Measuring Attitudes to and/or Prejudices against Groups

36-313, Fall 2021

9 November 2021, Lecture 20

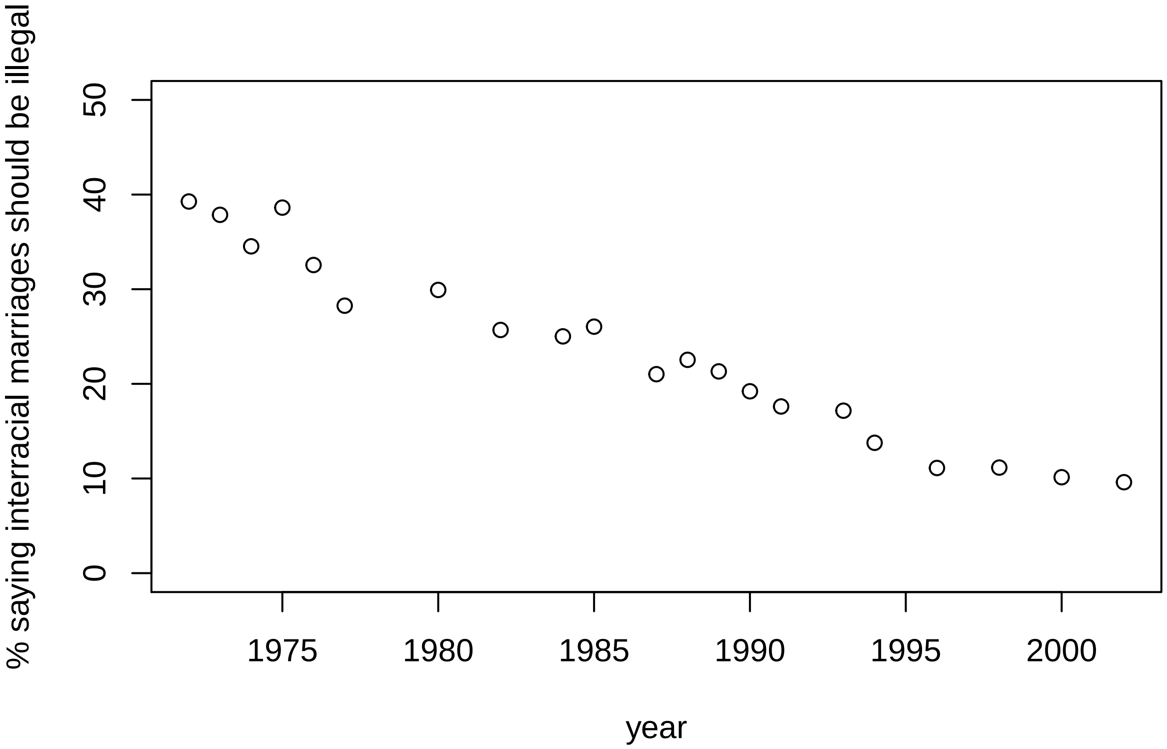
- The last few lessons have been implicitly about **measurement**
  - Tests like the SAT, and even more specific subject-matter tests, try to measure knowledge of facts and/or skills
    - "Face validity": They ask you to demonstrate that knowledge and/or skill
    - Process: It's hard to see how you 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
  - IQ testing is much less successful as an effort at measuring "general intelligence"
    - No theory about what should or should not be on such a test, despite a lot of time
    - 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)
    - Lots of other explanations for patterns of correlation among sub-tests, which would also explain why IQ scores correlate with lots of life outcomes
  - 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.

1. 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*.
  - 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:
  - "Do you feel favorable or unfavorable towards X?" (That is, a binary question)
  - Ordinal scales, of the "strongly disagree, disagree, neutral, agree, strongly agree, no opinion" form, or the "strongly disapprove, disapprove, neutral, approve, strongly approve, no opinion" form. 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 very sensible, and so means don't make a lot of sense.
    - 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 rather dubious 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.
      - 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 estiamate all the parameters.
    - 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.)
    - It's 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.
  - 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.

- 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-taker 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-taker 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 survey 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, etc. (Alba 2020), but we *also* have reasons to think that *some* of the change visible in that graph is increasing desirability bias.

Not quite so explicit attitude measures

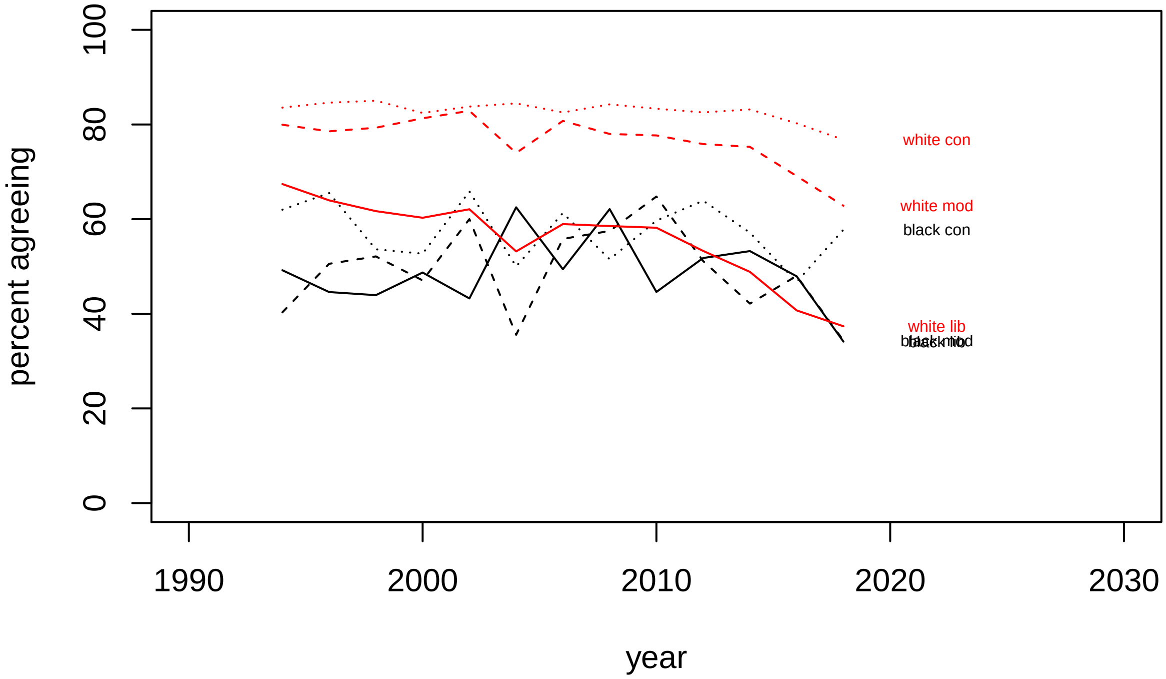
- One way to try to get around desirability bias is to ask questions which aren't so *directly* about attitudes towards the group in question, so that people will feel more comfortable giving honest answers.
- A very prominent instance of this, since the 1980s, has been the use of "modern racism" or "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:

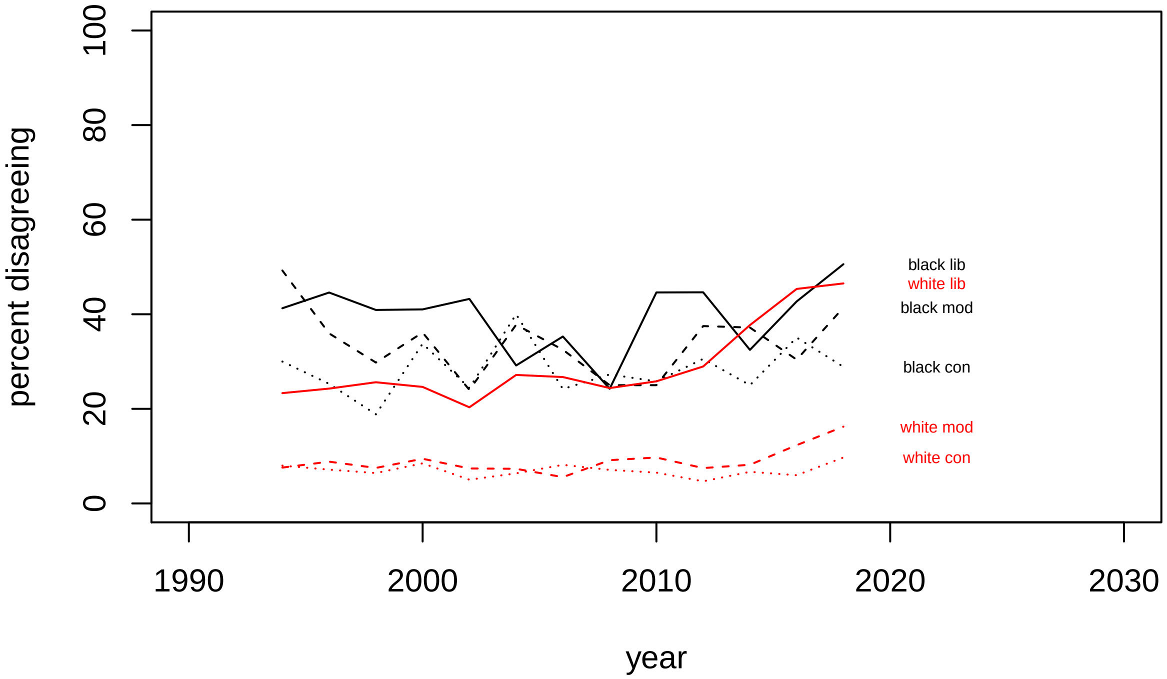
[1] In the homework, you'll see examples of the kinds of questions used to measure racism in the 1940s and 1950s, which were *much* more blunt.

> 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. This 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.
  - It is not clear that this is *still* what the question meant in 2016, or means in 2021, or will mean in 2030.
  - It’s also not clear that it 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:





Here I've sub-divided respondents by race, and by their political views as liberals, moderates and conservatives[2]. I've also collapsed "strongly agree" and "agree" into a single level of "agree", and likewise with "disagree". 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 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)[3]. It is *imaginable* that this is all because of deeply internalized anti-black racist views on the part of 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 right now.)
- All of which said, "modern racism" scales *do* at least predict certain kinds of behavior (Cramer 2020).[4]

[2] 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.

[3] Figures like this inevitably suggest that the same group of people is changing their minds about the question being asked. 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 identify themselves as liberals. Panel or longitudinal data, tracking the same people over time, is by far the best way to settle such doubts.

[4] 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).

Implicit attitude tests

- One big problem with asking people about their attitudes, and even with asking them questions that hint about 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.
- Outstanding example: 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.
  - So if X is (strongly) associated with Y, and we ask people to do something where they need to associate A with B, that should be fast.
  - But if we ask people to d something where they need to associate 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
  - 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] . To be clear, faster reactions in phase 1 than in phase 2 are supposed to indicate the extent to which you have the "A good, B bad" association
  - A/B pairs where this has been used: black/white, male/female, Japanese/Korean, insects/flowers, ...
- 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
  - It's not at all clear what this is measuring.
    - What's being *calculated* is the difference in 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")
    - *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*

[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.)

    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. 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
  - 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
    - 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 group differences. 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.
  - 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.

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

[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.)

measures racial resentment. But whether it *actually* does so is a complicated and debatable scientific hypothesis[8]. 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.

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.

On conflicts over the IAT, I've given additional references on the class homepage. 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]).

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.

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.

[8] 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.

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.

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