OLSONZALTMANASSOCIATES

A Marketer's Guide to Implicit Measures: What to Know and What to Ask

EXAMINING CURRENT BIOMETRIC TECHNIQUES IN RESEARCH

by Dr. Andrew Baron
Consumer Neuroscience Advisor, Olson Zaltman Associates
Assistant Professor of Psychology and Affiliate, National Core for Neuroethics, at the University of British Columbia
July 2014

With millions of dollars in play, research firms are striving to distinguish themselves. Realizing a recent market trend toward including more quantitative measures of unconscious thinking that drive consumer behavior, many companies are developing their own proprietary tools promising to reveal the deep thoughts and feelings individuals harbor outside conscious awareness. Companies are also employing and adapting tools from the behavioral sciences to measure the unconscious mind. The aim of this paper is to introduce you to some of the important questions worth thinking about when deciding which tool to select for your marketing needs.

We frequently see variants of the Implicit Association Test (IAT; Greenwald et al., 1998), the Affect Misattribution Procedure (AMP; Payne et al., 2005), Semantic Priming (SP; Meyer & Schvaneveldt, 1971), as well as proprietary measures developed internally by independent research firms and applied to a variety of marketing issues. Collectively, these tools are called response latency tests (RLTs) because they measure responses in milliseconds using a computer. In marketing, they're often referred to as Implicit Tests because they are thought to capture unconscious thoughts and feelings.

The crucial question any client should ask is whether the particular tool or even the particular version of a tool matters when deciding how to address your marketing questions.

The position in this article is to state emphatically: *It matters. A lot.* In fact, it is incredibly easy to compromise the validity of these tools (the ability to measure what you want), especially when you employ methods that have not been proven in the academic literature. We outline below what steps you can take to ensure you're selecting the right tool.

"Implicit tests are not interchangeable with one another."

Implicit tests are not interchangeable with one another – they differ in strengths and weaknesses. And, the way the method is implemented and the data are analyzed means the difference between arriving at conclusions you can trust versus those you could have flipped a coin to reach. Thus, it is essential to understand how to differentiate among the various RLT offerings – what is it good at measuring, what is it bad at measuring and, most importantly, when does it fail to reliably measure anything? Below, the important issues you need to consider when selecting a tool are outlined. In a forthcoming article, I will discuss specific strengths and weaknesses in the academic literature among several popular measures of implicit tests (e.g., IAT, AMP, Priming) and their fit with different marketing questions.

When selecting a tool to address any research question, it is essential to consider two core issues. **First,** to what extent has a tool (and the corresponding analytic strategy for data generated by that tool) been supported by academic research? Put simply, does the tool measure what it claims to measure? **Second**, how much about the process of collecting and analyzing data is considered proprietary? These two issues are certainly intertwined and below I'll unpack their unique importance.

The importance of academically vetted methods: The behavioral sciences could not be more concerned with ensuring the validity of the tools employed by research scientists. In fact, each time an empirical article is submitted for publication, the authors must convince the editors that their methodology and analytic strategies are well supported by previous research. If the method or analytic strategy is novel, then the authors must go to considerable lengths to demonstrate the soundness of both. This usually requires extensive reliability testing (e.g., measuring the same constructs at two different points in time among the same individuals to demonstrate consistency in what is being measured). Statistical benchmarks exist for what is considered weak, good and great reliability and if there isn't great reliability then that new method won't become widely adopted in the field and will likely only be published in a low-tier journal.

The researchers must also demonstrate the internal consistency of a method (ensuring that the association strengths measured aren't unduly influenced by the first few or last few trials, but that all the trials provide converging information). The issue of internal consistency is critical because it bears on the total number of trials required to reliably capture a particular thought or feeling. This involves comparing the first half and the second half of the data collected for a particular measure and ensuring that it passes acceptable statistical levels of consistency. If a measure has low internal consistency, then it means it is not really measuring what it is supposed to measure. For example, let's say I am using the IAT to measure unconscious thoughts about my favorite brand of toothpaste. On many trials my responses are governed by those unconscious thoughts but sometimes my responses might also be influenced by extraneous things like my dominant (left) hand (leading me to make more errors sometimes by pressing my left hand too frequently) and sometimes my responses might be influenced by a distraction (phone ringing) or something to do with a particular stimulus presented on the screen (the word refreshing begins with the letter R and I keep thinking to respond with my *right* hand). When a measure is low in internal consistency then these various extraneous factors affecting responses unduly influence our measurement of implicit associations. A measure high in internal consistency means that there is a minimal amount of noise (unexpected influences on responding) affecting our ability to measure the desired associations.

"Even if you select a tool that has met high standards for reliability and validity, there are still several ways in which the implementation of that method might lead to misleading data."

When a method is modified from how it is traditionally used in academic research to accommodate research budgets or to seem more competitive (by reducing the number of trials or increasing the number of associations measured), the internal consistency and reliability of the method must be demonstrated anew. As but one example, with the IAT, published research has identified a critical number of trials to achieve appropriate levels of reliability and internal consistency and has spoken to the limits of how many associations can be reliably measured in one sitting by a participant. **Deviating from these standards means you are no longer accurately measuring a person's implicit associations.** Data cannot be trusted when you make

modifications to an established method unless research has shown that with those modifications the tool still meets the same criteria for reliability and validity.

Predictive validity (how we know a method is measuring something real that influences behavior) is another critical component of any method. It is not enough to ask people to respond quickly on a computer for something to constitute an implicit measure of some particular thought or feeling. To be valid, a measure of any thought or feeling must predict sensible things (e.g., specific behaviors, other thoughts and feelings). For example, if I have three tasks that I call measures of implicit preference and none of them are correlated with one another, then I would have reason to suspect that one or more of these tasks isn't measuring what I think it is (e.g., Cunningham et al., 2004). However, if one of these measures more strongly predicts a behavior (e.g., choice) determined by regression analysis (or some other appropriate statistical analysis), then this would mean it has better predictive validity than the other tools and therefore is doing a better job of capturing an aspect of implicit preference related to decision making (e.g., Greenwald et al., 2010).

It is important to note that even if you select a tool that has met high standards for reliability and validity, there are still several ways in which the implementation of that method might lead to misleading data. Research by Mitchell et al. (2003) has underscored how implicit measures are sensitive to the specific language used to represent the concepts measured. In their study, they examined an implicit preference for one race over another using the IAT. They asked participants to categorize dark-skinned faces and light-skinned faces into one of two categories. In Condition 1, the categories were labeled African-American and European-American. In Condition 2, the categories were labeled Athlete and Politician. Again, all the face stimuli were the same across conditions. Depending on which labels were selected, participants showed a strong preference for light-skinned faces (when the labels referred to race) or a strong preference for dark-skinned faces (when the labels referred to occupation). In a marketing context, this means that when selecting the appropriate labels and stimuli for your study, it is crucial that you leverage deep insights that accurately reflect the meanings you want participants (and eventually consumers) to co-create.

"It is crucial that you leverage deep insights that accurately reflect the meanings you want consumers to co-create."

Lastly, when employing RLTs there are a variety of ways to analyze the data (e.g., Signal Detection Theory, examine error rates and use chi-square analyses, directly compare latency differences, compare effect sizes, use structural equation modeling, QUAD modeling, or repeated-measures ANOVAs). Some are appropriate while others will lead you to false conclusions resulting in financially costly decisions (Greenwald et al., 2003). One strategy could tell you that nothing is significant. Another might tell you that Option A is preferred while yet another technique might reveal a completely different conclusion. This is precisely why it is not acceptable to employ just any analytic strategy and why it is essential that there be transparency in which analytic strategy a company utilizes.

The importance of transparency: It is incredibly easy for clients to be misled about the validity of a measurement tool. Here's why: People are susceptible to believing individuals when they think those people have more knowledge than them (e.g., professional experience, academic degrees). This phenomenon is similar to what we see with the "seductive allure of neuroscience explanations" – a phenomenon made very salient by research demonstrating that people are more apt to believe a statement when it is accompanied by an image of the brain than when the statement appears by itself - even when the brain image adds no explanatory value (Weisberg, et al., 2008; McCabe & Castell, 2008).

We have a natural bias to think that if someone sounds confident, appears to have more expertise or can say that they used some form of neuroscience in their work, then what they say must be true. This is why many clients are at great risk for being taken advantage of with their limited research budgets.

The risk with opacity should be obvious – if a company discovers that with a conventional approach there is no clear answer for their client, then they might fear that the client will not return with future business. The company may then decide to cherry pick some other analytic strategy that happens to give them a quantitative difference to report. There is even a name for this approach in the field – it is called p-hacking or significance hacking and many major academic journals now safeguard against this. The problem if this occurs is that any decisions based on these reports are completely meaningless. For each RLT in the field, there are optimal analytic strategies (e.g., for IATs they involve either calculating D-scores or using QUAD models). In many cases, when researchers conducted the reliability and validity studies described above they were simultaneously establishing which analytic strategy is the most robust in ensuring the reliability and validity of the tool (e.g., Greenwald et al., 2003). Thus, there should be skepticism when companies create their own analytic strategies that bypass the procedures that passed peer-review and have been published in quality journals.

Understand the risks associated with proprietary methods to measure implicit thoughts. If a novel tool was truly superior to an existing, academically validated tool, it either would have been discovered already or have been quickly adopted by the academic community. There are thousands upon thousands of social and cognitive psychologists globally, all vying to set themselves apart from their peers and to find new and, most importantly, **better** ways of measuring phenomena in their field. This contrasts with the few dozen very competitive behavioral neuroscience infused market research firms around the globe. Personally, I would trust an independent group of thousands who are not financially biased in their choices **and**, critically, whose work is under constant scrutiny from the peer-review process to establish the state-of-the-art in measurement techniques.

Proprietary tools are often designed to have a lot of sex appeal. Lots of animations and cool colors is almost naturally appealing, even entertaining. However, these features can often be unwanted distractions. Many validated tools from the behavioral sciences are often designed intentionally to avoid such frills because such they can actually interfere with what you are trying to measure. Let's consider an example. If I wanted to measure

consumer liking of a brand and I am using a tool that has been made to feel fun and cool, like a computer game with captivating animations and sounds, then I am likely activating a positive mindset, which will have the undesired effect of skewing the responses to be appear more favorable toward the brand than they truly are. Consistent with this view, there is the well-documented effect of misattribution of arousal (e.g., Dutton & Aron, 1974) whereby if you induce a certain emotional state in a person (e.g., positivity), they may unknowingly attribute that positivity to an unrelated object they are asked to evaluate. For example, research has shown that if you give a person a warm coffee cup to hold and ask them to make a judgment about someone else, they will rate that person more positively. And, if they are holding a cold coffee cup during the task, then they will rate that person more negatively (Williams & Bargh, 2009). This is one of the primary reasons the IAT and other well-validated implicit tools are designed to feel somewhat dry when you take one of these tests - since you're measuring associations, you want to limit activation of other unrelated associations.

SUMMARY

When correctly applied, implicit RLTs can reveal, perhaps better than any other kind of methodology, how much a person thinks or feels something. With this insight in hand, companies can often make critical choices, backed up with hard numbers, that precisely quantify how a person subconsciously thinks or feels along some dimension (e.g., like/don't like, trust/don't trust, want/don't want) concerning their brand, product, or advertisement compared to another brand, product, or advertisement.

However, not every measure administered on a computer that captures response latency is considered a valid measure of implicit thoughts and feelings. To be a true measure of implicit association, certain criteria for reliability and validity must be met and described thoroughly in peer-reviewed publications (preferably in top-tier journals).

Establishing the reliability and validity of a method is an onerous job that requires years of investigation to determine such fine-tuned details as the boundaries of what a measure is good (and not good) at measuring. For example, the AMP is better at measuring attitudes (feelings that the brand is liked) than it is at measuring implicit thoughts (beliefs that the brand is *healthy*). The IAT was traditionally used to measure relative associations (how much I like Brand A compared with Brand B), however advances in method and data analytic approaches now allow you to measure associations equally well in isolation (e.g., how much you like Brand A without a comparison to Brand B; Sriram & Greenwald, 2009; Sherman et al, 2008; Karpinski & Hilton, 2001). Research has also established which analytic strategy is more appropriate given the nature of the study design, and how well the method is able to predict certain behaviors (and thoughts and feelings). Just because a method is called an IAT, AMP, or SP, doesn't mean that any version of that tool is correctly administered. Nor does it mean that the particular analytic strategy for that tool is consistent with what the field says is the appropriate approach required to maintain sufficient reliability and validity.

This isn't a matter of degree – like one version of a tool is a little better than the other. It's truly a matter of kind – **a method is applied correctly and will measure what you want** or **its application is baseless** and one might as well make their decisions by picking an answer out of a hat. Here's what you can do to guard against this concern:

If you are considering using a vendor that offers an implicit test, I would recommend that you ask some of the following questions:

1. Is this method used substantially in academic literature?

2. If so, does your version of this tool differ in how it is used in the literature (trial numbers, number of associations measured, how the data are analyzed)? And, if yes, can you share any documentation (published papers) that speaks to the reliability and validity of these changes with this tool?

3. Can you explain why you chose the particular data analysis approach you'd employ and what significance level are you using? Sometimes vendors will just say that one association is stronger than another but if that isn't significant at an acceptably high level - 95%, 90%, 80% confidence, then the risk that you're making an incorrect assessment is greater. The client has to decide what level of confidence they're comfortable with. Field standards are typically 90% or greater (95% in academia). In practice, it is helpful to still know what is significant at varying levels (80%, 70%) as long as it is clearly marked.

4. Is there evidence that this method predicts behaviors (decision-making, etc)? There is an extensive literature with the IAT detailing behaviors it predicts. But if a vendor offers a modified IAT you want to be sure that that new version also has been shown to predict those desired behaviors.

5. Are the published papers supporting use of this method written largely by people at the company or do they come from academic sources published in peer-reviewed journals (preferably top-tier journals)? Even better, does the company work with academics who actively publish in peer-reviewed sources with these methods?

6. Who are the science experts behind the administration of the tool (design and analysis)? Advanced degrees are great, but they are certainly not enough. You want to know that these individuals are actively publishing in peer-reviewed journals using such methods as this is unbiased evidence of their expertise. And, in the same way you would not settle for a spine surgeon when you need heart surgery, academic psychologists have very specialized niches and implicit cognition is one such niche. Make sure this was the focus of their training and that they were trained by recognized experts in that field.