A Misattribution Approach to Implicit Attitudes:
Implications for Measurement and Theory

Keith Payne
University of North Carolina

Well, thank you to the conference organizers for organizing this conference, and thank you for the invitation to be here. I think that I'd like to start my talk today by stating the obvious, which is that attitudes matter in election studies, though it goes without saying that attitudes towards political candidates matter in shaping vote choices. Attitudes towards the act of voting itself — that may go without saying, but I think that I'm the third person who's pointed it out this morning, so I'm glad that we all shared that basic starting point.

But even if you take a step back and look a little bit more broadly, social attitudes towards specific policies make a big difference in shaping vote choice; for example, in the elections that are underway right now we see attitudes towards the Iraq War having a big difference in shaping vote choice. But also if you take a further step back, attitudes towards social attitudes in general like intergroup attitudes make a big difference in shaping both attitudes towards specific policies and attitudes towards candidates. So, for example, there has been a long-running discussion about the extent to which racial attitudes shape and motivate attitudes towards policies like affirmative action. Now, the American National Election Study has been collecting data on self-reported intergroup attitudes for a very long time. This is actually a mistake. That should be 1948, as I learned this morning.

What I really want to focus on today, similar to the theme of Brian [Nosek]'s talk, is implicit measures of people's social attitudes, and this is particularly interesting when it comes to racial attitudes and somewhat controversial social attitudes, because this is a domain in which implicit and explicit measures often look very different. So we know very well that racial attitudes have been changing in America over the last century, and in particular since the Civil Rights movement so you can see this in just about any measure, any indicator of racial attitudes you want to look at. I just pulled these data as an example. This is attitudes towards racially integrated housing asked in a number of ways. For example, "Is it okay with you to have black neighbors living next door, or living in great numbers in the same place where you live, or do you support an open housing law?" and so forth.

So it's very clear that over time we've seen great increases toward racial egalitarianism in the American public, but this is all in terms of self-reported explicit attitudes. What I want to focus on is implicit measures of attitudes. And for those of you who are not familiar with implicit measures, what I mean here is that we're looking at measures that don't directly ask somebody to self-report their attitude. Instead we asked people to do some task, some objective task, and then we infer from their behavior on that task what their attitude must be. Now, the advantage of doing things that way is that because you're not asking people to introspect and self-report, you're not limited by self-knowledge or being aware that "this is my attitude." Another advantage

of that way of arranging things is that it can be very difficult for people to self-present on these kinds of measures, because they're concentrating on doing a task and don't necessarily know or control what inferences you're making about their attitudes indirectly.

Now, Brian talked about the IAT, which is one of the most commonly used indirect measures of attitudes today. There are other measures such as evaluative priming that Russ Fazio has developed and others. Nosek's "go, no-go" association task and variants of other cognitive tasks like the Simon task and variations on the Stroop task.

In general, I think that most social psychologists would agree that implicit tests of race bias, in particular, often show more biases than people report on their self-reports. So one way of looking at this question is to think about how far back do we have to go before our explicit attitudes represent what we see today in terms of our implicit attitudes. So we have to go back a long time before people are willing to self-report the same degree of biases that we see on a lot of implicit tests today.

But it's important to keep in mind here that you're going to draw different conclusions, depending on the different kinds of measures that you use. For example, as Brian just overviewed, when you use the IAT you tend to find a large majority of Americans showing anti-black or pro-white biases. You tend to find a low, but positive correlation between implicit and explicit measures of racial attitudes when you're using that instrument. You can contrast that with using the evaluative priming procedure.

In Russ Fazio's work, he reports showing across lots of studies over time — now this is with college students — about 50 percent of his subjects show pro-white or anti-black bias, and that means 50 percent are showing either no bias or a pro-black bias. So the distributions are very different, if you compare the different kinds of measures here.

Now, there are lots of reasons that that might be, and there is active debate and controversy and discussion going on in the field about how to interpret these differences between measures. We can talk about that later if you'd like to, but I bring this up to point out that the way you measure these things matters. Not just for sort of dry psychometric analyses, but also for the kinds of conclusions that you're going to take away when you ask the same questions. So because there's no one gold standard for how to measure implicit attitudes, it's important to have a multi-pronged approach in which you use complementary measures. I think that the way that we're going to answer this question, ultimately, is by triangulating with different kinds of measures.

So what I'd really like to do today is to describe a new kind of approach to implicit measurement that my colleagues and I have been developing and describe what it can add to or how it can complement the other

methods that we already have available, and then I'm going to end by pointing out some questions that I think that this method is particularly well suited to answering, especially in the context of studies of election.

So the measure that I'm going to talk about is what we call the "affect misattribution procedure," and sometimes I abbreviate it AMP or "amp" for short. Now, how this differs from most other implicit measures out there — one way that it differs is that rather than looking at reaction time, this is an indirect measure of evaluative content. That is, we're looking at an evaluation, per se. Now, that ends up having some important consequences, and I think that there are a couple of advantages of doing it this way. One thing is that it's fast, simple and easy to administer. I'll walk you through the procedure in a moment, but when we do it in the laboratory where we have an hour experimental session, and we can have as many trials as we want, we only find it necessary to use about 36 to 48 trials, and that takes about three, four to five minutes to do.

If we were going to include this in sort of a large-scale survey, it could be easily enough reduced to a one or two-minute task, and the reason we can do that is because it tends to have high reliability. So with 36 trials the reliability is quite good. It wouldn't hurt to reduce it to something like 12, if you really needed to. Another strength that I think this method has is the high degree of specificity. I'm going to show you some data on separating within a general evaluation, just separating specific emotional responses that people might have.

So this is what the affect misattribution procedure looks like. It's a priming task and it's based on early work by Murphy & Zajonc. We flash a prime picture like this baby here, briefly but visibly, so in the study that I'm going to tell you about — it's 75 milliseconds. So if you're looking at a word for 75 milliseconds, that's very fast, but a colored picture you see very clearly at 75 milliseconds. So it's briefly, but none of this stuff that I'm going to talk about today is subliminal. Then we flash an ambiguous picture. In all of the studies that I'm using today we used Chinese pictographs, and then we asked people to make a pleasantness judgment about the pictograph. With most of the studies I'm going to talk about today is a binary judgment — "Is it more or less pleasant than the average Chinese pictograph?" So that's obviously a very ambiguous thing to ask a non-Chinese speaker, right? So just to give you some examples of what these things look like, the participants have no idea whether they've seen the same one before or not, but nonetheless they go ahead and make a decision. This is what some of the unpleasant items looked like in the first study I'm going to show you, and here are some of the pleasant items. These are taken from a norm set of images.

So in this first study, we had two groups, because when doing something as simple as evaluating this pictograph — not under time pressure and not with subliminal priming — one's intuition is that, "Sure the prime might have some effect, but people are easily going to be able to correct for that." You can see the item, you know exactly what is going on, and so you might think that people will correct for it. In this study, we have one group with no warnings and another group with a warning that they should correct. So we

wanted to explicitly test how well people can adjust for the impact of that prime on their evaluation pictograph. So in the "no warning" condition, you can read it for yourself as it just simply says evaluate the Chinese character. We do always include instruction that you shouldn't judge all of the characters as pleasant or unpleasant, but instead it's more or less pleasant than the average Chinese character. So you get some distribution in responses, because in general, people think that they're pleasant.

Now, if you look at the "warning" instruction they get the same instructions plus the specific warning. It's important to note that having just seen a positive image can sometimes make you judge the Chinese character more positively than you otherwise would. Having just seen a negative image can make you judge the Chinese character more negatively. Because we're interested in studying how people can avoid being biased, please try your absolute best not to let the real life images bias your judgments of the Chinese characters. So we tell them exactly how the prime is likely to influence them, and we urge them not to do it. And we even say, "This study is about how you can not do that." So if they're ever going to correct, we expect this to be a condition in which they're going to do that.

So what you can see here is the proportion of pleasant judgments of the pictographs and those are overwhelmingly more pleasant judgments after we flash pleasant primes, as compared to unpleasant — and the neutral prime is just a gray square that we've included as a baseline here. The affective warning is vanishingly small. Here it's about three percent and not close to significant. So what we see is that people's evaluations of those primes strongly influence the evaluations of the pictographs, and there doesn't seem to be a lot that participants can do to overcome that. We've done some other studies looking at exactly why that is, and we can talk about that if you want, but for now the point that I want to make is that when you're using items that you know people have pleasant or unpleasant attitudes towards, the effect size is quite large. The reliability is very good. Here we just scored each trial as an item and did internal consistency…*[Portion contains technical difficulties for about 3 min.]*

—*(Recording resumes:)* From study to study, the neutral prime sort of bounces around more so than some of the other things, so I wouldn't put too much weight on exactly where the neutral prime is here. Especially when you're using faces, it's difficult to know exactly what a neutral comparison is for a face. Another way to look at the data is to score—*[technical difficulties]* And so if you look at that as an individual difference measure and look at how it correlates with these various explicit attitude questionnaires, we see across the board strong positive correlations between their performance on the AMP task here, and what they're reporting as their attitudes toward the candidates. If you look at the vote choice, there's also a strong ability to predict how people are going to vote, and again, a high degree of reliability for the AMP itself. So it seems to be functioning well as an individual difference measure of attitudes, functioning like we would expect it to be.

The next question that we want to look at is whether if we move away from the political domain — a domain where people are perfectly willing as we just talked about in Brian's talk, perfectly willing and able to express their political viewpoints on a self-report measure — do you find similar findings when you move into a more controversial domain, where implicit attitudes actually become more useful for reasons of social desirability, self-presentation and so forth?

So in the next study we looked at racial attitudes, and in this one we used white and black male faces as primes, and also matched on a number of features like attractiveness. We included an explicit measure of feelings towards blacks and whites, and we used included *(unintelligible*) Devine's measure of motivation and control prejudice. This is essentially a measure of how much people say that they're interested in responding in a non-prejudiced way.

Overall, if you look at the implicit/explicit correlation—well, the mean results seem to have disappeared, I can tell you. We found basically in-group bias for both white and black subjects. Whites showed a pro-white bias, and blacks showed a pro-black bias. If you look at the individual difference correlations with the explicit feeling thermometer and performance on this procedure, you find a fairly strong positive correlation. It's stronger than you often see in implicit/explicit studies on the topic of race, at least. But it's important to think separately about those who are high on the motivation and control prejudice versus those who are low on the motivation and control prejudice.

So if you separate those out, what you see — this is regression sloped plotted at one standard deviation above and below the mean of motivation and control prejudice. For those who are high on the motivation and control prejudice — these are folks who say, "No, it's really important to me not to be the kind of person who expresses any kind of prejudice" — their attitudes towards African Americans were quite positive across the board, and completely uncorrelated with the AMP. So what the feeling thermometer is telling you and what the AMP is telling you, are totally unrelated to each other. They're totally independent pieces of information for those guys.

But if you look at the low motivation subjects, and these are the folks who say, "My attitudes are what they are and I don't really care who knows," you see a very strong correlation between implicit and explicit attitudes. So strongly you start thinking that they may be getting at very, very similar constructs.

So sometimes we want to know more than just a person's global evaluation. Sometimes it's informative to know exactly how an issue or a person or a candidate makes a person feel. What I want to look at next is whether the measure can be used to disentangle specific emotional reactions that go together to make up someone's over global evaluation. So here we're going to look at whether the measure can take apart specific emotional reactions drawing on some of Elliott Smith's work on intergroup specific emotions.

The idea here is that two people might be at the exact same point on a general attitude measure from very favorable to unfavorable, for example, but they might be at that same point for very different reasons. So for example if you're talking about attitudes towards Muslims, two people might give you the same mildly or moderately negative attitude, but one because they're afraid of them, and the other one because they're very angry at them. Those two kinds of different emotional reactions are going to result in very different kinds of action tendencies. They're going to predict a fear kind of response. It might predict avoidance or defensive behaviors, or as an anger kind of response might predict much more aggressive behaviors.

So what we did to see if we could separate these emotions implicitly is to arrange a task with four different kinds of judgments about the pictographs, one judgment per block. So in one block they were told to guess the meaning of the Chinese pictograph, and they were told that the best strategy for doing this is that people have this ability across languages to sort of guess with above chance accuracy about what words in other languages mean just based on how they make you feel. And that they should pay attention to however they feel about the pictograph to make their best guess about what the word means.

So within each block people got four different kinds of emotional primes — some were fear-inducing, some were disgust-inducing, some were happy and some were sexy. These are fairly representative examples of all of these, and so all four primes are presented randomly intermixed within each block, and then one specific judgment is made on each block.

So what I'm showing you here is the proportion of emotion-consistent judgments in each of the prime conditions, so the highlighted items here are ones that are significantly different from anything else in that row. For example, when primed with disgust, people were much more likely to say that the pictograph was in fact related to disgust than when primed with any of the other primes. When people were primed with fear, they were more likely to say that the pictograph was related to fear. When primed with desire, they were more likely to say that the pictograph meant something about sex. When they were primed with a generally happy prime, they were more likely to say that the pictograph was related to something meaning happy.

Now, you do have some valence effect here. So, for example, a fear prime increases ratings of disgust somewhat compared to neutral or happy primes, but over and above that you see specificity, because this is greater than that and this is greater than that and so forth, and so you have both sort of a general valence component here, and also emotion specificity.

So the reason for doing that first study was really just to see if the measure was sensitive to specific emotions, so that we could then harness that for predicting intergroup emotions in an interesting way. Here

we drew on intergroup image theory — Michelle Alexander and Marilyn Brewer and colleagues, and I should say that this work is actually a master's thesis of Nathan Arbuckle, who is a graduate student at Ohio State.

The background of this project according to intergroup image theory, the kinds of stereotypes that people have about different social groups depend on the kinds of appraisals that they make about the relative status of group, the relative power and goal compatibility — by goal compatibility I mean whether they're cooperative with you, or whether they're competitive towards you or hostile towards you. Those appraisals give rise to specific intergroup emotions, which then shape the kinds of stereotypic images that we have about out-group members. So, for example, if you fear a particular group, then that leads to this theory that's called the "barbarian image". That's a little bit funny terminology, but you can think of that as sort of like a common criminal — so low status, low power, but dangerous to you. Whereas anger gives rise to the enemy image — equal status, equal power, but still dangerous to you.

So Alexander, Brewer and colleagues have developed measures of these different images, so we used their questionnaires to measure the extent to which they had these different kinds of images specifically about African Americans. Now, if you look at the average here, I should say that we used black and white faces as primes, just like in the previous racial attitudes. Here we had them make specific emotion judgments like in the emotion study. So on average, black is compared to white faces primed fear, and they also primed anger. It's not a whopping effect, but these are both reliable, and there was a moderate inter-correlation between the two. So for people who had fear responses to blacks, they also tended to have anger responses to blacks, but that correlation is not so large that they're anywhere close to redundant. There are a lot of people who had mainly one kind of reaction or the other.

If you look at the individual differences, and how those relate to people's stereotypic images of African Americans, you see a unique pattern of significant correlations here so that fear is specifically related to that barbarian image, whereas anger is specifically related to the enemy image, just as intergroup image theory would predict. These patterns of relations have been shown before using explicit measures of intergroup emotions. We show the same patterns here using an implicit measure, and I should say that these patterns hold up even when you control for explicit judgments made in the same way.

So what I really tried to argue is that this approach effectively measures attitudes. It resists motivational pressures to self-present, as a good implicit measure ought to do. Now, when there is no pressure to hide or conceal or alter your attitudes, there's a very strong correlation between the AMP and explicit measures. But as that motivation goes up, the correlation goes down vanishing to zero. One thing that's unique about this measure, as far as I know, I think that this is the first implicit measure that's been used to separate specific emotional reactions within a particular valence and show that that can predict specific intergroup emotions

and stereotypes. Again, it's got some practical advantages like high reliability, easy and fast to administer and so forth.

The kinds of questions that I think this is particularly useful in looking at in terms of asking questions about elections, one is that it allows a very direct kind of comparison between implicit and explicit measures. So when the prime is a group label, and I didn't tell you about any of these studies, but we have some looking at when you'd simply flash the word "African Americans" or "white Americans" as primes, you can manipulate whether somebody is told, "Evaluate the pictograph, and whatever you do don't be influenced by the prime" in one condition, and in another condition, "Evaluate whatever the prime says, and whatever you do, don't be influenced by the pictograph." What you have there is a comparison of what's essentially a feeling thermometer. "What are your feelings towards African Americans?" And an indirect version of a feeling thermometer, "What are your feelings toward this pictograph?"

We find that the closer you align implicit and explicit measures in terms of these structural dimensions, the higher the implicit/explicit correlation goes. It would be very interesting to see since the feeling thermometer has been included for years in the National Election Studies, how this indirect versus direct comparison looks in the context of those historical levels.

The specificity of the task is also interesting for answering questions about what kinds of specific emotional reactions to candidates, to issue policies and so forth predict motor behavior over and above a general evaluation and one's specifics emotions towards social groups predict policy preferences like affirmative action or support for the Iraq War and so forth. So I think that I'm about out of time and I'll stop there. I'll thank my collaborators, and I'll thank you all. Yes, Elliott?

Q. *We did these same two studies, and I know that you know this, but just for everyone else here, we also replicated what you just reported with pictures that are prerated to elicit particular emotions. That worked beautifully. We were not able, as you were, to get the specific emotional reactions to black faces to work. We were doing a slightly different thing than you, though. We were simply trying to get the AMP responses on fear and anger to black faces to correlate with explicit self-reports of how afraid or how angry do I feel about blacks. It was a slightly different task than yours, but that did not work for us.*

Okay, now did you ask in the explicit question, "What are your emotions towards blacks?" or did you show them the same items and say, "How do you feel about this face?"

Q. *We asked about blacks in general.*

I think that Nathan also has that data in there. I think that he found that the correlation between implicit responses and those explicit questions were fairly low whenever we asked it in a general way, like you're describing. So I think that there were significant correlations in there, .25, .3, but it was no whopping correlation. Where we found large correlations is when we showed them the exact same items, and asked them to directly rate, "How do you feel about this black person?" Now, on the one hand you might say, "We're really interested in attitudes towards the group as a whole," but then that sort of begs the question of why when it comes to a laboratory task you would show faces to somebody and think that at the implicit level that that's answering your question. So I think that the match between what you ask people to evaluate is really important in when you'll find those implicit and explicit correlations and when you won't. Jack?

Q. *When you asked them to make the judgments of the individual faces, do you instruct them? Do you have a condition where you instruct them to try not to let racial bias influence those judgments? Otherwise, it's kind of an indirect measure in and of itself.*

Right. We have studies in which we manipulated things like social pressure, because even when you explicitly rate feelings towards these faces, it's not that explicit of a measure. They may not know that this is getting at racial attitudes or something like that. What we find is a relationship very much like the two graphs that I showed you with motivation and control prejudice. Whenever we make people feel safe and secure in expressing their attitudes confidentially, we find a big, strong correlation. But whenever we tell people that this study is about eliminating the scourge of racial prejudice, all of a sudden they tell us that they have wonderful positive attitudes towards the black faces, but it doesn't affect those indirect ratings via the pictographs at all.

Q. *Can you tell us about the relationship of the AMP to other implicit measures, like IAT or priming, and then what do you make of those correlations on (unintelligible) or whatever it happens to be?*

Yes. Good question. I've only done one study that actually includes the AMP as well as the IAT and evaluative priming. Now, that was on the study of attitudes towards alcohol. Now, what that found is essentially no correlation between the three implicit measures. We know that there are no or low correlations that have been found before between priming and the IAT, or at least unless you don't control for measurement error and so forth. Now, the three measures did, interestingly, have some unique, predictive value in predicting how much people drank. I don't necessarily think that no correlation among measures is going to be general. I think that attitudes towards alcohol among a sample of 18-year-olds is a little bit like attitudes towards race — that is that people vary in the extent to which they feel the need to self-present on that one.

I think that if you're going to look at attitudes towards political candidates, for example, I think that it would be so highly collinear that they would look like the same thing. We know the IAT has a super high correlation with explicit judgments of political candidates, and the AMP has a super high correlation with the same thing. So it's probably going to vary with the domain, and that might be related to the complexity and multidimensionality of the attitudes. If each one is picking up on something that's slightly different or partly different, the more complex the attitude becomes — the more different things that you have to pick up on and predict.

Q. *So if they are uncorrelated, do you have some thoughts about what each one is picking up that might be different? Like is one more affective or what? I know that you're going to have to speculate, but—*

Right. Well, as we've done it here, I think that the AMP is certainly very affective. I mean we're asking them to respond about their feelings towards the pictograph or the pleasantness and so forth. There is a large literature developing out there on what the IAT is tapping, and it seems to be complex enough to tap several different things. I don't know that I would say that the IAT is not affective. There is some data to suggest that it correlates with affective judgments better than more cognitive judgments, but I guess that I would say that it's going to depend with respect to the AMP what you asked participants to judge about those pictographs. So we have some studies where we look at stereotypes, for example, rather than affect. We tell them to judge whether the pictograph means aggressive or friendly, so we're picking up on stereotypes there that correlate with other measures of stereotypes, but not particularly well with judgments of affect. So it's kind of wheedling out of your question, but it's going to depend on what exactly was asked of participants to do. I think that's going to be the ultimate answer.

Q. *So in the area of candidate evaluation, I think (inaudible) implicit measures give us the same answer as the explicit evaluation, and that is if we're talking just about candidate evaluation. One of the points that you're trying to make here is that your implicit measure can also pick up specific emotional reactions. Is there any evidence at all that implicit emotional reactions to candidates are any different than explicit emotional reactions?*

That's a great question. I don't have any data to that effect, but I can certainly image situations, and I'm just speculating here, but in which people might not be willing to admit certain emotions. If President Bush makes you feel proud, people might say that. But if President Bush makes you feel dominant, people might not be willing to say that. I can imagine situations in which the specific emotions that people report might differ from the ones that we would pick up here.

Q. *Following up on that and it's maybe pretty obvious, but to the extent that the candidates become more diverse, that perhaps then the questions — the issues overlap. So when the candidate is an African American or a woman, then all of a sudden you may find that candidate evaluations actually don't correlate. Have you done anything that would allow you to draw any conclusions about nontraditional candidates and whether you get high correlations between explicit and implicit measures?*

I don't have any data speaking specifically to that question. It's a great one, though, and it goes back to the question of partisanship as a matter of changing over time. In this partisan atmosphere it seems like everything lines up neatly on a bipolar line because everybody is very polarized, whereas in times past, or in times future, it may be much more multidimensional rather than the candidates lining up in stereotyped ways. That's a good question.

Q. *—(Unintelligible) Have you used AMP with black and white faces to distinguish disgust, so that it would appear anger but also just anti-black disgust?*

Yes, one of the studies that Nathan had in his master's thesis actually included a disgust judgment. If I recall right, I think that on average he did find an affect such that black faces primed disgust to a greater extent than white faces, but it was quite small. It didn't correlate with any of these stereotyped ratings, and that might be for any of a number of reasons. It might be that people generally don't tend to have that emotional reaction in the racial domain, but it was also sort of highly correlated with the fear and anger type of judgments. So it may be that disgust has multiple meanings. Disgust for food is different than the type of disgust that you might have for a person, and so that's a complex one.

Q. *This is just a hypothesis, and it's sort of consistent, I think, with what Wendy was saying is that I think that you need a big data set. Like all of these things seem to say implicit and explicit measures of candidate preference line up. What if you picked out people who voted against their party affiliation? Would those people—would the switchers be better predicted by an implicit measure? You know, a liberal doesn't want to say that he voted for Bush or a Republican in 2002, but maybe they did. And a conservative who's disgusted today may not want to say that they're going to vote for a Democrat in the next election, but maybe they're going to and the implicit measure might pick those people up.*

Yes, that's a great question. It might be either because they are lying to you now or it may not, because it's just sort of an honesty, self-presentation sort of thing. It may be that people who have this conflict between their implicit and explicit reactions, as I think that Bruce is probably going to talk about — may have greater doubt or less certainty about those attitudes, and therefore are likely to be labile in their behaviors in the same way that we know that ambivalence in lots of ways leads to very labile behavioral predispositions.

Q. *Yes, it needs something like close to an actual voting behavior.*

Yes, that's a good point. I think that I'll draw us to a close at this point. *(End of Payne_1 file)*