Measuring Attitudes to and/or Prejudices against Groups

36-313, Fall 2021

9 November 2021, Lecture 20

Contents

1 Measurement: Some Generalities

The last few lessons have been implicitly about **measurement**. The goal of tests like the SAT, and even more specific subject-matter tests is to measure knowledge of facts and/or skills. There are several lines of evidence for thinking they achieve that goal, at least to some extent:

- "Face validity": The tasks the tests ask you to do are, in fact, ones that demonstrate that knowledge and/or skill
- Process: It's hard to see how you could score high on such tests if you *don't* have that sort of knowledge, and it's easy to understand how having that knowledge contributes to your getting a high score
  - Even so, test scores are going to also be influenced by many other things, like motivation, fatigue/stress, test-taking-skills (as opposed to the skills we're really trying to measure), ...
- Predictive validity: scores on these tests are correlated with other things which (we think) also demand those sorts of knowledge and/or skill, like SAT predicting college grades

Despite their great similarity (and common historical roots), things like IQ testing are much less convincing as efforts to measuring "general intelligence". There is no *theory* about what should or should not be on such a test, the way we have criteria for what should or should not be on an arithmetic or geometry test. (And this is despite a lot of time and effort on the part of intelligence-testers who have been sensitive about this absence.) Attempts to explain what "general intelligence" might be, and how it might contribute to getting a high score on a test, usually turn out to be vague generalities or viciously circular (you're able to solve a lot of problems because you have a lot of problem-solving ability). There are, also, many other explanations for patterns in test-results which do not involve there being any such thing as "general intelligence", which would also explain why IQ scores correlate with many life outcomes.

This suggests some lessons about measurement:

- What you're trying to measure should, in fact, exist
- That variable should be *a* cause of the measured values
  - The *other* causes of the measured values should, ideally, be so much random noise
- Saying "this is how we measure X" doesn't make it so

Today we're looking at measuring attitudes towards groups, and/or prejudice against (or even for) groups. This is a *very* complicated and tricky thing, once you think it through. For any one person, we're talking about their idea or conception of some social group, which is an abstraction in their mind, *and* then the emotional tone or coloration or dispositions which go along with that. (If somebody thinks group X is unfairly discriminated against and oppressed, but that a consequence of that discrimination is that members of group X are aggressive and lots of them are driven to lives of crime and are dangerous to be around, is that a sympathetic or unsympathetic attitude towards the group?) How are attitudes about a group, in the abstract, connected to attitudes towards any particular person who is (someone thinks) a member of the group? How are either kind of attitude connected to behavior? And now we want to do *statistics* on this, so we want to gather measurements from lots of people and hope they are, somehow, comparable.

There are lots of difficulties for every way people have tried to do this. That doesn't mean it's not worth attempting, but it's important to be clear about those difficulties, and the limitations of what's been attempted.

2 Explicit Attitude Measures

The simplest way you could find out what somebody thinks and feels about a social group is to ask that person. If you approach them the right way, they will often give you an answer! Since we're rarely interested in what one particular person's attitudes, this is usually done as part of some survey with a sampling scheme.

If you ask someone what they think and feel, they will often tell you — *in words*. This is a problem for statisticians. Dealing with free-form text as data is difficult; dealing with free-form text from hundreds or thousands of people as data is very difficult. (If some people say auto mechanics are "cheats", and other people say they are "swindlers", is that expressing the same attitude, or importantly different ones?) Attitude surveys therefore try to force people to respond in set, stereotyped ways, which make the data analysis easier. Some prominent ones are binary questions, ordinal scales, and numerical scales.

2.1 Binary questions

are simple: "Do you feel favorable or unfavorable towards X?" The difficulty is that lots of people might have more complicated feels.

2.2 Ordinal or Likert Scales

These are questions where the allowed answers are of the form "strongly disagree, disagree, neutral, agree, strongly agree, no opinion", or the "strongly disapprove, disapprove, neutral, approve, strongly approve, no opinion". These are called **Likert scales** not (as I thought as an undergrad) because they're about how much you like something, but after Likert (1932), which first systematically used them in attitude measurements. (Those examples are of the common five-point scale; you can imagine how the three-point and seven-point scales go.)

It's important to realize that ordinal data are in fact ordinal, so it makes sense to *rank* them, and we can, e.g., talk about a median value, but most arithmetic operations aren't dubious, and so means don't make a lot of sense. It's even not altogether clear that my "agree" and your "agree" on the same question have the same meaning. Maybe my intensity of feeling when I say "agree" would actually correspond to what you mean when you say "strongly agree". But the hope is that most people in the population in question mean roughly the same things by these familiar words and phrases.

Also, there are usually many Likert-type questions, and then we often want to reduce them to some sort of over-all estimate of the attitude. At this point people will sometimes start to do unwise things like assigning numerical scores 1–5 to the levels and summing them or averaging them across questions. A statistically more sophisticated procedure would be to say that each person i has an attitude α_i, and the probability of giving response k on question j is then some function $f(\alpha_i, j, k)$. If we think that different questions are better or worse at "tapping in to" the underlying attitude, so each question has a β_j, we'd then have response probabilities of the form $f(\alpha_i, \beta_j, k)$, and we could try to jointly estimate the α_i's and the β_j's. This will usually involve some assumptions about specific algebraic forms for the function f. If this sounds rather like the item response theory we talked about for achievement tests, that's no coincidence. Note that if our model *assumes* there's a single, one-dimensional variable α which drives the responses, simply estimating that model won't check those assumptions. (Estimation is not goodness-of-fit.)

(A typical model here would be what's called the "graded response model" (GRM), which would say that the probability of giving response k *or higher* is $\frac{e^{a_j(\alpha_i - \beta_{jk})}}{1+e^{a_j(\alpha_i - \beta_{jk})}}$. The β_{jk} parameter is basically a the threshold for responding k or higher on question j, and the a_j parameter says how sensitive question j is to the trait being measured over-all. From these cumulative probabilities, we can find the probability of any one response by subtraction. This gives us the likelihood, and the log of that is what we maximize to estimate all the parameters.)

2.3 Quantitative scales

There are also attempts to try to get more directly quantitative measures of attitudes out of people, like "feeling thermometers" where you're supposed to say how "warmly" you feel about a group, on a familiar temperature scale. The obvious problem here is that we don't really know if my answering "70" really means the same thing as your answering "70". (This would be less of a problem if we want to do "within-subject" comparisons, of how relatively warmly I feel about different groups.) My suspicion is that things like this aren't actually any more quantitative than a Likert scale, but I will admit to not being a specialist on this.

2.4 Drawbacks to Explicit Attitude Measures

There are a number of obvious difficulties with surveys which ask people what they think and feel. Some of these are common to all surveys, others specifically related to these issues.

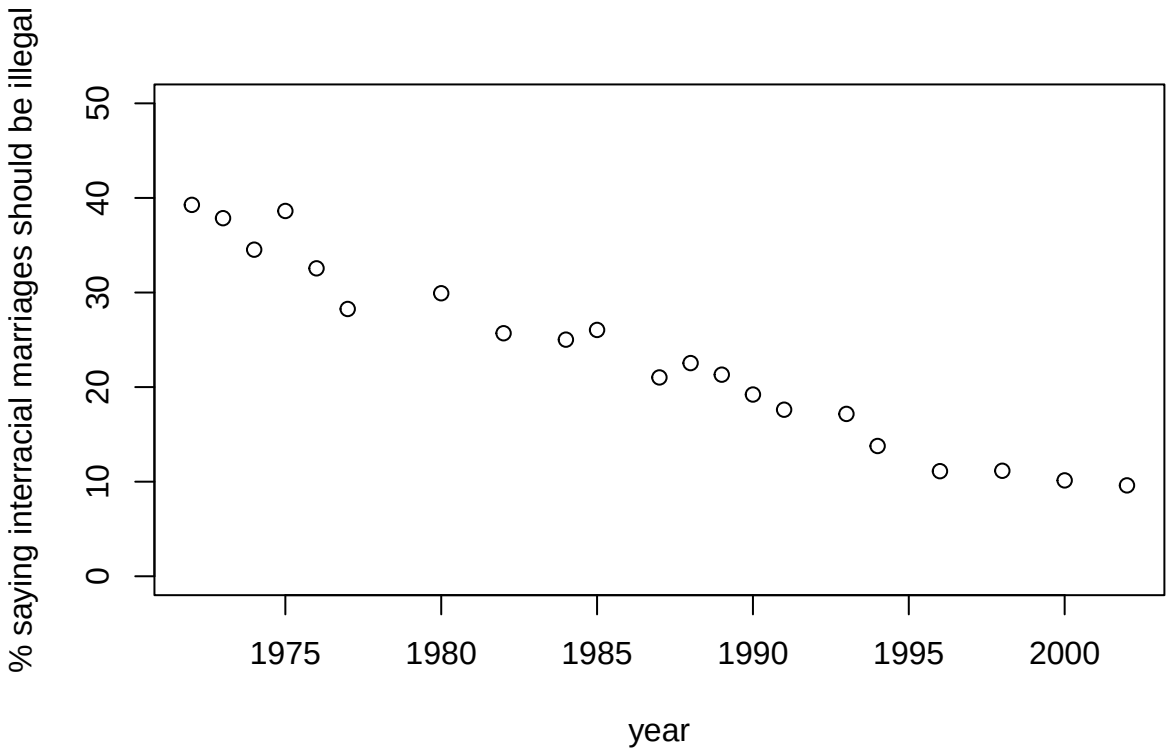
People who answer surveys may be systematically different from those who don't. In particular, certain attitudes may be more or less common among survey-responders than in the population at large. So while the results are accurate about those who answer, they don't generalize to the population of interest. This is one of the various forms of **selection bias**.

Different ways of phrasing a question can elicit different answers. Whether this is because those different phrasings carry subtly but importantly different meanings, or because people are suggestible and manipulated by "framing", is a difficult question. More practical responses to the problem are to include multiple versions of the "same" question, perhaps flipping randomly whether it's asked positively or negatively.

- Telling a survey-worker what you think and feel about some group is a social interaction with another human being, and people can be *very* deliberate about how they appear to others. In particular, they are prone to lying to make themselves look good in the eyes of others. The technical phrase for this is **desirability bias**. This is a problem here, because if I think that such-and-such an attitude is widely disliked, I'm not likely to admit to holding that attitude. (I might be more or less willing to admit it to a stranger like a survey-worker than to someone I know.)

(The trouble desirability bias creates for us gets worse because desirable attitudes are different in different times and places. Attitudes that are seen as desirable among young online-magazine writers in Brooklyn will not be the same as those endorsed among middle aged fundamentalist Mormon farmers in Idaho. This makes comparisons over time, space, and social groups especially difficult.)

As a concrete example, there are long-running surveys which have asked questions about approval of interracial marriage over many decades. In the General Social Survey (GSS), the text of the question reads "Do you think there should be laws against marriages between (Negroes/Blacks/African-Americans) and whites?" (The text has changed slightly over the decades to reflect trends in group names. Also, the question wasn't asked in every year of the survey.)



What you can see from the figure is that the percentage of people *saying* that marriages between blacks and whites should be illegal declined dramatically between 1972 and 2002. Such marriages were, in fact, illegal in many states, until all those laws were over-ruled by a 1967 Supreme Court case, wonderfully named *Loving versus Virginia*. (Notice that this was just five years before the GSS started asking this question.) The percentage of those saying such marriages should be illegal dropped something like ten points in less than ten years, so *either* a lot of people changed their attitudes towards this, *or* a lot of people changed what they were willing to say. Or, of course, some of both. Now, in this particular case, we have other reasons to think attitudes really have changed, because interracial marriages have become much more common (and the children of such marriages do no worse in life than others and are not despised as mongrels, etc.), etc. (Alba 2020). But we *also* have reasons to think that *some* of the change visible in that graph is increasing desirability bias.

3 Not quite so explicit attitude measures

One way to try to get around desirability bias is to ask questions which aren't so *directly* about attitudes towards the group in question, so that people will feel more comfortable giving honest answers.

A very prominent instance of this, since the 1980s, has been the use of "modern racism" or "racial resentment" tests or scales. The idea here is that, by the 1980s, it was not acceptable in public polite company to express openly racist attitudes of the kind that had been common in the 1950s and 1960s (to say nothing of earlier)[1]. Remember that the civil rights acts were in 1964 and 1965, that interracial marriage only became legal everywhere in the US in 1967, etc. — there were plenty of Americans in 1980 or 1985 who had been quite openly racist fifteen or twenty years earlier, but who now felt like they couldn't be *openly* racist. "Modern racism" or "racial resentment" scales are a series of Likert questions where someone could give answers which don't *necessarily* commit them to views like "the reason black people are poor is that they're lazy and dumb", but *hint* at it. A typical item on the scale asks people to agree or disagree with the following:

> Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.

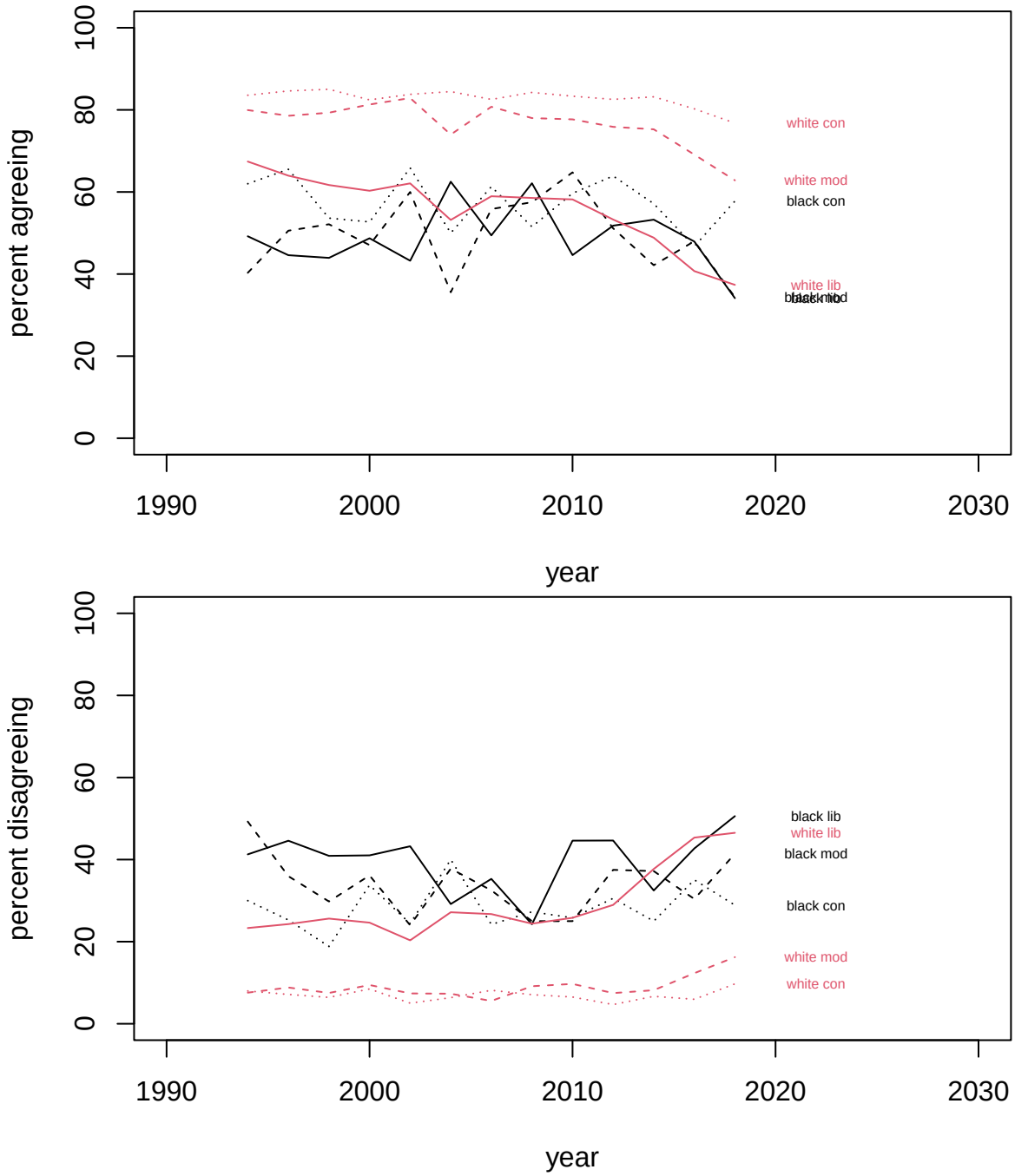
and so on through other, similar items. (For some items, like this one, agreement is the more racist direction, but for others, e.g., "Over the past few years, blacks have gotten less than they deserve", disagreement.) These scales or tests are *intended* to get at attitudes towards blacks, and specifically at *racist* attitudes towards them. These scales certainly predict certain kinds of behavior (Cramer 2020),[2] but exactly what that means is where things get tricky.

If you were a white racist in 1985, and got asked to agree or disagree with that statement, it's easy to imagine you'd endorse it. With legal discrimination ended only 20 years before, it's also plausible to imagine that *most* people who endorsed it would be white racists, though that's a shakier (and see the figure below). (Also: it's not exactly certain we've gotten away from desirability bias! You have to expect that at least *some* people see what the test-makers were driving at with these questions, and adjust their responses accordingly.) But it is not clear that this is *still* what the question meant in 2016, or means in 2022, or will mean in 2030.

It's also not clear that this statement has the same meaning to all groups. As it happens, the "work their way up" question has been asked in the GSS since 1994, so we can look at trends in it:

[1] In the homework, you'll see examples of the kinds of questions used to measure racism in the 1940s and 1950s, which were *much* more blunt.

[2] A very subtle point can arise here. A typical finding is that people who score higher on a "modern racism" scale are also less likely to vote for black candidates. This is often interpreted as high scorers being prejudiced against blacks. But it's also logically possible that *low* scorers are prejudiced *in favor* of blacks. It seems to me very hard to distinguish between these alternatives, since it's not like we have some known-to-be-complete-unprejudiced people whose support for a given black candidate could be used as a reference level. Indeed it seems so hard to distinguish between these alternatives that I doubt whether the question is meaningful at all. But see Agadjanian et al. (2021).



Here I've sub-divided respondents by race, and by their political views as liberals, moderates and conservatives[3]. I've also collapsed "strongly agree" and "agree" into a single level of "agree", and likewise with "disagree". (You can play with my code to break these out if you like; it doesn't change things much but does make the graphs much "busier".) You can see that whites are more likely to agree with this work-their-way-up statement than are blacks who declare the same political views, which fits with the idea that this statement is tapping into racism of some kind. But something like 40% of black *liberals* agree with the statement. White liberals are now more likely to *dis*-agree with this statement than the average of all black people (though not black liberals)[4]. It is *imaginable* that this is all because of deeply internalized anti-black racist views on

[3]The GSS has a 7-point Likert scale for political views, from very liberal to very conservative, so I've grouped 1–3 as "liberal" and 5–7 as "conservative". I'm also ignoring people who say it doesn't fit them, don't answer, etc.

[4]Figures like this inevitably suggest that the same group of people is changing their minds about the question being asked.

the part of many black people, but it's also imaginable that some black people, at least, give this statement a different meaning than it was originally intended to have by the test-makers. Maybe they take it as an expression of black pride, maybe they understand "special favors" differently, maybe they even take it as a statement of defiance. But if black people can understand this statement differently, maybe other people can too!

There is currently controversy about whether, *today*, the "modern racism" scale really measures *racism*, or what's called the "just world assumption", that people by and large end up with what they deserve. One line of argument in favor of this is that if you take the same statements and swap out other groups for "blacks", e.g., "Nepalese", you get very similar results. You even get similar results if you swap in "whites" or "Americans" (Carney and Enos 2017). If this is right, then these tests might not be measuring racism, at least not any more, but a different attitude towards inequality. People acting on such attitudes might help perpetuate or create racial inequalities, but it would seem a stretch to say that the *attitude* was *racist*. (I do want to emphasize that this is very much a live point of debate among experts right now, so it's unwise for non-experts to hold any very strong opinion here.)

This *may* be the case, but there can again be selection issues: the set of people who self-identify as "black conservatives" or "white moderates" changes over time. It *could* be that lots of white liberals have come to disagree with the statement over the last ten years, but it's also possible that white people who agree with it have ceased to call themselves "liberals". Panel or longitudinal data, tracking the same people over time, is by far the best way to settle such doubts.

4 Implicit Attitudes

One big problem with asking people about their attitudes, and even with asking them questions that *hint at* their attitudes, is that they can see what you're getting at, and at least some of them will adjust their answers to serve their goals, not to tell you the truth. Another big problem is that, people may not *know* their own attitudes, and/or that their conscious attitudes may not actually be what shape their behavior[5].

This has led people to try to find *indirect* ways of measuring attitudes. Ideally:

- The attitude is a cause of the performance on the test
- We don't directly ask about the attitude
- It's hard for people to control how they respond, and so hard to fake.

The outstanding example of this is the **implicit association test** (IAT) (Greenwald, McGhee, and Schwartz 1998; Greenwald, Nosek, and Banaji 2003). The basic idea is that very strongly learned associations between two concepts can be activated very quickly and automatically, without conscious thought, leading to fast reactions. But if people have to over-ride those very strongly learned associations, that *does* take conscious thought, which is slow (and error-prone). So if someone (strongly) associates idea X with idea Y, and we ask them to do something where they need to link X's and Y's. But if we ask people to do something where they need to link X with Z, that should be slow, *especially* if Y and Z are somehow opposed. Reversing this, if we find that linking X with Y is faster than linking X with Z, *maybe* we can conclude that X and Y were already more strongly associated in someone's mind than X and Z.

This leads to the scoring procedure introduced by Greenwald, Nosek, and Banaji (2003) and used about a bazillion times since then:

- In phase 1, you have to press one key (say `a` on the left) if the computer shows you a picture from group A *or* a positive word, and a different key (say `l` on the right) if you see a picture from group B *or* a negative word
- Phase 2, you press the first key if you see a picture from group B *or* a positive word, and the other key if you see a picture from group A *or* a negative word.

The difference between your average response time in phase 1 and in phase 2 is supposed to measure how strongly you associate group A with positive things and group B with negative things, versus associating group A with positive things and group B with positive things.[6] That is, to be clear, to be clear, faster reactions in phase 1 than in phase 2 are supposed to measure the extent to which you have the "A good, B bad" association. Association strengths are supposed to cause to reaction times, so we can "read back" from reaction times to associations. Attitudes are supposed to cause associations, so there's an extra step of inference there. The A/B pairs where this has been used include: black vs. white, male vs. female, Japanese vs. Korean, insects vs. flowers, and many others.

This has been a *hugely* influential procedure, not just in psychology but "in the wild" of daily life. Unfortunately there are a lot of problems.

We can start with saying that it's not at all clear what someone's score on the IAT has *measured*. What's been *calculated* is the difference in average response times between phase 1 (A/good, B/bad) and phase 2 (A/bad, B/good). That stronger associations imply faster responses is plausible, but it rests on some psychological ideas which can certainly be disputed (they're not really detailed enough to call a "theory"). We certainly don't have a theory which lets us quantitatively connect association strength with reaction times, or even an understanding of what other confounding factors would show up in reaction times.

[5]For example, I had an acquaintance in my 20s who insisted that what he was looking for in a girlfriend was spiritual companionship from a fellow Catholic. But it was a bit of a joke in our circle that what his girlfriends all *actually* had in common were very similar figures and red hair. I don't think he was lying, or even deceiving himself, but there was clearly something going on *other than* his expressed, conscious attitudes.

[6]I haven't seen psychologists address what should happen for someone who likes A but has no particular feelings about B, or who doesn't like B but really, really hates A. It'd seem like at most the test could measure *relative* attitudes towards the two groups. But it's a huge literature and I could easily have missed this corner of it. (If any reader can send me a pointer, I'd appreciate it.)

But even if this procedure measures associations, it doesn't tell us where the associations come from. Someone's personal attitudes *might* lead to associations between positive and negative *words* and the objects of those attitudes. *Maybe* people who are racists for white people and against black people *therefore* have an association between "caress" and "Hank", and an association between "crash" and "Latisha". (These are actual examples from Greenwald, McGhee, and Schwartz (1998).) How those associations would be built up is not exactly clear. (They certainly wouldn't actually encounter those names paired with those words many times in their experience.) It would seem to have to be a very indirect association. But (as many people have pointed out) another place such associations could come from is simply someone's *knowledge of* cultural stereotypes, even if their own attitudes are very different[7]. It's also quite possible that the linkages here are so indirect, and so many other things can affect reaction times, that there's no straight-forward interpretation *at all* of the difference in reaction times.

Leaving that to one side, the **reliability** of the IAT is very low. Measurement theorists say a measurement is "reliable" if it gives the same, or very similar, values on repeated measurement. A scale which gives wildly different readings every time you step on it is not a reliable measure of weight. A scale which gives the *same* measure if you step on it, step off, and then step back on is "reliable" in this sense, even if (say) it's always off by 25%, so long as the error is always in the same direction. Reliability, in this sense, is basically the opposite of measurement noise, *not* necessarily accuracy[8].

A common measure of reliability is the "test-retest correlation", the correlation coefficient for re-doing the test after some time has passed. For the IAT, typical values of this, after a few weeks, are about 0.4 (Machery 2021, sec. 4). (For something like the SAT or an IQ test, the test-retest correlation at that time-interval would be more like 0.8 or 0.9, and even something as dubious as "narcissism" would clock in around 0.7.)

If a measure has low reliability, it's a bad idea to base any important judgments or decisions on a single measurement. A low-reliability measure of individuals *might* still give useful information about groups averages. So (for example) even if the IAT is an unreliable measure of how sexist (or racist, etc.) any individual is, aggregating lots of unreliable measurements might still tell us whether (say) doctors or lawyers are more sexist. Some of the original proponents of the IAT now take more or less this line, though not always consistently. Similarly, you could imagine averaging many IATs of the same person taken over time, in the hope that they'll all fluctuate around their true level of bias. (I don't know if anyone has advocated this seriously.) Reporting results to, or on, any one individual about how biased they are on the bias of a single test this unreliable seems *scientifically* irresponsible.

But reliability may also be beside the point, because IAT scores do not, in fact, do a great job of predicting behavior, and changes in IAT scores do not seem to lead to changes in behavior (Machery 2021, sec. 5 and 6):

> [T]here is no sugarcoating it; At the individual level, indirect measures are poorly predictive of behavior, and their incremental validity [over and above explicit measures], while not null, is very limited. Predictive validity could be higher in some contexts, but compelling evidence is lacking. The limitation of the significance of indirect measures to a narrow context undermines their social significance and is definitely at odds with the ambitions of their inventors.

[7] There are, for instance, negative stereotypes of white people which are common in American culture, most prominently that they're boring and uptight: their food is bland, they're bland, they bad at sports and dance and anything else that uses their bodies, they're sexually repressed, they try to make everyone else as boring as they are, etc. Basically anyone who grew up in the US, exposed to popular culture, has seen many instances of these stereotypes. It would be very interesting to know if a version of the IAT can detect *these* associations. (Again, for all I know this has been done somewhere in the huge IAT literature, and I'd appreciate any pointer from readers.)

[8] See much more on this below.

5 Morals

1. Measurement is hard; measuring slippery, complicated things is *very* hard.
2. It is easy to be mislead by the names people give their procedures into thinking that measurement has been achieved. If something is *called* a "racial resentment test", then it's easy to *presume* that it measures racial resentment. But whether it *actually* does so is a complicated and debatable scientific hypothesis[9]. Measurement is an *achievement*, not a *presumption*.
3. The fact that all the ways of measuring attitudes I've covered have big problems doesn't mean we should give up; but it does mean that we can't say this is a solved problem and build on its results.

[9]It might be better if we gave tests like this random identifying strings, so that the question of whether the BGJHD test measures racism, the belief that everyone gets what they deserve, or something else, might be less heated, and less pre-judged. Cf. McDermott (1976) on the dangers of "wishful mnemonics" in artificial intelligence research.

6 "Reliability", "Validity", "Precision", "Accuracy"

Scientists and statisticians have spent a lot of time thinking about measurement, and evolved some fairly sophisticated theories about it. Some of this theorizing has been done in the physical sciences, and some in the social sciences, most especially psychology. We won't go fully into the details, but both traditions have developed a distinction between measurements being "reliable" or "precise" on the one hand, and being "valid" or "accurate" on the other. (Psychologists tend to talk about "reliability" and "validity", physicists about "precision" and "accuracy".) Reliability or precision is about whether repeating the same measurement, under the same condition, will give the same result. Validity and accuracy are about whether we're measuring what we claim to be measuring, and coming close to the truth.

To illustrate a little, here are 5 measurements of my body temperature (in Farenheit), made with the same thermometer at one-minute intervals as I was working on these notes:

97.2, 97.1, 97.0, 97.3, 97.5

Body temperature fluctuates, but probably not by half a degree in two minutes. It's reasonable to think my actual temperature wasn't changing over this period. The difference in these measurements is then just noise. (Maybe I held the thermometer differently each time and that alters the results, maybe the electronic circuits in the thermometer keep drifting into and out of alignment, etc.) When we talk about the **precision** of a measurement, we're talking about the extent to which we'd get the same values out of repeating the measurement. Usually we report some notion of *im*-precision, like the standard deviation of the measurements (0.192 degrees F) or their variance (0.037 degrees F squared).

When we talk about the **accuracy** of a measurement, we're talking about how close it gets to the truth. This is often quantified by some notion of *in*-accuracy, like the mean squared error (MSE) or root-mean-squared (RMS) error. As you remember, the MSE is equal to the variance plus the square of the bias. More generally, physicists tend to distinguish between **systematic** sources of error — ones which show up repeatedly, the way bias shows up as a difference between the average measurement and the truth — and **statistical** sources of error, which follow some distribution but don't repeat. We can't get at *systematic* error sources, like bias, *just* from looking at the distribution of measurements. We *can* get at the magnitude of merely-statistical error by looking at that distribution. If my body temperature is *really* 98.6, then the fact that the average thermometer reading was 97.2 F suggests a bias of 1.38 degrees Farenheit. (This in turn would imply an MSE of 1.93 degrees Farenheit squared, and an RMSE of 1.39 degrees Farenheit.) But if my actual temperature is really 97.4 degrees, there would be a different bias and a different RMSE. Checking instruments and procedures against cases with *known* values — of temperature, of mass, of radioactivity, etc. — is an important part of doing high-quality measurement in the physical sciences for just this reason.

So: physicists like to gauge precision using the variance or standard deviation of measurements, and to gauge accuracy using the MSE or RMS error. Since, as I just reminded you, MSE is variance plus bias squared, precision limits accuracy, more exactly variance is a *lower bound* on MSE. (And so the standard deviation is a lower bound on RMS error.)

When psychologists talk about the "reliability" of a measurement, they mean basically the same thing as what physicists mean by "precision". However, psychologists like to gauge reliability by the *correlation* between measurements that should be identical. This is *related* to the variance of measurements, but it's not the same. Say that the true variable (which we're trying to measure) is F, but our measurement is $X = m(F) + \epsilon$, where m is some function[10] and ϵ is mean-zero noise[11]. Then the physicists' notion of precision is $\mathrm{Var}[\epsilon]$, or perhaps $\mathrm{Var}[\epsilon|F = f]$. If we have two separate measurements, $X = m(F) + \epsilon$ and $X' = m(F) + \epsilon'$, the

[10] The function m is intended to capture systematic errors and distortions. In an ideal case, $m(f) = f$, the identity function. (This is the case of what psychologists call "classical test theory".) Otherwise, there is a bias, since $\mathbb{E}[X|F = f] - f = m(f) - f \neq 0$. In many real-world measurement situations, for instance, it's very common for $m(f)$ to become increasingly nonlinear, usually increasingly flat, as f moves out of some range.

[11] If we *define* $m(f) \equiv \mathbb{E}[X|F = f]$, and further *define* $\epsilon \equiv X - m(F)$, then one can show (Shalizi, n.d., ch. 1) that $\mathbb{E}[\epsilon|F = f] = 0$ for all f, and so both that $\mathbb{E}[epsilon] = 0$ and that $\mathrm{Cov}[\epsilon, F] = 0$. It's not, however, usually the case that ϵ is *statistically independent* of F.

covariance between them will be

$$\begin{aligned}
\mathrm{Cov}\left[X, X'\right] &= \mathrm{Cov}\left[m(F) + \epsilon, m(F) + \epsilon'\right] & (1)\\
&= \mathrm{Var}\left[m(F)\right] + \mathrm{Cov}\left[\epsilon, m(F)\right] + \mathrm{Cov}\left[\epsilon', m(F)\right] + \mathrm{Cov}\left[\epsilon, \epsilon'\right] & (2)\\
&= \mathrm{Var}\left[m(F)\right] & (3)
\end{aligned}$$

assuming the noise terms are uncorrelated with each other and with $m(F)$. So how much covariance there is between two measurements will depend on the variance of the true values. Reducing this covariance to a correlation would mean dividing by $\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[X'\right]}$. Under the assumptions I've just made, $\mathrm{Var}\left[X\right] = \mathrm{Var}\left[X'\right] = \mathrm{Var}\left[m(F)\right] + \mathrm{Var}\left[\epsilon\right]$, so

$$\mathrm{Cor}\left[X, X'\right] = \frac{\mathrm{Var}\left[m(F)\right]}{\mathrm{Var}\left[m(F)\right] + \mathrm{Var}\left[\epsilon\right]}$$

Under the simplifying assumption that $m(F) = F$, so there's no bias,

$$\mathrm{Cor}\left[X, X'\right] = \frac{\mathrm{Var}\left[F\right]}{\mathrm{Var}\left[F\right] + \mathrm{Var}\left[\epsilon\right]}$$

The implication is that when psychologists report the reliability of a measurement by giving the correlation between two separate measurements, that correlation only holds in a certain population, with a certain distribution of the true value F. Notice that as $\mathrm{Var}\left[m(F)\right] \to 0$, *every* measure will become "unreliable", i.e., approach zero correlation[12]. At the other extreme, as $\mathrm{Var}\left[m(F)\right] \to \infty$, *every* measure will become perfectly reliable (the correlation will approach 1).

Just as "reliability" means basically the same thing as "precision", but it's typically reported differently, "validity" for a psychologist means basically the same thing as "accuracy". But validity, too, would typically be reported by a psychologist as a covariance or correlation, in our symbols as $\mathrm{Cov}\left[X, F\right]$ or $\mathrm{Cor}\left[X, F\right]$.

$$\begin{aligned}
\mathrm{Cov}\left[X, F\right] &= \mathrm{Cov}\left[m(F) + \epsilon, F\right] & (4)\\
&= \mathrm{Cov}\left[m(F), F\right] & (5)\\
\mathrm{Cor}\left[X, F\right] &= \frac{\mathrm{Cov}\left[m(F), F\right]}{\sqrt{(\mathrm{Var}\left[m(F)\right] + \mathrm{Var}\left[\epsilon\right])\mathrm{Var}\left[F\right]}} & (6)
\end{aligned}$$

If we can make the simplifying assumption that there is no bias, $m(F) = F$, or that there is a constant bias, $m(F) = F + b$, then

$$\begin{aligned}
\mathrm{Cov}\left[X, F\right] &= \mathrm{Var}\left[F\right] & (7)\\
\mathrm{Cor}\left[X, F\right] &= \frac{\mathrm{Var}\left[F\right]}{\sqrt{(\mathrm{Var}\left[F\right] + \mathrm{Var}\left[\epsilon\right])\mathrm{Var}\left[F\right]}} & (8)
\end{aligned}$$

We saw above that "precision bounds accuracy", because variance bounds MSE. Similarly, comparing this last expression for the correlation between the measured value and the truth with the correlation between measurements should explain the saying "reliability bounds validity". But, again, "validity" in this sense will be a function of $\mathrm{Var}\left[F\right]$, increasing as the latter increases.

Notice, by the way, that validity just isn't something that can be calculated from the distribution of measurements alone, while reliability *is*.

6.1 Does the target of measurement exist?

Accuracy or validity, in this sense, *presumes* that what we're trying to measure actually exists. *This is not obvious.* In the case of temperature, we're trying to answer a question which in ordinary, pre-scientific

[12] In particular, repeated measures, *no matter how precise*, on a calibrated case with a known value of F will have zero reliability! It *is* true that we're imagining a situation where the *differences* between measurement outcomes are all so much noise...

language we'd put something like "how hot is it?" But "hot", while a simple word children use and understand correctly, turns out to actually be a very complicated concept! If you've taken a physics or chemistry course, you've probably been forced to learn distinctions between the temperature of a body, its heat content, its heat conduction, etc., which weren't fully sorted out until the discovery of thermodynamics in the middle of the 1800s. These concepts were forced on scientists, gradually over centuries, as they tried to give sensible and exact answers to "how hot is it?" question.[13]

As I mentioned in class, there used to be a theory — advanced by very serious physicists and chemists — that "fire" was a chemical element called "phlogiston". More exactly, the theory was that substances which can burn contain this element, that when they burn they release this element into the air, and the air has only a limited capacity to absorb it, which was why fires go out unless their air supply is renewed. This theory was widely accepted, even though after experiments showed that some substances *gained* mass by being burned, which would imply that phlogiston had *negative* mass. (Some quite eminent scientists were prepared to accept that.) People tried to develop ways of measuring the phlogiston content of bodies and of air. These may have been more or less precise and reliable, as measurement theorists use the words, but none of these were accurate or valid, because phlogiston does not exist[14].

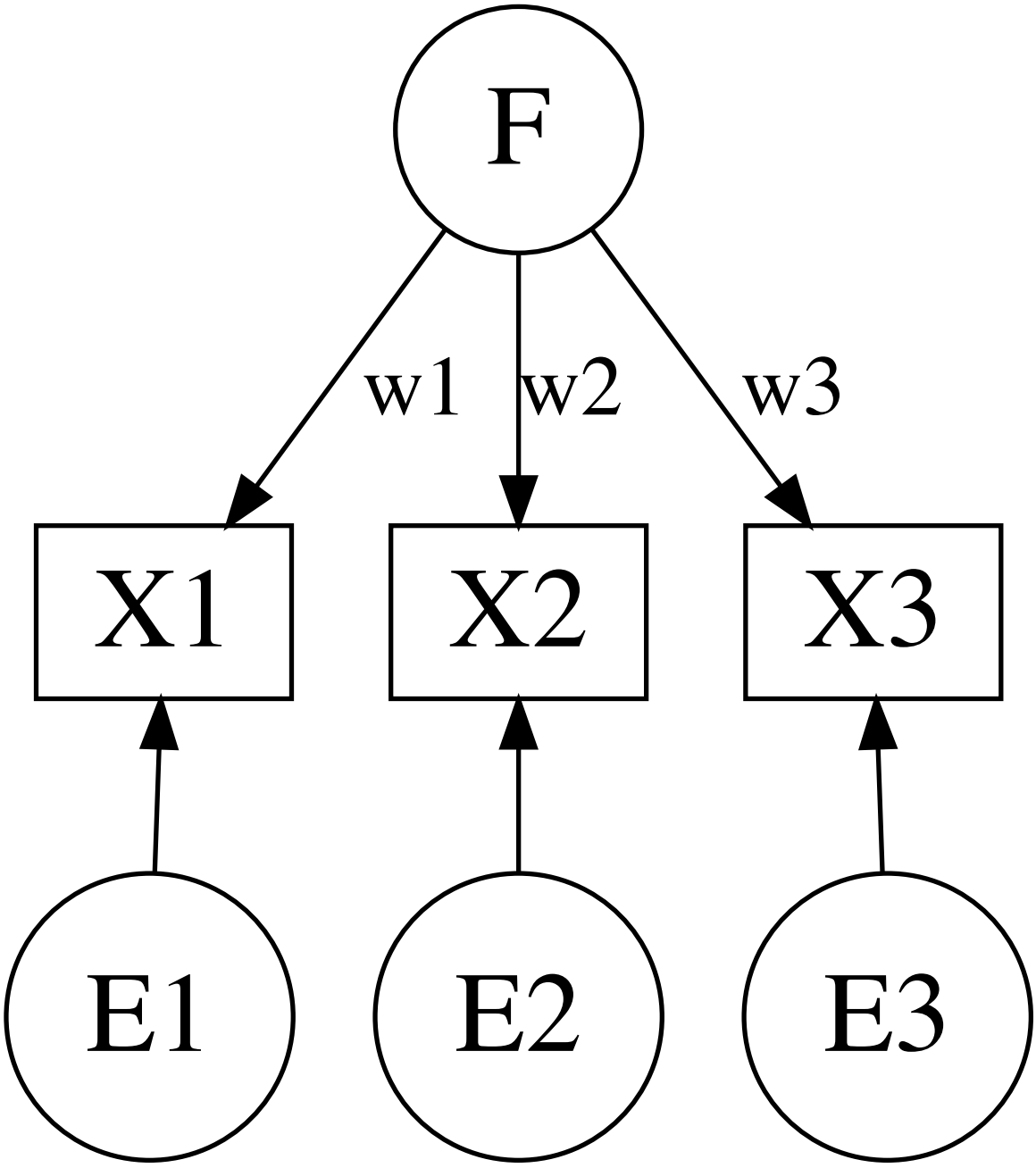
In the realm of psychological measurement, and social measurement more generally, it is often unclear if the variables we're interested in are like "hot", or like "temperature", or like "amount of phlogiston".

[13] If you heat an oven to 400 degrees and keep it there for a while, the air inside the oven and a metal rack in the middle will both be at that temperature. If you stick your bare hand in the oven for a second and wave it around in the air, you'll be fine. If you were to be so foolish as to grab the rack and hold on for that same second, you'd get a burn. (Don't grab a hot oven rack with your bare hand!) Why the difference, when both the air and the rack have the same temperature?

[14] The actual explanation is that burning is a process of very rapid combination with oxygen which releases heat; air has limited oxygen content which can be exhausted by combustion; substances can gain mass on burning because the extra mass comes from the oxygen. So "fire" is not a distinct kind of substance or element at all, and the phlogiston theory had everything almost exactly backwards.

7 Factor Models: Synthesizing Multiple Measurements

This is a good place to lay out some of the math of one of the most common measurement models, which has come to be called the **factor model**, or **factor analysis**. The classic case is shown by the next figure:



Here we have one latent or unobserved variable F, which we're trying to measure, and three observed variables X_1, X_2, X_3 (also called "manifest variables" or "observables" or "indicators"). The idea is that each observable is linearly[15] related to the latent variable:

$$X_j = w_j F + \epsilon_j$$

[15]There are nonlinear variants as well, but overwhelmingly when people talk about "factor models", they have this sort of linear relationship between latent variables and observables in mind.

where ϵ_j is mean-zero noise, uncorrelated with F, $\mathrm{Cov}\left[\epsilon_j, F\right] = 0$, and uncorrelated across the observables, $\mathrm{Cov}\left[\epsilon_j, \epsilon_k\right] = 0$ (unless $k = j$). The coefficients w_j are known as "weights" or "loadings", and show how strongly each observable is related to the latent variable. More exactly:

$$\begin{aligned} \mathrm{Cov}\left[X_j, F\right] &= \mathrm{Cov}\left[w_j F + \epsilon, F\right] & (9) \\ &= w_j \mathrm{Cov}\left[F, F\right] = w_j \mathrm{Var}\left[F\right] & (10) \end{aligned}$$

Now at this point I am going to make three customary assumptions, which don't really limit the model but do simplify the book-keeping:

- $\mathbb{E}\left[F\right] = 0$
- $\mathrm{Var}\left[F\right] = 1$
- $\mathbb{E}\left[X_j\right] = 0$

The first two assumptions are on the latent variable, and we can always *make* them true by changing the scale of the units we use to describe it and where we set the zero point for that scale. The third assumption is on the observables, we can always make it true by centering them, i.e., subtracting off their mean. So these really are "without loss of generality" assumptions. Using them, we can boil down the covariance between the observable X_j and the latent variable F to

$$\mathrm{Cov}\left[X_j, F\right] = w_j \tag{11}$$

Similarly, we can find the covariance between any two observables:

$$\begin{aligned} \mathrm{Cov}\left[X_j, X_k\right] &= \mathrm{Cov}\left[w_j F + \epsilon_j, w_k F + \epsilon_k\right] & (12) \\ &= w_j w_k \mathrm{Var}\left[F\right] + w_j \mathrm{Cov}\left[F, \epsilon_k\right] + w_k \mathrm{Cov}\left[\epsilon_j, F\right] + \mathrm{Cov}\left[\epsilon_j, \epsilon_k\right] & (13) \\ &= w_j w_k + 0 + 0 + 0 = w_j w_k & (14) \end{aligned}$$

A genuinely-restrictive assumption is usually added at this point: the noise variances don't change with F, $\mathrm{Var}\left[\epsilon_j | F = f\right] = \mathrm{Var}\left[\epsilon_j\right] \equiv \psi_j$.

7.1 Measuring the factor using the observables

Suppose we take any one of the observables, say X_j. How should we use it to measure F? Let's suppose that we know the weights or loadings w_j. Then one obvious possibility would be X_j / w_j. This would be unbiased:

$$\mathbb{E}\left[X_j / w_j | F\right] = \mathbb{E}\left[(w_j F + \epsilon_j)/w_j | F\right] = F + \frac{1}{w_j}\mathbb{E}\left[\epsilon_j | F\right] = F$$

and have certain amount of variance:

$$\mathrm{Var}\left[X_j / w_j | F\right] = \frac{1}{w_j^2}\mathrm{Var}\left[\epsilon_j | F\right] = \frac{1}{w_j^2}\psi_j$$

Overall, the variance of estimates obtained this way will be

$$\begin{aligned} VarX_j/w_j &= \mathrm{Var}\left[F + \frac{1}{w_j}\epsilon_j\right] & (15) \\ &= \mathrm{Var}\left[F\right] + \frac{1}{w_j^2}\psi_j > \mathrm{Var}\left[F\right] & (16) \end{aligned}$$

Let's call this the "scaling" estimator, since it just re-scales the observed values.

The scaling estimator, while simple and natural, is *not* the most accurate one possible. To find the most accurate linear estimator, we need to optimize. Any linear estimate will take the form $\hat{F} = bX_j$, so we want to minimize the MSE, $\mathbb{E}\left[(F - \hat{F})^2\right]$. Thus we want

$$\begin{aligned}
\beta &= \underset{b}{\operatorname{argmin}}\, \mathbb{E}\left[(F - bX_j)^2\right] && (17)\\
\mathbb{E}\left[(F - bX_j)^2\right] &= (\mathbb{E}\left[F - bX_j\right])^2 + \mathrm{Var}\left[F - bX_j\right] && (18)\\
&= 0^2 + \mathrm{Var}\left[F\right] + \mathrm{Var}\left[-bX_j\right] + 2\mathrm{Cov}\left[F, -bX_j\right] && (19)\\
&= \mathrm{Var}\left[F\right] + b^2\mathrm{Var}\left[w_j F + \epsilon_j\right] - 2b\mathrm{Cov}\left[F, w_j F + \epsilon_j\right] && (20)\\
&= 1 + b^2(w_j^2 + \psi_j) - 2bw_j && (21)\\
\frac{d}{db}\mathbb{E}\left[(F - bX_j)^2\right] &= 2b(w_j^2 + \psi_j) - 2w_j && (22)\\
\beta &= \frac{w_j}{w_j^2 + \psi_j} && (23)
\end{aligned}$$

Since $\psi_j > 0$, the optimal coefficient β will be smaller (closer to zero) than $1/w_j$. How *much* smaller depends on how much of the observable is noise; the more noisy the measurement, the more the coefficient should be "shrunk" towards zero. I will leave it as character-building exercises to calculate $\mathbb{E}\left[\beta X_j | F\right]$, $\mathrm{Var}\left[\beta X_j | F\right]$, $\mathrm{Var}\left[\beta X_j\right]$, and the difference in MSEs between this optimal estimate and the simple scaling estimator.

(βX_j is the optimal *linear* estimator, but might there be a more accurate non-linear one? As it happens, no, not under these assumptions, though that's surprisingly tricky to show — at least, I don't know of a way of doing so that's short enough to include here.)

7.2 Two (or more) observables

What if we want to use *two* of the observables *together* to estimate F? If we use X_j and X_k, scaling would give us two estimates, X_j/w_j and X_k/w_k, and we could try doing some sort of average, say a weighted average that gives more weight to the less-noisy observable. But this is very ad hoc; why don't we try looking for the optimal linear combination again? We'll now have two coefficients to adjust, because our estimate will take the form $b_j X_j + b_k X_k$. Now we *could* go through this all again, but if you do so you will quickly get a pair of simultaneous linear equations for the optimal coefficients:

$$\begin{aligned}
\beta_j(w_j^2 + \psi_j) - \beta_k w_j w_k - w_j &= 0 && (24)\\
\beta_k(w_k^2 + \psi_k) - \beta_j w_j w_k - w_k &= 0 && (25)
\end{aligned}$$

Trying to solve these by elementary algebraic manipulations is possible but a pain. Trying to do the same thing with three or more observables is even more of a pain — so painful, in fact, that it was one of the motivations for the Ancestors to develop linear algebra! Trying to

There is a much more transparent and general procedure, which however requires you to take a factor on trust:

> If F is a random variable, $\mathbb{E}\left[F\right] = 0$, and $\vec{X}$ is a vector of random variables, $\mathbb{E}\left[\vec{X}\right] = 0$, then the optimal linear coefficients, $\beta = \mathrm{argmin}_b\, \mathbb{E}\left[(F - b \cdot \vec{X})^2\right]$, are given by $\beta = \mathrm{Var}\left[\vec{X}\right]^{-1} \mathrm{Cov}\left[\vec{X}, F\right]$.

(See Shalizi (n.d.), ch. 2.) Applied here,

$$
\begin{aligned}
\beta &= \begin{bmatrix} \mathrm{Var}[X_j] & \mathrm{Cov}[X_j, X_k] \\ \mathrm{Cov}[X_j, X_k] & \mathrm{Var}[X_k] \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}[X_j, F] \\ \mathrm{Cov}[X_k, F] \end{bmatrix} && (26)\\
&= \begin{bmatrix} w_j^2 + \psi_j & w_j w_k \\ w_j w_k & w_k^2 + \psi_k \end{bmatrix}^{-1} \begin{bmatrix} w_j \\ w_k \end{bmatrix} && (27)\\
&= \frac{1}{(w_j^2+\psi_j)(w_k^2+\psi_k) - w_j^2 w_k^2} \begin{bmatrix} w_k^2 + \psi_k & -w_j w_k \\ -w_j w_k & w_j^2 + \psi_j \end{bmatrix} \begin{bmatrix} w_j \\ w_k \end{bmatrix} && (28)\\
&= \frac{1}{(w_j^2+\psi_j)(w_k^2+\psi_k) - w_j^2 w_k^2} \begin{bmatrix} w_j(w_k^2+\psi_k) - w_j^2 w_k \\ w_k(w_j^2+\psi_j) - w_j w_k^2 \end{bmatrix} && (29)
\end{aligned}
$$

You can imagine how unhelpful the parallel formulas are for three or more observables. But the important point is that there *are* formulas, and so we can calculate the optimal β coefficients knowing the variances and covariances of the observables (which are all functions of the loadings w_j and the noise variances ψ_j) and the covariances of the observable and the latent variable (which are all just the loadings w_j).

7.3 Some implications of the one-factor model

We have already worked out some basic implications of the one-factor model for the covariances among the observables:

$$\mathrm{Cov}[X_j, X_k] = w_j w_k \tag{30}$$

unless $j = k$, in which case

$$\mathrm{Cov}[X_j, X_j] = \mathrm{Var}[X_j] = w_j^2 + \psi_j \tag{31}$$

If we have p observables over all, there is a $p \times p$ matrix which gives all the variances and covariances among observables. In the one-factor model, this has a very peculiar structure:

$$\mathrm{Var}[X] = \mathbf{w}^T \mathbf{w} + \psi \tag{32}$$

where $\mathbf{w}$ is the $1 \times p$ matrix of the w_j's, and ψ is the diagonal matrix of the ψ_j's. A matrix like $\mathrm{Var}[X]$ is called "rank one plus noise", and it's a very special and restricted kind of matrix; most covariance matrices are not of this form.

One way to see how peculiar these covariances are is to consider groups of four variables, say X_j, X_k, X_l, X_m (all distinct). We have

$$
\begin{aligned}
\mathrm{Cov}[X_j, X_l] &= w_j w_l && (33)\\
\mathrm{Cov}[X_j, X_m] &= w_j w_m && (34)\\
\frac{\mathrm{Cov}[X_j, X_l]}{\mathrm{Cov}[X_j, X_m]} &= \frac{w_l}{w_m} && (35)
\end{aligned}
$$

But also

$$
\begin{aligned}
\mathrm{Cov}[X_k, X_l] &= w_k w_l && (36)\\
\mathrm{Cov}[X_k, X_m] &= w_k w_m && (37)\\
\frac{\mathrm{Cov}[X_k, X_l]}{\mathrm{Cov}[X_k, X_m]} &= \frac{w_l}{w_m} && (38)
\end{aligned}
$$

So

$$\frac{\mathrm{Cov}[X_j, X_l]}{\mathrm{Cov}[X_j, X_m]} = \frac{\mathrm{Cov}[X_k, X_l]}{\mathrm{Cov}[X_k, X_m]} \tag{39}$$

Notice that this equation doesn't involve the loadings, just covariances between observables, which are things we can calculate directly from the data. This is called a "tetrad equation", because it involves four of the variables, and is one of the peculiar implications of the one-factor model.

To be clear on the logic, *if* the one-factor model is right, *then* this equation should hold for any four observables. Turned around: if this equation does *not* hold, for even one tetrad of variables, the one-factor model cannot be right. Of course, when we estimate covariances from data, we need to make allowances for the error in those estimates, but in principle this gives us a way to tell when the one-factor model breaks down.

(You might well ask whether the tetrad equations hold if *and only if* the one-factor model is right, i.e., if there are other models which also imply those equations. The answer is that models with more than one factor do *not* imply those equations, but that there are non-factor models which do.)

7.4 Fitting the one-factor model

If we want to use our factor model to measure the underlying latent variable, we need to find the loadings w_j. How can we do that?

Let's go back to the equation for covariance between two observables:

$$\mathrm{Cov}\left[X_j, X_k\right] = w_j w_k \ (j \neq k) \tag{40}$$

If we have p different observables, there are p loadings $w_1, w_2, \ldots w_p$ to find, but there are $p(p-1)/2$ distinct covariances among the observables. So there will be more equations than unknowns. We could select p of the equations and solve them for the unknowns, but that throws away the information in the other equations. A more satisfying approach is to estimate the w_j's by (nonlinear) least squares,

$$\hat{w} = \underset{w}{\mathrm{argmin}} \sum_{j \neq k} \left(\widehat{\mathrm{Cov}\left[X_j, X_k\right]} - w_j w_k\right)^2 \ .$$

Of course, sample covariances between observables aren't true covariances, and some covariances will be more accurately estimated than others, so we might really want to do *weighted* nonlinear least squares, but that requires using some sampling theory to work out the standard errors of $\widehat{\mathrm{Cov}\left[X_j, X_k\right]}$.

An alternative is to go back to the equation for the variance matrix as a whole:

$$\mathrm{Var}\left[X\right] = \mathbf{w}^T\mathbf{w} + \psi \tag{41}$$

If we could find ψ, then

$$\mathbf{w}^T\mathbf{w} = \left(\mathrm{Var}\left[X\right] - \psi\right)$$

so we could find $\mathbf{w}$ by a kind of matrix square root, which computers are now very good at, though in a way impossible to explain without linear algebra[16].

Conversely, if we knew the loadings w_j, finding the noise variances ψ_j would be easy:

$$\psi = \mathrm{Var}\left[X\right] - \mathbf{w}^T\mathbf{w} \tag{42}$$

This suggests an iterative scheme, where we start with a guess at ψ, use that guess to estimate $\mathbf{w}$, then use that get a better estimate of ψ, and keep alternating until convergence. A common starting guess for ψ_j is the variance of the residuals from regressing X_j on all the other X's[17].

There are other ways to estimate factor models. If one is willing to assume that F and the ϵ_j's are all Gaussian, then it turns out to be fairly straightforward to write out a likelihood function and maximize it, though it still involves making a starting guess about the ψ_j's. There are also more modern (and algorithmic) approaches to low-rank matrix approximation than nonlinear least squares.

[16] $\mathrm{Var}\left[X\right]$, being a variance matrix, is symmetric and "positive semi-definite", $u^T\mathrm{Var}\left[X\right]u \geq 0$ for any p-dimensional vector v. Any such matrices has an **eigendecomposition**, $\mathrm{Var}\left[X\right] = \mathbf{v}^T\mathbf{d}\mathbf{v}$, where $\mathbf{v}$ is the matrix whose columns are the eigenvectors of $\mathrm{Var}\left[X\right]$, and $\mathbf{d}$ is the diagonal matrix of eigenvalues, all of which are non-negative. One can show that $\mathrm{Var}\left[X\right] - \psi$ is *also* a symmetric and positive semi-definite matrix. (Symmetry is easy, PSD is a bit more work.) So it, too, will have an eigendecomposition. But, because $\mathbf{w}^T\mathbf{w}$ is a rank one matrix, $\mathrm{Var}\left[X\right] - \psi$ can have only *one* positive eigenvalue, all others being zero, say λ_1, with corresponding eigenvector v_1. Then setting $w = \sqrt{\lambda_1}v_1$ gives us our solution. (If there is some rounding and/or estimation error and $\mathrm{Var}\left[X\right] - \psi$ isn't *exactly* rank one, take its leading eigenvalue and eigenvector, and hope the other eigenvalues are all very close to zero.)

[17] If we could regress X_j on F, the variance of the residuals would converge on ψ_j as $n \to \infty$. The other observables won't be quite as informative about X_j as F would be, so this starting point should tend to over-estimate the noise variances.

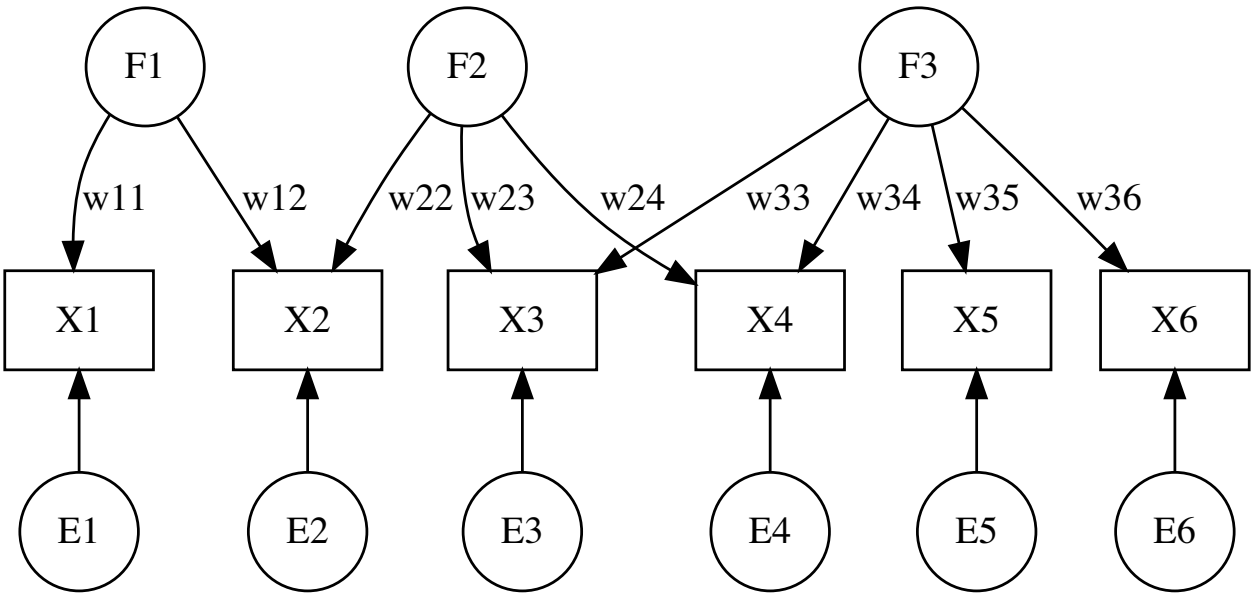
7.5 Model checking

We saw above that the one-factor model imposes some pretty severe restrictions on the covariance matrix. These are testable. In particular, we can form a test statistic which is something like $\sum_{jk} (\mathrm{Var}[X]_{jk} - (\mathbf{w}^T\mathbf{w} + \psi)_{jk})^2$ and work out its distribution under the null hypothesis that the one-factor model is indeed correct, though this might involve some simulation or bootstrapping[18].

If the one-factor model fails tests like this, it indicates that (with high confidence) the model is wrong. If the one-factor model *passes* a test like this, and the test had high power to detect violations of these restrictions, that *is* evidence of a kind in favor of the one-factor model. The reason I hedge with "of a kind" is that there are *other* models which *also* imply the same restrictions on the covariance matrix (see further reading). The reason I hedge with "and the test had high power" is that the power of any test grows with sample size, and many tests of this form would require *very* large sample sizes to detect even moderately large departures from the one-factor model's assumptions.

7.6 Multiple Factors

If a one-factor model doesn't fit, a natural thought is that there might be another source of shared variance among at least some of the observables, i.e., at least one other factor. This gets us into *multiple* factor models. Graphically, they look like the next figure:



The basic equation is now

$$X_j = \sum_{i=1}^{q} w_{ij} F_i + \epsilon_j$$

where there are q distinct factors, $F_1, F_2, \ldots F_q$, and the noise ϵ_j is still mean zero and uncorrelated with all the factor variables (and all the other noises, $\mathrm{Cov}[\epsilon_j, \epsilon_k] = 0$ unless $j = k$). As in the one-factor case we can, without any real loss of generality, assume that $\mathbb{E}[F_i] = 0$, and that $\mathrm{Var}[F_i] = 1$. It is less obvious,

[18] If we're willing to make assumptions on the distribution of the factor and the noise terms, e.g., that they're Gaussian, we can calculate the log-likelihood ratio between the null hypothesis that the one-factor model is right vs. the alternative of an unrestricted covariance matrix. This will generally be a more powerful test, and has the advantage of having a known large-sample distribution (twice the log-likelihood ratio is χ^2 with $p(p+1)/2 - 2p$ degrees of freedom), *when those assumptions hold*.

but still true, that we can without loss of generality assume the factors are uncorrelated with each other, $\mathrm{Cov}\left[F_i, F_i'\right] = 0$ unless $i = i'$, but this is nonetheless true[19].

The q-factor model still implies restrictions on the covariance matrix of the observables. If we gather up all the loadings into a $q \times p$ matrix, say $\mathbf{w}$, then we get

$$\mathrm{Var}\left[X\right] = \mathbf{w}^T\mathbf{w} + \psi$$

The matrix $\mathbf{w}^T\mathbf{w}$ has rank $q < p$, so the over-all covariance matrix is said to be "low rank plus noise".

Estimation of factor scores from observables proceeds much as before — we can find the optimal linear estimator using the variance matrix of the observables, and the loading-implied covariances between each observable and each factor. Estimation of the factor loadings w_{ij} and noise variances ψ_j also proceeds much as before, but with some additional mathematical complications[20].

I said above that, *mathematically*, we lose no generality by assuming the latent factor variables are uncorrelated with each other. If you start with a model where the factors are correlated, I can always construct another model with uncorrelated factors, and our two models imply *exactly* the same distribution over observables. This means that nothing about the *data* could decide between them. *Substantively*, however, if one factor is what we're trying to measure and another factor is some confounding variable that systematically messes with our measurements, it may make more sense to use two (or more) correlated factors. A great deal of ingenuity has been spent, over the last century, in coming up with procedures which try to transform factor models so that their loading matrices seem simple, plausible, have lots of zeroes (are "sparse"), or are otherwise aesthetically pleasing. Little to nothing is known about when these procedures converge on the truth (assuming there even is such a thing as a true factor structure).

[19]The trick is that we start with q correlated factor variables, we can always construct q linear combinations of those variables which are uncorrelated, and express the observables as linear functions of those new, uncorrelated variables; this is observationally indistinguishable from the original model. See the chapter on factor models in Shalizi (n.d.) for details.

[20]Notice that in the one-factor model, if we replace all the loadings w_j by $-w_j$, nothing changes about the distribution of observables. With two or more factors, if we replace the loading matrix $\mathbf{w}$ by $\mathbf{ow}$, where $\mathbf{o}$ is a $q \times q$ matrix such that $\mathbf{o}^T\mathbf{o} = \mathbf{I}$, then again nothing changes about the distribution of observables. Matrices with this property are called "orthogonal" matrices, and outstanding examples are ones which flip around coordinate axes (as in the one-factor case) and rotate coordinate axes around the origin. Because of the latter, this issue is called the "rotation problem". There isn't really a *solution* to it, because it's the algebra's way of saying we can use any coordinate system we like for the hidden variables and our choice of coordinates doesn't *really* change anything. For more, see the chapter on factor analysis in Shalizi (n.d.).

8 Further reading

Measuring attitudes is a specific form of psychological measurement. On psychological measurement in general, I strongly recommend Borsboom (2005);Borsboom (2006). Zeller and Carmines (1980) is a straightforward and readable, though now slightly old-fashioned, introduction to psychological and social measurement, making a lot of use of factor models; it has a lot of sound things to say, but is very optimistic about what those models can achieve.

On conflicts over the IAT, I've given additional references on the class homepage. Among these, I will put in a specific plug for Machery (2021).

Measurement in psychology has a long and contested history, which has included some truly startlingly bad ideas being very widely endorsed. Michell (1999) is, as its subtitle says, "a critical history" (sometimes too critical: [http://bactra.org/reviews/michell-measurement.html]). Borsboom (2005) is, again, an excellent guide here. Taylor (1997) is a standard introduction to the way people in the physical sciences think about measurement.

On factor models, the relevant chapter of Shalizi (n.d.) has a lot more from this point of view, including references to *non*-factor models which have the same implications for the covariances between observables. For further statistical details, I recommend Bartholomew (1987).

References

Agadjanian, Alexander, John M. Carey, Yusaku Horiuchi, and Timothy J. Ryan. 2021. "Disfavor or Favor? Assessing the Valence of White Americans' Racial Attitudes." Electronic preprint, SSRN/3701331. https://doi.org/10.2139/ssrn.3701331.

Alba, Richard. 2020. *The Great Demographic Illusion: Majority, Minority, and the Expanding American Mainstream*. Princeton: Princeton University Press.

Bartholomew, David J. 1987. *Latent Variable Models and Factor Analysis*. New York: Oxford University Press.

Borsboom, Denny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge, England: Cambridge University Press.

———. 2006. "The Attack of the Psychometricians." *Psychometrika* 71:425–40. https://doi.org/10.1007/s11336-006-1447-6.

Carney, Riley K., and Ryan D. Enos. 2017. "Conservatism and Fairness in Contemporary Politics: Unpacking the Psychological Underpinnings of Modern Racism." Unpublished manuscript. https://scholar.harvard.edu/files/renos/files/carneyenos.pdf.

Cramer, Katherine. 2020. "Understanding the Role of Racism in Contemporary US Public Opinion." *Annual Review of Political Science* 23:153–69. https://doi.org/10.1146/annurev-polisci-060418-042842.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74:1464–80. https://doi.org/10.1037//0022-3514.74.6.1464.

Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji. 2003. "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm." *Journal of Personality and Social Psychology* 85:197–216. https://doi.org/10.1037/0022-3514.85.2.197.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 140:1–55.

Machery, Edouard. 2021. "Anomalies in Implicit Attitudes Research." *Wiley Interdisciplinary Reviews: Cognitive Science* forthcoming:e1569. https://doi.org/10.1002/wcs.1569.

McDermott, Drew. 1976. "Artificial Intelligence Meets Natural Stupidity." *ACM SIGART Bulletin* 57:4–9. https://doi.org/10.1145/1045339.1045340.

Michell, Joel. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Cambridge, England: Cambridge University Press.

Shalizi, Cosma Rohilla. n.d. *Advanced Data Analysis from an Elementary Point of View*. Cambridge, England: Cambridge University Press. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV.

Taylor, John R. 1997. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Second. Sausalito, California: University Science Books.

Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the Social Sciences: The Link Between Theory and Data*. Cambridge, England: Cambridge University Press.