In D. Reitter & F. E. Ritter (Eds.), Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016). University Park, PA: Penn State.

The Minimalist Interference Model of the Implicit Association Test Predicts Working Memory Confounds

Michael Paul McDonald (mpmcd@uw.edu)
Department of Psychology, Guthrie Hall, 119A
Seattle, WA 98195 USA
Andrea Stocco (stocco@uw.edu)
Department of Psychology, Guthrie Hall, 119A
Seattle, WA 98195 USA
University of Washington

Keywords: interference, implicit, IAT, ACT-R, response competition

Introduction

The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) is an indirect measure of association between concepts (e.g. race) and attributes (e.g. pleasant/unpleasant). Subjects classify concepts by category and attributes by valence as rapidly and accurately as possible, using the same response keys for concepts and attributes. Typically one key pairing is easier accomplished than with the other. It is assumed that this facility is due to an association in the subject's mind between the concepts and attributes. For instance, subjects who perform more rapidly using the white/pleasant and black/unpleasant key mapping on a Race IAT are assumed to have a positive association with White and/or a negative association with Black. This is termed an *implicit preference* for White over Black.

However, research employing the IAT has dramatically outpaced research *on* the IAT. The method yields scores with favorable psychometric properties (Cunningham, Preacher, & Banaji, 2001; Greenwald, Nosek, & Banaji, 2003), and analyses have demonstrated predictive validity by correlating IAT scores with behavioral outcomes (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Greenwald, Banaji, & Nosek, 2015). However, fundamental questions remain about the mechanism of effect, and the degree to which scores on the IAT represent underlying associations or attitudes. Understanding the mechanism of effect will allow us to better interpret D scores generated by the IAT.

Assumptions and Limitations

The end result of the IAT is the *D score*, constructed to reflect an indirect measurement of relative association strength between the target concepts and attributes. The scores are assigned positive or negative signs to indicate the direction of association. The magnitude indicates strength of effect. In the case of a Race IAT, a D near 0 would suggest neutrality, while a substantially positive D score would indicate implicit White preference, and a substantially negative score would indicate implicit Black preference.

This interpretation rests on the assumption that the IAT effect is enabled, or at least predominately contributed to, by underlying relative associations between the concepts and attributes in the subjects' minds. That is, for a subject to have a highly positive D score on the Race IAT, they must have a greater association of the concept "White" with positive than "Black" with positive, or an equivalent negative pairing.

The IAT is perhaps the most widely completed cognitive task ever developed. Through the Project Implicit website, tens of thousands of IATs are conducted each month. Complete experimental data for millions of IATs may be used for comparison to computational model results.

Presented here is a minimalist computational model of the IAT, using associations within declarative memory as the primary source of response interference.

The Minimalist Interference Model of the IAT

The minimalist interference model of the IAT (MIMI) generates an IAT effect by constructing associative interference between chunks in declarative memory. Each stimulus chunk is directly associated with its category, and the concept and attribute categories are also associated (see Figure 1). Spreading activation differentially primes the associated attribute when classifying target stimuli, thus making retrieval of the correct category more difficult. The IAT effect produced is a function of retrieval interference, and the extremity of the resulting D score is monotonic with the magnitude of disparity between underlying associations.

The effect produced yields mean latencies, but lacking SDs and realistic human noise, a D score cannot be produced (the D score, modeled after Cohen's d, is a ratio of the mean differences to the pooled standard deviation). Response time in the compatible condition is 771ms, and 819ms in the incompatible condition, with the overall mean response time at 795ms - comparable to the mean response time observed in

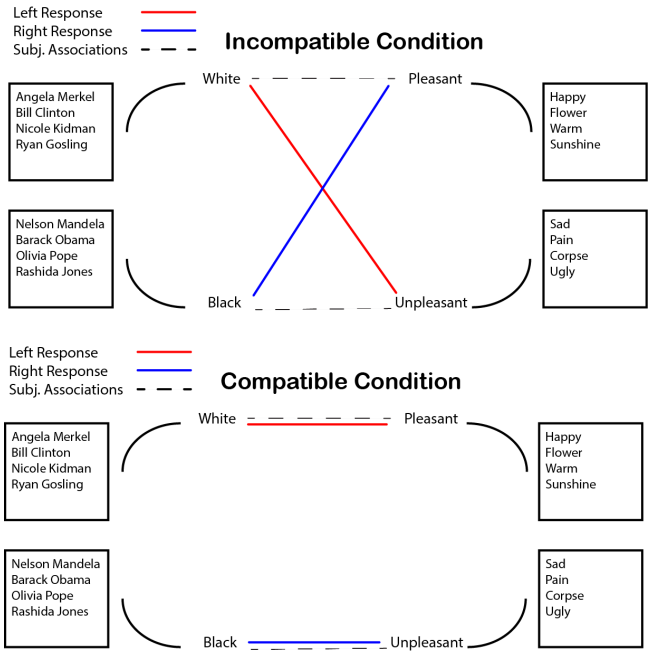

Figure 1. Diagram of associations between chunks, categories, attributes, and rules. In the compatible condition, the association created by the rules is congruent with the underlying associations in the subject's mind. In the incompatible condition, the associations created by the rules are incongruent with those in the subject's mind. This association facilitates retrieval of the rule category in the compatible condition, and detracts from performance in the incompatible condition.

the lab, about 790ms (Greenwald et al., 2003).

The modeled effect relies specifically on the underlying associations between the concepts and attributes (e.g. White and pleasant). Associations between stimuli and their parent classes are flat. The IAT effect has been shown to depend on the association between the concept category and the attributes, rather than between the exemplars and attributes (De Houwer, 2001). For example, in subjects that display automatic White preference, this preference is still evident when using uniformly negative White stimuli (e.g. Ted Bundy) and positive Black stimuli (e.g. Nelson Mandela).

When the model perceives a stimulus, it retrieves a rule with a property in common with the properties of the stimulus. This retrieval can be complicated by lingering activation from previous retrievals. For instance, if a White stimulus appears after classifying a pleasant word, the White-pleasant association may trigger a retrieval of an incorrect rule (e.g. Black-pleasant > respond with right key). Even if this doesn't cause an incorrect retrieval, the closer levels of activation make the retrieval more difficult (i.e. take longer) than in the compatible condition.

Discussion

The minimalist model approximates normal subject performance, but does not adequately explain the interference effect. MIMI assumes that the underlying strategy used by subjects is unchanged, and that differential latency between conditions is a result only of retrieval interference. While this model provides useful insight into the role of retrieval interference caused by spreading activation, it is not expected that this model accurately reflects reality. For instance, this model does not reproduce more extreme D scores in subjects that experience the compatible condition first (seen in human subjects), as the interference produced is be the same regardless of order of experience (since no learning is involved).

It is hoped that these models will lead to a more cogent understanding of the underpinning mechanisms of the IAT and other implicit interference effects. This understanding will in turn give greater insight into the proper interpretation of metrics such at the IAT's D score.

References

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001, March). Implicit Attitude Measures: Consistency, Stability, and Convergent Validity. *Psychological Science*, *12*(2), 163–170.

De Houwer, J. (2001, November). A Structural and Process Analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443–451.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Mierke, J., & Klauer, K. C. (2003). Method-Specific Variance in the Implicit Association Test. *Journal of Personality and Social Psychology*, *85*(6), 1180–1192.

von Stülpnagel, R., & Steffens, M. C. (2010). Prejudiced or Just Smart? *Zeitschrift für Psychologie / Journal of Psychology*, *218*(1), 51–53.