This is an Accepted Manuscript of an article published by Taylor & Francis in Addiction Research & Theory on 19/09/2022, available at: https://doi.org/10.1080/16066359.2022.2123474.

Towards the nature of automatic associations: Item-level computational semantic similarity and IAT-based alcohol-valence associations

Thomas E. Gladwin [a*]

a. No affiliation.

* Corresponding author. Email: thomas.gladwin@gmail.com.

Running head: Semantic similarity and alcohol associations.

Abstract

Automatic associations involving alcohol have been proposed to play a role in drinking behaviour. Such associations are often assessed using implicit measures such as the Implicit Association Test (IAT). Neural network language models provide computational measures of semantic relationships between words. These model-based measures could be related to behavioural alcohol-related associations as observed using the IAT. If so, this could provide a step towards better understanding of the nature of automatic associations and their relationship to behaviour. The current study therefore aimed to test whether there is a systematic covariation over items between model-based and behaviour-based associations. Analyses were performed for two single-target IATs from a previously published study. One task involved alcohol-versus-non-alcohol drinks and positive associates, and the other alcohol-versus-non-alcohol drinks and negative associates. The GenSim library and a pretrained word2vec model were used to calculate a relative computational association between specific items from the positive and negative categories, respectively, and the alcohol versus non-alcohol word sets. In both tasks, a significant covariance between items' computational and behavioural measures of association was found over participants. The results thus add to the information on the relationship between neural network language models and psychological associations. They may provide methodological strategies for task design and data analysis. Models of semantic associations connect computational linguistics and social-cognitive psychology and may provide a theoretical link between measures of alcohol-related associations using verbal stimuli and alcohol-related cognition and behaviours.

Keywords: IAT; machine learning; item-level; word2vec; alcohol-valence.

Natural language processing involves machine learning methods that generate a representation of the meaning of words. The neural network language model (NNLM) approach to this, such as the word2vec algorithm, trains a neural network to predict the surrounding words of a given word (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). This produces a hidden layer of, e.g., 300 nodes, the weights to which can be used as a vector that represents this kind of word meaning, and that can be used to connect vector calculus to meaning. For instance, using these vectors, "monarch" plus "female" minus "male" does indeed produce a vector close to that for "queen", while "monarch" plus "male" minus "female" is closer to "king". This approach not only provides an interesting computationally defined conceptual framework for meaning, but also a range of methods of potential interest to psychological questions, leveraging textual data with emotional or psychiatric content (Min et al., 2021; Tausczik & Pennebaker, 2010; Yin et al., 2019). However, it remains to be determined to what extent, and in which ways, such computational models reflect actual human cognition. In particular, do they have the potential to better understand the nature of automatic associations such as those involved in dual process models of addiction (McClure & Bickel, 2014; Stacy & Wiers, 2010)?

For this to be the case, it is necessary to show that computational similarities between semantic vectors are related to psychological associations as measured via behavioural effects. One approach to such measurement is the Implicit Association Test, IAT (De Houwer et al., 2009a; Greenwald et al., 1998). This is a reaction time task in which participants categorize stimuli. Critically, multiple categories are assigned to the same response. For instance, the task could involve using a left-hand response button that represents the categories "insects" and "dirty" and a right-hand response button that represents "flowers" and "pretty". This mapping of categories to responses would be considered congruent, as associated categories are mapped to the same response. The incongruent mapping would have "insects" and "pretty" mapped to one response button, and "flowers" and "dirty" mapped to the other. The IAT presents blocks alternating between such congruent and incongruent mappings, and incongruent mappings result in slower reaction times and higher error

rates than congruent mappings. We briefly note that although the validity of the IAT, in particular as a measure of individual differences, is debated (Blanton et al., 2009; De Houwer et al., 2009b), basic congruence effects such as those involving insects versus flowers are strong and replicable (Gladwin et al., 2012; Greenwald et al., 1998); the current study concerns widespread, population-level associations and general psychological processes rather than individual differences. However, previous research has shown relationships between what the IAT measures and self-reported behaviour related to alcohol use. Risky drinking is associated with stronger associations between alcohol and approach (Ostafin & Palfai, 2006; Palfai & Ostafin, 2003) and weaker associations between alcohol and negative concepts (Houben et al., 2009, 2010; Wiers et al., 2002). Individual differences between IAT measures and attentional bias have been found, stronger alcohol-negative associations being associated with attentional avoidance of alcohol stimuli (Gladwin & Vink, 2018).

Previous research has already shown connections between computational semantic similarity and behavioural associations. Computational similarities between words represented as word-embeddings, or vectors in language models, have been shown to mirror a range of prejudices expressed in behavioural data (Caliskan et al., 2017). This was noted to represent a risk in societal uses of machine learning, but also demonstrates that such meaning representations reflect psychological patterns. Going a step further, a covariation was demonstrated, over studies, between linguistic similarities between items based on co-occurrence of words in a large corpus of textual data and behavioural effects in lexical IATs (Lynott et al., 2012). It was noted that multiple causal interpretations of such findings are currently possible. Nevertheless, if a strong empirical foundation can be established that shows that behavioural data can be explained by semantic models, then connecting computational models to implicit measures could progress our theoretical understanding of the nature of “associations” in implicit biases.

The current study aims to continue this line of inquiry in the alcohol context, by testing at the item level whether alcohol-related associations measured via the IAT are related to variations in NNLM-

based semantic similarities involving specific words in IATs. The data used for this concerned IAT-derived alcohol-valence associations. This particular dataset involved two variants of the IAT, one of which assessed alcohol-related associations with positive stimuli, e.g., “Social”, “Exciting”, and “Acceptance”, and one with negative stimuli, e.g., “Violent”, “Disapproval”, and “Failure” (Gladwin & Vink, 2018). The hypothesis was that item-specific congruence effects on behavioural performance measures would be associated with the computed relative similarity of the valence items with the sets of alcohol versus non-alcohol stimuli.

Methods

The data and source code used in the analyses are available at https://figshare.com/s/835d5be0307f69fbe7c4.

Psychological data and score calculation

The behavioural data came from a previously published study, where task details and results are available (Gladwin & Vink, 2018). The data that could be used for the current analyses consisted of 75 participants, one of which was rejected in quality checks for extremely low accuracy, leaving 74 participants. We briefly repeat the task description for convenience here. A variant of the IAT was used in which there was one contrasting pair of categories - alcoholic versus non-alcoholic - but a single valence category. This version of the IAT has the form of a “single-target” IAT, STIAT (Dickson et al., 2013), where in this case the “target” is the valence. Two different STIATs were used: one had the valence category “positive”, the other “negative”; in both cases, the stimuli were chosen to be theoretically associated with alcohol. Positive words were: “Confident”, “Social”, “Exciting”, “Relaxing”, “Acceptance”, “Worthwhile”, and “Success”. Negative words were “Dangerous”, “Violent”, “Boring”, “Disgusting”, “Disapproval”, “Hangover”, and “Failure”. In both task versions, alcohol words were the beverages (familiar to the study population) “Beer”, “Wine”, “Heineken”, “Amstel”, “Grolsch”, “Whiskey”, and “Gin”; and non-alcohol words were the beverages “Juice”, “Tea”, “Coffee”, “Water”, “Cassis”, “Milk”, and “Cola”. (In the original study, the separate STIATS

were used to allow the assessment of potential simultaneous positive and negative associations with alcohol versus non-alcohol.) Critically, blocks in the STIAT vary in the mapping of categories to response buttons. Alcohol and non-alcohol stimuli always required responses using opposite buttons. On congruent blocks, the valence category was mapped to the alcohol key; on incongruent blocks, the valence category was mapped to the non-alcohol key. Congruence thus refers to the assumed association between alcoholic, relative to non-alcoholic, drinks and the positive or negative concepts.

The behavioural measures of association between alcohol-versus-non-alcohol and the respective valences were (1) the difference in reaction time (RT) on accurate trials and (2) the difference in accuracy, between congruent and incongruent blocks when responding to valence words. That is, e.g., if there is an automatic association between alcohol and positive, participants should be faster and more accurate to respond to the word “Confident” when it is mapped to the Alcohol response button than when it is mapped to the Non-alcohol response button. Behavioural performance measures were derived only from valence words as these were always associated with a shared response button, whereas the change in mapping from congruent to incongruent blocks for alcohol and non-alcohol words was confounded with a shift from being the only category mapped to one button to sharing a response button. For the current study, for each valence word the difference, for reaction time and accuracy, between responses in congruent and incongruent blocks were calculated, per participant. The median reaction time over trials and the mean accuracy over trials were used. A negative score for RT means that responses were faster in the congruent block and a positive score for accuracy means that responses were more accurate in the congruent block.

Modelling of semantic associations

The basis of the quantification of semantic associations was the word2vec algorithm (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). This algorithm uses the interrelationship between words in a large corpus of sentences to represent words as a 300-dimensional vector. In the current

analyses, the GenSim library was used (Řehůřek & Sojka, 2010), with word vectors based on the Google News corpus, GoogleNews-vectors-negative300. The cosine similarity between word2vec vectors was used a measure of similarity between a given pair of words.

These similarities were used to calculate an alcohol-association score between each individual valence word and the sets of words of the alcohol and non-alcohol stimulus sets. For a given valence word (e.g., “Exciting”), the score was calculated as its mean similarity with the set of alcohol words minus its mean similarity with the set of non-alcohol words. The similarity scores were, for the positive words: "confident": .0082, "social": -.0082, "exciting": .028, "relaxing": .013, "acceptance": -.0051, "worthwhile": .0028, "success": .0058; and for the negative words: "dangerous": -.030, "violent": .0078, "boring": -.030, "disgusting": -.0067, "disapproval": -.025, "hangover": .033, "failure": -.046.

Preprocessing

The z-score of all reaction time difference scores, over all participants and all words, was calculated. Reaction times with an absolute z-score over four were considered outliers and removed. It was also checked, per participant, whether at least three words within a task and block type had a valid RT difference score (i.e., at least one accurate trial, that was not an outlier) and whether the overall accuracy was at least .5. The resulted in one participant being rejected. We acknowledge that this procedure differs from that often used in traditional IAT analyses.

Statistical tests of relationships between behavioural and computational associations

The statistical tests were performed for the positive and the negative task variants separately, and for RT and accuracy separately. Per task and for RT and accuracy, within each participant, the covariance between the valence words’ behavioural and computational association score was calculated as that participant’s behavioural-computational association score. A one-sample t-test was performed to determine whether the mean of these per-subject covariances was significantly

non-zero. That is: is there, over all participants, a consistent direction of a measure of their within-subject association between their behavioural and computational association scores?

Results

The mean reaction times were, for the positive task, 648 (SD = 130) ms on incongruent blocks and 668 (SD = 148) ms on congruent blocks; and for the negative task, 681 (SD = 115) ms on incongruent blocks and 654 (SD = 112) ms on congruent blocks. The mean accuracies were, for the positive task, .91 (SD = .096) on incongruent blocks and .85 (SD = .12) on congruent blocks; and for the negative task, .86 (SD .12) on incongruent blocks and .91 (.093) on congruent blocks. For the positive task, a within-subject effect of block type indicating a reversed association between alcohol and positive concepts was found for accuracy ($t(73) = -4.09$, $p = .0001$), but no effect was found for RT ($t(73) = 0.51$, $p = .61$). For the negative task, a within-subject effect of block type indicating an association between alcohol and negative concepts was found for accuracy ($t(73) = 2.74$, $p = .0077$) and RT ($t(73) = -2.51$, $p = .01$). We briefly note that in the prior findings (Gladwin & Vink, 2018), a number of associations were found between the behavioural data and questionnaire scales concerning risky drinking, social drinking motives, and enhancement drinking motives, as well as with attentional bias measures.

The relationships between computational and behavioural associations scores are shown in Figure 1. For the alcohol-positive task, there was a significant association between computational and behavioural association scores for RT (mean = -0.29, $t(73) = -2.03$, $p = .046$): that is, items with stronger computational associations with alcohol had stronger congruence effects, i.e., faster reaction times on congruent versus incongruent blocks. There was no association for accuracy (mean = -0.00, $t(73) = -1.46$, $p = .15$).

For the alcohol-negative task, there was a significant association for RT (mean = -1.22, $t(73) = -4.19$, $p < .001$), in the expected direction. The association for accuracy was also significant for this task variant (mean = 5.32, $t(73) = 5.32$, $p < .001$), in the expected direction.

<Figure 1 around here.>

Discussion

The current study aimed to determine whether variations in computational similarity, based on a NNLM, is associated with behavioural effects of congruence versus incongruence. This was indeed found to be the case in two Implicit Association Test variants involving alcohol-related associations. Items with stronger computational associations with alcohol also had stronger behavioural congruence effects, for alcohol-positive as well as alcohol-negative associations. The results suggest some interesting directions for theory and methods.

From the perspective of the features of NNLM and the word2vec algorithm, the current results provide further evidence that such methods capture at least some aspects of the kind of psychological processing involved in implicit measures. This empirically extends knowledge on the method's ability to usefully quantify meaning to the domain of dual process models of automatic versus reflective cognitive processing (Bargh, 1994; Deutsch & Strack, 2006; Schneider & Shiffrin, 1977). The model could thus conceivably be used to support a range of applications involving "nudging" or cognitive bias modification (Gladwin et al., 2016; MacLeod, 2012). Such interventions aim to influence behaviour via inducing changes in automatic processes; for instance, approach-avoidance retraining for alcoholism has been shown to reduce the chances of relapse (Eberl et al., 2013; Rinck et al., 2018; Wiers et al., 2011), with statistical mediation via implicit measures (Eberl et al., 2013; Gladwin et al., 2015). Where interventions involve verbal stimuli, in communications or as stimuli in training paradigms, stimulus selection could leverage semantic similarities to optimize the desired associations.

Conversely, from the perspective of theories of automatic associations, the results may provide a stepping stone to more precise, computational models of associations in human cognition. While the claim here is of course not that the current results imply that the brain implements any particular specific algorithm, and the relationship between corpora-based linguistic models and cognition is complex (Wingfield & Connell, 2019), using a semantic model provisionally may generate more specific hypotheses than very general concepts of associations. That is, rather than visualizing an association as two circles connected by a line, where the circles represent words and the line represents their association, an association becomes a relatively similar pattern of weights in a neural network layer following a certain kind of training. Whether the brain could implement such training in an analogous way is a question for further computational neuroscientific research, although we note that there certainly exists promising research into neurally plausible analogues for computational neural networks (Hao et al., 2020; Jocham et al., 2011; Suri & Schultz, 1999).

Such modelling of semantic relationships also serves to emphasize the fundamental question of the relationship between task-specific processes involving verbal stimuli and putatively-related automatic processes underlying behaviour (Lynott et al., 2012). In the context of alcohol use, drinking behaviour and experiences may affect the meaning assigned to certain stimuli and the words associated with them; e.g., for a social drinker, the feature of a drink or situation being alcohol-related could imply it acquires the meaning of enjoyable, friendly, and so on. For coping drinking, the meaning of alcoholic could include a reduction in stress. Our models of such associations in meaning in this broader psychological sense could be analogous to semantic vectors, except for being acquired from learning experiences rather than a corpus of texts. The reinforcing effects of alcohol could then play a role in biasing the acquisition of alcohol-related meanings (Gladwin & Wiers, 2012; Robinson & Berridge, 1993).

It must be acknowledged and emphasized that such generalizations from the linguistic to the psychological context are speculative and the current study provides only a step towards lines of

research in such directions. However, there is a natural hypothetical causal mechanism connecting semantics and emotional states, namely appraisal (Bayliss et al., 2016; Cutuli, 2014; Hajcak & Nieuwenhuis, 2006). If meaning is biased in a certain way, this could affect the appraisal of stimuli and hence emotional responses. Further, verbal stimuli are psychologically associated with other cognitive and emotional processes (Ashley et al., 2013; Cane et al., 2009; Frings et al., 2010; Herbert & Kissler, 2010; Stanford et al., 2001), and so semantic associations could form a bridge between more general stimuli and psychological states. Finally, we note that semantic models may fit into different models of the IAT, such as the quad model (Beer et al., 2008; Conrey et al., 2005). The quad model notes that the IAT is not a process-pure task and attempts to disentangle four sub-processes, and has been used to attempt to better understand relationships between alcohol and aggression (Gladwin et al., 2018). One of the sub-processes is the automatic activation of an association. In the quad model, this is not further formalized beyond the role of the parameter in a multinomial model fitting performance data. It seems that computational models of semantic similarities could be, at least conceptually, plugged in to the model within this parameter.

The results also suggest a use of machine learning in improving implicit measures. One of the issues with implicit methods is the variation in effects from trial to trial due to which specific stimuli are used. Considering the specific stimuli used in the tasks has been found to improve psychometric properties; for instance, internal reliability of an attentional bias measure involving alcohol was improved by using personalized stimuli (Christiansen et al., 2015). Exploratory research found that only certain stimulus categories showed mediating mechanisms of a cognitive training intervention for alcoholism (Gladwin et al., 2015). At the item level of the IAT, certain words chosen to fit in a category could have stronger congruence effects than others. Using algorithms such as those used in the current study would allow researchers to use the strategy of selecting word sets based on a priori computational similarities that are more likely to result in sensitive methods able to detect effects of interest; or word sets with less strong overall similarities that could be more suitable to evoke individual differences (Hedge et al., 2017). Furthermore, this approach would allow a

principled re-analysis of existing data that could focus on subsets of stimuli predicted to be more able to evoke effects.

The current study provides an initial step in this direction but, of course, has limitations. The current results concern specific alcohol-related associations, with certain sets of positive and negative stimuli. Future research is needed to determine whether the connection between computational and behavioural measures of association generalizes. Further, the original study (Gladwin & Vink, 2018) concerned a convenience sample, which was considered appropriate to the aims but did not allow strong inferences concerning populations within which effects would or would not be expected; future research could focus on more specific populations and potential differences between them. Similarly, the corpus on which the word vectors used in the current study were based involved only one particular, although well-known, vocabulary; other corpuses could in principle generate models with different results. Finally, we did not compare different types of models: while the current results showed strong associations involving computational and behavioural similarities at item level, future research might find different models that are even better at predicting behavioural patterns.

In conclusion, the current study demonstrated relationships, at the item level, between computational measures of association, based on quantified semantic similarity, and behavioural measures of association, based on reaction time in Implicit Association Tests. The results add to the evidence that neural network language models and the word2vec algorithm captures interesting aspects of meaning, may support theoretical development around implicit associations, and provides directions for methodological improvements in implicit measures design and data (re-)analysis.

Data availability statement

The script and data that support the findings of this study are openly available in IAT_CBA at https://figshare.com/s/835d5be0307f69fbe7c4.

Declaration of Interest

The author reports no conflicts of interest.

References

Ashley, V., Honzel, N., Larsen, J., Justus, T., & Swick, D. (2013). Attentional bias for trauma-related words: Exaggerated emotional Stroop effect in Afghanistan and Iraq war veterans with PTSD. *BMC Psychiatry*, *13*, 86. https://doi.org/10.1186/1471-244X-13-86

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency and control in social cognition. In R. Wyer & T. Srull (Eds.), *Handbook of Social Cognition* (2nd ed., pp. 1–40). Erlbaum.

Bayliss, A. P., Tipper, S. P., Wakeley, J., Cowen, P. J., & Rogers, R. D. (2016). Vulnerability to depression is associated with a failure to acquire implicit social appraisals. *Cognition and Emotion*, 1–9. https://doi.org/10.1080/02699931.2016.1160869

Beer, J. S., Stallen, M., Lombardo, M., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *NeuroImage*, *43*(4), 775–783. https://doi.org/10.1016/j.neuroimage.2008.08.033

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *The Journal of Applied Psychology*, *94*(3), 567–582; discussion 583-603. https://doi.org/10.1037/a0014665

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cane, J. E., Sharma, D., & Albery, I. P. (2009). The addiction Stroop task: Examining the fast and slow effects of smoking and marijuana-related cues. *Journal of Psychopharmacology*, *23*(5), 510–519. https://doi.org/10.1177/0269881108091253

Christiansen, P., Mansfield, R., Duckworth, J., Field, M., & Jones, A. (2015). Internal reliability of the alcohol-related visual probe task is increased by utilising personalised stimuli and eye-tracking. *Drug and Alcohol Dependence*, *155*, 170–174. https://doi.org/10.1016/j.drugalcdep.2015.07.672

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Cutuli, D. (2014). Cognitive reappraisal and expressive suppression strategies role in the emotion regulation: An overview on their modulatory effects and neural correlates. *Frontiers in Systems Neuroscience*, *8*, 175. https://doi.org/10.3389/fnsys.2014.00175

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009a). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009b). Theoretical claims necessitate basic research: Reply to Gawronski, Lebel, Peters, and Banse (2009) and Nosek and Greenwald (2009). *Psychological Bulletin*, *135*(3), 377–379. https://doi.org/10.1037/a0015328

Deutsch, R., & Strack, F. (2006). TARGET ARTICLE: Duality Models in Social Psychology: From Dual Processes to Interacting Systems. *Psychological Inquiry*, *17*(3), 166–172.

Dickson, J. M., Gately, C., & Field, M. (2013). Alcohol dependent patients have weak negative rather than strong positive implicit alcohol associations. *Psychopharmacology*, *228*(4), 603–610. https://doi.org/10.1007/s00213-013-3066-0

Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2013). Approach bias modification in alcohol dependence: Do clinical effects replicate and for whom does it work best? *Developmental Cognitive Neuroscience*, *4*, 38–51. https://doi.org/10.1016/j.dcn.2012.11.002

Frings, C., Englert, J., Wentura, D., & Bermeitinger, C. (2010). Decomposing the emotional Stroop effect. *Quarterly Journal of Experimental Psychology (2006)*, *63*(1), 42–49. https://doi.org/10.1080/17470210903156594

Gladwin, T. E., den Uyl, T. E., & Wiers, R. W. (2012). Anodal tDCS of dorsolateral prefontal cortex during an Implicit Association Test. *Neuroscience Letters*, *517*(2), 86–82. https://doi.org/10.1016/j.neulet.2012.04.025

Gladwin, T. E., Rinck, M., Eberl, C., Becker, E. S., Lindenmeyer, J., & Wiers, R. W. (2015). Mediation of Cognitive Bias Modification for alcohol addiction via stimulus-specific alcohol avoidance association. *Alcoholism, Clinical and Experimental Research*, *39*(1), 101–107. https://doi.org/10.1111/acer.12602

Gladwin, T. E., Salemink, E., Garritsen, H., Noom, S., Kraaij, G., & Wiers, R. W. (2018). Is alcohol-related aggression related to automatic alcohol–power or alcohol–aggression associations? *Psychology of Violence*, *8*(2), 229–237. https://doi.org/10.1037/vio0000139

Gladwin, T. E., & Vink, M. (2018). Alcohol-related attentional bias variability and conflicting automatic associations. *Journal of Experimental Psychopathology*, *9*(2). https://doi.org/10.5127/jep.062317

Gladwin, T. E., Wiers, C. E., & Wiers, R. W. (2016). Cognitive neuroscience of cognitive retraining for addiction medicine: From mediating mechanisms to questions of efficacy. In *Progress in Brain Research. Vol. 224* (pp. 323–344). https://doi.org/10.1016/bs.pbr.2015.07.021

Gladwin, T. E., & Wiers, R. W. (2012). Alcohol-Related Effects on Automaticity due to Experimentally Manipulated Conditioning. *Alcoholism: Clinical and Experimental Research*, *36*(5), 895–899. https://doi.org/10.1111/j.1530-0277.2011.01687.x

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

Hajcak, G., & Nieuwenhuis, S. (2006). Reappraisal modulates the electrocortical response to unpleasant pictures. *Cognitive, Affective & Behavioral Neuroscience*, *6*(4), 291–297.

Hao, Y., Huang, X., Dong, M., & Xu, B. (2020). A biologically plausible supervised learning method for spiking neural networks using the symmetric STDP rule. *Neural Networks*, *121*, 387–395. https://doi.org/10.1016/j.neunet.2019.09.007

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1–21. https://doi.org/10.3758/s13428-017-0935-1

Herbert, C., & Kissler, J. (2010). Motivational priming and processing interrupt: Startle reflex modulation during shallow and deep processing of emotional words. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *76*(2), 64–71. https://doi.org/10.1016/j.ijpsycho.2010.02.004

Houben, K., Nosek, B. A., & Wiers, R. W. (2010). Seeing the forest through the trees: A comparison of different IAT variants measuring implicit alcohol associations. *Drug and Alcohol Dependence*, *106*(2–3), 204–211. https://doi.org/10.1016/j.drugalcdep.2009.08.016

Houben, K., Rothermund, K., & Wiers, R. W. (2009). Predicting alcohol use with a recoding-free variant of the Implicit Association Test. *Addictive Behaviors*, *34*(5), 487–489. https://doi.org/10.1016/j.addbeh.2008.12.012

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *The Journal of Neuroscience*, *31*(5), 1606–1613. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Lynott, D., Kansal, H., Connell, L., & O'Brien, K. (2012). Modelling the IAT: Implicit Association Test Reflects Shallow Linguistic Environment and not Deep Personal Attitudes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). https://escholarship.org/uc/item/5fj441tg

MacLeod, C. M. (2012). Cognitive bias modification procedures in the management of mental disorders. *Current Opinion in Psychiatry*, *25*(2), 114–120. https://doi.org/10.1097/YCO.0b013e32834fda4a

McClure, S. M., & Bickel, W. K. (2014). A dual-systems perspective on addiction: Contributions from neuroimaging and cognitive training. *Annals of the New York Academy of Sciences*, *1327*(1), 62–78. https://doi.org/10.1111/nyas.12561

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *ArXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546

Min, H., Peng, Y., Shoss, M., & Yang, B. (2021). Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. *Journal of Applied Psychology*, *106*(2), 214–229. https://doi.org/10.1037/apl0000886

Ostafin, B. D., & Palfai, T. P. (2006). Compelled to consume: The Implicit Association Test and automatic alcohol motivation. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *20*(3), 322–327. https://doi.org/10.1037/0893-164X.20.3.322

Palfai, T. P., & Ostafin, B. D. (2003). Alcohol-related motivational tendencies in hazardous drinkers: Assessing implicit response tendencies using the modified-IAT. *Behaviour Research and Therapy*, *41*(10), 1149–1162.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rinck, M., Wiers, R. W., Becker, E. S., & Lindenmeyer, J. (2018). Relapse prevention in abstinent alcoholics by cognitive bias modification: Clinical effects of combining approach bias modification and attention bias modification. *Journal of Consulting and Clinical Psychology*, *86*(12), 1005–1016. https://doi.org/10.1037/ccp0000321

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research. Brain Research Reviews*, *18*(3), 247–291.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Stacy, A. W., & Wiers, R. W. (2010). Implicit cognition and addiction: A tool for explaining paradoxical behavior. *Annual Review of Clinical Psychology*, *6*, 551–575. https://doi.org/10.1146/annurev.clinpsy.121208.131444

Stanford, M. S., Vasterling, J. J., Mathias, C. W., Constans, J. I., & Houston, R. J. (2001). Impact of threat relevance on P3 event-related potentials in combat-related post-traumatic stress disorder. *Psychiatry Research*, *102*(2), 125–137.

Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*(3), 871–890.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science*, *22*(4), 490–497. https://doi.org/10.1177/0956797611400615

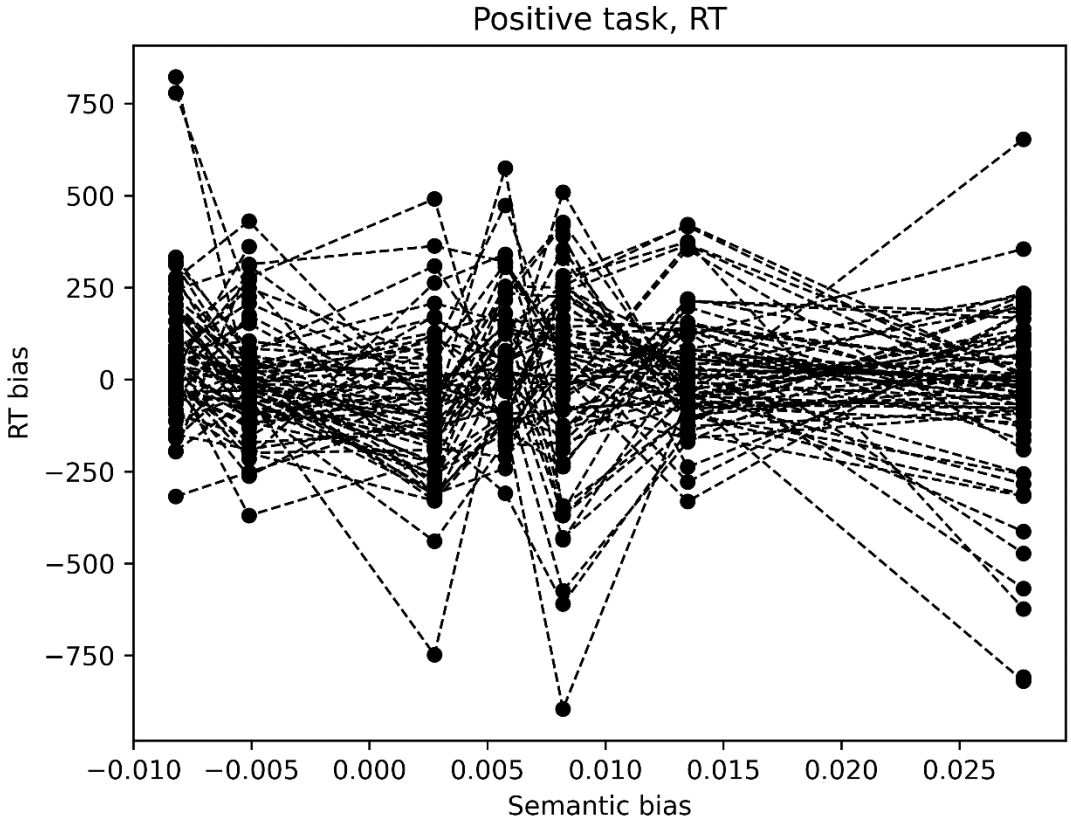
Wiers, R. W., van Woerden, N., Smulders, F. T. Y., & de Jong, P. J. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of Abnormal Psychology*, *111*(4), 648–658.

Wingfield, C., & Connell, L. (2019). *Understanding the role of linguistic distributional knowledge in cognition*. PsyArXiv. https://doi.org/10.31234/osf.io/hpm4z

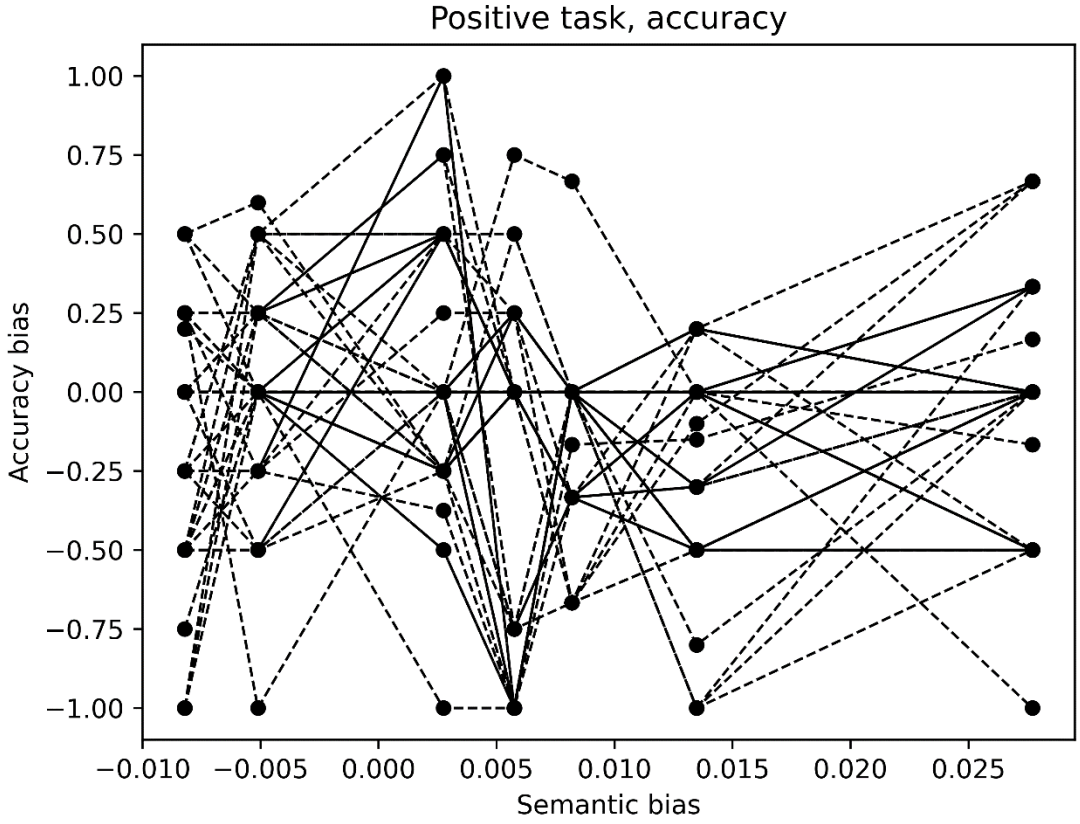
Yin, Z., Sulieman, L. M., & Malin, B. A. (2019). A systematic literature review of machine learning in online personal health data. *Journal of the American Medical Informatics Association: JAMIA*, *26*(6), 561–576. https://doi.org/10.1093/jamia/ocz009

Figure 1. Relationships between computational and behavioural association scores

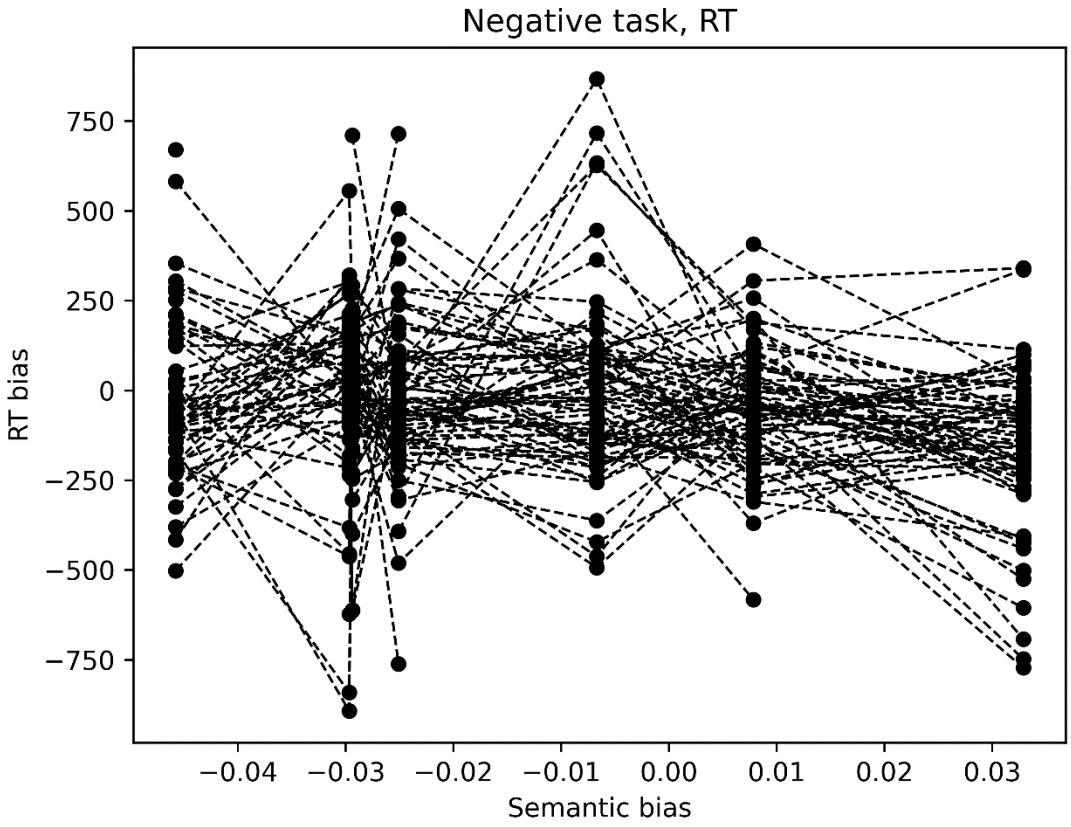
A. Positive task, RT



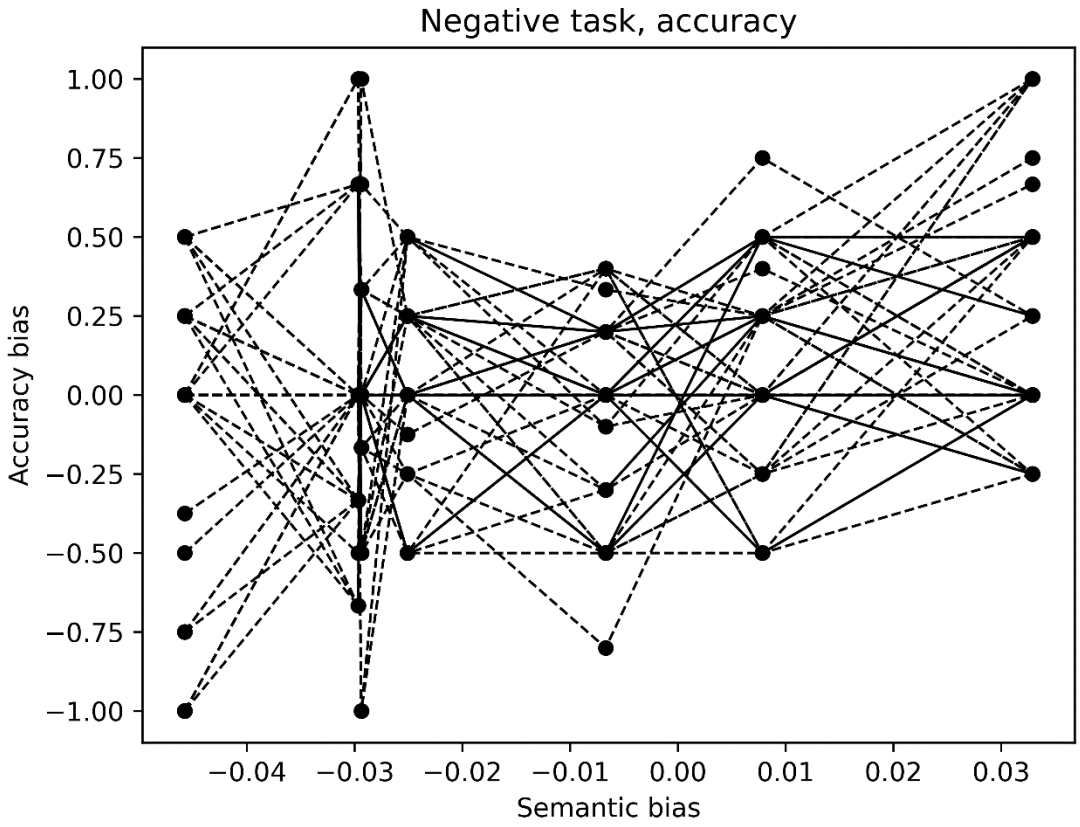
B. Positive task, accuracy



C. Negative task, RT



D. Negative task, accuracy



Note. The plots show the behavioural-computational associations for different words, for the positive and negative tasks and for RT and accuracy. Each line on the plots represents one participant; the lines connect points which represent the combinations of computational and behavioural association scores for each word. The order of words from lowest to highest computational association scores words was, on the positive task: 'social', 'acceptance', 'worthwhile', 'success', 'confident', 'relaxing', and 'exciting'; and on the negative task: 'failure', 'dangerous', 'boring', 'disapproval', 'disgusting', 'violent', and 'hangover'.