should be **well formulated.** This means item formulation should be **short** and **simple**. Things like **double negation**, **unfamiliar words** and overly **complicated formulations** should be avoided.
The item "I do not take an unfavorable stance toward physical interaction involving fur-

stroking with creatures of the feline genus" will probably confuse many respondents.

Formulations should also be **unambiguous**. Take the item "Cats are funny creatures": Does this mean that cats make you laugh, or that you think they're strange? **Double-barrelled** questions are another source of ambiguity. Take the question "I love to pet and pick up cats". How should people respond who like to pet cats but don't like picking them up?

Also, items should be neutral, **not suggestive**. An item like "Don't you agree that most cats are friendly?" could influence impressionable people to give a more favorable answer than they would have given otherwise.

Items should be **answerable** for all respondents. For example, people who have never been in direct contact with a cat cannot answer items 1, 2 and 3 of my questionnaire. Perhaps I should have included a **filter** question, asking if people have ever physically interacted with a cat.

Extreme wording should also be avoided. Using words like 'never' or 'always' make items very hard to moderately agree or disagree with. Take the item "Cats are purely evil creatures". Even moderate cat haters will probably disagree with this statement.

Of course **response options** also need to be unambiguous and consistent. Response options need to be **exhaustive** and **mutually exclusive**; exhaustive means all respondents should be able to reflect their position on the property. If you ask people their age and only provide the options twenty to thirty, thirty to forty and forty to fifty, then teenagers and people over fifty cannot respond.

These categories are also not mutually exclusive; people who are thirty or forty will have a hard time choosing between categories. Changing the options to zero to thirty, thirty-one to forty and forty-one and over fixes both problems.

Of course there is much more to item, scale and questionnaire construction than the points I've discussed here. There are many more scales types, item types and response options, and different ways to implement these, for example by combining self-report with ratings by others. Also, there are many more aspects to consider, such as item order and how to deal with sensitive questions. We've only just scratched the surface here.

4.09 Measurement: response and rater bias

When people respond to items, our hope is that their observed scores are determined mostly by their 'true' score on the measured property: The score we would obtain if we had a perfectly valid and reliable instrument. Of course no instrument is perfectly valid or reliable. There's always some degree of random and systematic error. A special type of systematic error occurs when respondents show a systematic bias in their responses to items. These biases are referred to as **response sets** or **response styles**.

I'll first discuss the most common response styles or biases that occur in self-report measurement. These are acquiescence, social desirability, extreme response styles and bias towards the middle.

Acquiescence refers to the tendency to agree with all statements, regardless of their content. An easy way to spot this bias is to include some negatively phrased items. Consider my cat fondness questionnaire. Items two and three are negatively worded. Someone who agrees with statements one, four and five, but also with items two and three, isn't answering consistently.

Social desirability causes a biased pattern that is a little harder to detect. A social desirability bias affects the responses of people who tend to present themselves more favorably or in more socially acceptable ways. A social desirability bias can occur if a scale measures a property that is considered socially sensitive, or measures a property that is relevant to someone's self-image.

It's possible to detect a social desirability bias by adding 'social desirability' items such as: "I've never stolen anything in my life" or "I've never lied to anyone". The idea is that every one has stolen something or lied at least once in their lives, if only as a child stealing from or lying to their peers. If people strongly agree with these items there's a fair chance that their responses to other questions are biased towards responses that are more socially accepted.

An **extreme response style** is much harder to detect. This bias can occur for example when respondents don't want to think about exactly *how* strongly they agree or disagree with an item, they'll just choose the most extreme option. So unlike the acquiescence bias, participants' responses are consistent, just more extreme than their true value.

Bias towards the middle is highly similar to an extreme response style; only the tendency here is to choose a *less* extreme response option. For example, some respondents might never use the most extreme response options because they want to seem nuanced.

The ultimate version of bias to the middle occurs when there is an uneven number of response options, and a respondent always chooses the middle option. This response

pattern can be detected by including some extra extremely strong items, such as "cats are purely evil creatures".
Cat lovers will strongly disagree, but even people who like cats just a little should show some disagreement with this statement. If they respond with the middle category to all items, including these extremely worded items their response pattern is inconsistent.

Biases due to the mere act of responding to a test or questionnaire can also occur when a rater observes and rates behavior of others. There are many rater biases. I will just discuss the halo effect and generosity and severity errors here.
The **halo-effect** occurs when a positive or negative rating on one dimension spills over to other dimensions of behavior that are rated or evaluated. A well-known example is that more attractive people are generally rated as more intelligent or better at their job.
A **generosity error** or **leniency effect** occurs when the rater is overly positive or kind in their ratings. The opposite, a systematic bias towards very strict or negative rating is referred to as a **severity error.**

It can be hard to detect, let alone avoid halo-effects and generosity or severity errors. One approach is to use several raters that are trained to use clearly defined coding schemes. Checking the inter-rater reliability and average ratings for each rater can help to detect any systematic bias between raters, but of course it is very hard to detect bias that is shared by all raters!

4.10 Measurement: Other measurement types

I've focused on measurement using **questionnaires**, **surveys** and **tests**, but of course there are many other ways to measure social and psychological constructs. In biology, medicine and psychology **physical** measures are very common. Think of things like electrical skin conductance to measure arousal, eye tracking to measure focus of attention, EEG and fMRI to register brain activity and reaction times to assess cognitive ability.

Another way to measure is through observation, a method frequently used in sociology, psychology and educational sciences. **Observational measurement** might seem simple, but there is more to it than just observing and recording all the behavior that you see. Systematic observation involves careful registration of specific behavior.
Researchers employ coding schemes that specify categories of behavior and their criteria. They specify what the behavior in each category looks like, how long it should be displayed and under what circumstances it should occur.

A researcher also needs to decide on the time frame to be coded. Will we view an entire hour of videotaped behavior or will we sample five two-minute intervals? If we have taped material of an hour for each of sixty participants, then the two-minute intervals might be a better idea!

Other important issues are training and calibration of observers. Coding schemes can be complex and target behavior can be difficult to spot. So it's a good idea to have more than one observer and to train observers until they show enough agreement when coding the same material.
Agreement between different observers, called the inter-rater reliability, should be high of course. If reliability is low than at least one of the observers codes the behavior differently from the rest. Or, even worse, the behavior cannot be interpreted consistently.

Let's move on to a related form of measurement. **Trace measurement** assesses behavior indirectly through physical trace evidence. An example is counting the number of used tissues after a therapy session to represent how depressed a client is. Sometimes a property can be represented with measurements that were already collected by others. We refer to this as **archival data**. An example of archival research is to use census data, collected by a national research institute, on income and voting behavior. We could investigate whether areas with a higher average income are associated with more votes for conservative political parties. Trace measurement and especially archival data are frequently used in political sciences and sociology.

Content analysis is a technique that shares characteristics with observational, archival and trace measurement. Like observational measurement, content analysis consists of structured coding, but of elements in a text. The text can consist of newspaper articles, blogs, narratives or transcription of interviews.
Content analysis can be used, for example, to see if conservative and liberal politicians argue differently, by identifying the number of emotional and rational words they use in newspaper interviews. Of course this is a simple example. Text can be coded automatically according to very complex schemes using computer software.

A final measurement method I want to discuss is interviewing. In a **structured interview,** the questions, the question order and response options are pre-determined. This type of interview - be it face-to-face, through telephone or Skype - is not much different from using a survey. The response rate of interviews is higher, but it can be more difficult to get unbiased answers to sensitive questions.

Unstructured or **open interviews** are very different. An open interview is considered a qualitative method. Now although the focus here is on quantitative methods, I quickly describe open interviews because they are used very often. In an open interview the interviewer starts off with a general topic and usually has a set of points to be addressed but the interview is not limited to these points.

Questions are open-ended and there is little structure, so the conversation can lead anywhere depending on the respondent's answers. The questions that will come up and the range of answers are undetermined. Of course this makes it much harder to compare and aggregate data from different respondents. I won't go into other qualitative methods here, but you should know there are other methods available such as case studies, focus groups, oral histories, participatory observation and many, many more.