Social Cognition, Vol. 38, Supplement, 2020, pp. S26–S41

ACCURATE BY BEING NOISY: A FORMAL NETWORK MODEL OF IMPLICIT MEASURES OF ATTITUDES

Jonas Dalege
Santa Fe Institute and University of Amsterdam

Han L. J. van der Maas
University of Amsterdam

In this article, we model implicit attitude measures using our network theory of attitudes. The model rests on the assumption that implicit measures limit attitudinal entropy reduction, because implicit measures represent a measurement outcome that is the result of evaluating the attitude object in a quick and effortless manner. Implicit measures therefore assess attitudes in high entropy states (i.e., inconsistent and unstable states). In a simulation, we illustrate the implications of our network theory for implicit measures. The results of this simulation show a paradoxical result: Implicit measures can provide a more accurate assessment of conflicting evaluative reactions to an attitude object (e.g., evaluative reactions not in line with the dominant evaluative reactions) than explicit measures, *because* they assess these properties in a noisier and less reliable manner. We conclude that our network theory of attitudes increases the connection between substantive theorizing on attitudes and psychometric properties of implicit measures.

Keywords: implicit measures, indirect measures, attitudes, network theory

The study of indirect measures of attitudes (i.e., measuring attitudes without prompting individuals to reflect on their attitudes) has been one of the most active areas within social psychology in the last two decades. The appeal of these measures is that they might provide insights into implicitly measured attitudes (i.e., spontaneous judgments of the attitude object). While this focus on implicitly

The authors thank Denny Borsboom, Jamie Cummins, Jan De Houwer, Jens Lange, Pieter Van Dessel, and Frenk van Harreveld for comments and discussions.

J. D. was partly supported by a grant from the National Science Foundation (BSC-1918490).

Correspondence concerning this article should be addressed to Jonas Dalege, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. E-mail: j.dalege@gmail.com

measured attitudes has produced many interesting results, fundamental issues in implicit measurement remain. First, there are theoretical debates about the construct that implicit measures assess. For example, researchers debate whether implicit measures assess stable representations and whether these representations are qualitatively different from representations assessed by explicit measures (e.g., Gawronski, Morrison, Phills, & Galdi, 2017; Hahn & Gawronski, 2019; Schimmack, 2019; Vianello & Bar-Anan, 2020). Second, implicit measures have some unsatisfactory psychometric characteristics—implicit measures vary in the degree to which they are reliable, and results regarding validity are also far from optimal (e.g., Bar-Anan & Nosek, 2014; Cummins, Hussey, & Hughes, 2019; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Schimmack, 2019). Here, we model implicit measures using our network theory of attitudes (Dalege et al., 2016; Dalege, Borsboom, van Harreveld, & van der Maas, 2018) and by this show that the theoretical debates on implicit measures and their psychometric issues are intertwined.

IMPLICIT MEASURES OF ATTITUDE

In line with De Houwer (2006), we use the term *implicit measure* as referring to the outcome of a measurement and the term *indirect measure* as referring to the measurement procedure. More specifically, we use the term *implicit measure*, in contrast to explicit measure, for a measurement outcome that is the result of evaluating the attitude object in a quick and effortless manner (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006; Hahn & Gawronski, 2019). We use the term *indirect measures*, in contrast to direct measures (e.g., typical self-report attitude questionnaires), as a procedure that assesses individuals' attitudes without asking them to reflect on their attitudes.

Three of the most popular indirect measures of attitudes are the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), evaluative priming tasks (Fazio, Jackson, Dunton, & Williams, 1995), and the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005). In the IAT, individuals have to categorize two opposing attitude objects; here, we use the example of categorizing ingroup members versus members of a stigmatized group and positive versus negative words. In one block, participants have to categorize ingroup members with positive words and members of the stigmatized group with negative words. In the other block, the stimuli are switched so that participants have to categorize members of the stigmatized group with positive words and ingroup members with negative words. The IAT score is based on the difference in performance between the two blocks. For example, faster reaction times in the first block than in the second block would indicate that the individual has a more positive attitude toward the ingroup than toward the stigmatized group. The IAT can be seen as an indirect measure, because it assesses participants' attitudes by differences in reaction times without directly asking individuals about their attitudes. Whether scores on the IAT represent implicitly measured attitudes depends on whether individuals do not reflect on their attitudes assessed by the IAT. For example, if individuals recognize that the IAT assesses their attitudes toward the stigmatized

group and try to not show negative attitudes toward the stigmatized group, the IAT scores represent a more explicit measure (cf., Fiedler & Bluemke, 2005).

In a typical evaluative priming task (Fazio et al., 1995), attitudes are assessed by using two contrasting attitude objects as primes (e.g., pictures of members of a stigmatized group vs. pictures of members of the ingroup) that are followed by positive or negative target words, and participants have to judge the target words as negative or positive as quickly as possible. The participants' attitude is then inferred by differences in reaction times. The AMP (Payne et al., 2005) works in a similar way but with the difference that individuals have to judge a Chinese ideograph shown after the prime. The participants' attitudes toward the prime is then inferred by their judgments of the Chinese ideographs. In both evaluative priming tasks and the AMP, attitudes are thus assessed indirectly by making attitude-related stimuli task irrelevant. Similar to scores on the IAT, scores on evaluative priming tasks and the AMP only represent implicit measures if individuals' reflections on the primes do not influence the scores. For example, it has been argued that the AMP might not succeed in measuring participants' attitudes implicitly, because participants rate the primes instead of the Chinese ideographs (Bar-Anan & Nosek, 2012).

Here, we apply our network theory of attitudes (Dalege et al., 2016, 2018) to implicit measures. The resulting model is based on the fact that indirect measures assess attitudes without asking participants to reflect on their attitudes. In our network theory of attitudes, a central determinant of attitude dynamics is the entropy of an attitude, a measure of stability and consistency derived from thermodynamics. In thermodynamics, entropy is scaled by (inverse) temperature, so that low temperature leads to low entropy. In our network theory of attitudes, we use an analogue of inverse temperature that determines attitudinal entropy. This analogue of inverse temperature subsumes different psychological processes, such as reflecting on the attitude, paying attention to an attitude object, or thinking about an attitude object. In the remainder of this article we refer to these processes as attitudinal entropy reduction.

A NETWORK THEORY OF ATTITUDES

The central premise of our network theory of attitudes holds that attitudes are properties that emerge from lower level evaluative reactions (Dalege et al., 2016). Specifically, attitudes are defined as networks of interacting evaluative reactions, representing the nodes in attitude networks. Based on the tripartite model of attitudes (Rosenberg, Hovland, McGuire, Abelson, & Brehm, 1960), these nodes represent evaluative reactions that take the form of feelings, beliefs, and behaviors vis-à-vis an attitude object, see Figure 1. These evaluative reactions represent current states of the nodes that are in principle observable (if only to the person having them). For example, an attitude network toward a stigmatized group might consist of negative feelings such as fear; beliefs counteracting these feelings such as support of equal rights; and behaviors toward the stigmatized group, including spontaneous behavior, such as reacting fearfully when one encounters a member

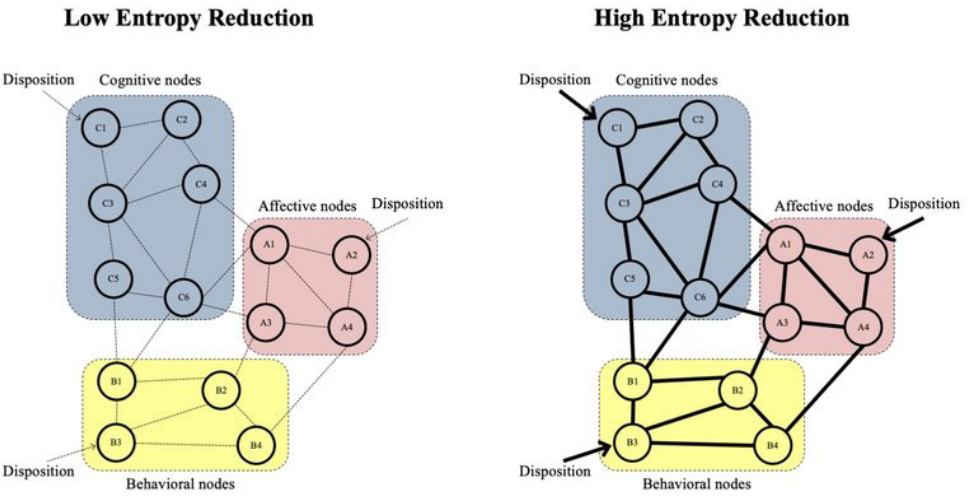

FIGURE 1. A graphic representation of our network theory of attitudes (Dalege et al., 2016, 2018). Edges represent interactions between attitude nodes and arrows represent dispositions of nodes that are influenced by factors outside the network (note that we only display dispositions for three nodes and omit the dispositions of the other nodes). Thicker and solid lines represent stronger influence of dispositions and interactions between attitude nodes under higher attitudinal entropy reduction.

of the stigmatized group and deliberate behavior such as voting for a member of the stigmatized group.

Edges in attitude networks represent interactions between evaluative reactions at the functional level. A positive edge between two evaluative reactions indicates that showing one reaction increases the likelihood of also showing the other reaction, while a negative edge indicates that showing one evaluative reaction decreases the likelihood of showing the other reaction (e.g., fearful reactions toward a stigmatized group decrease the likelihood that one shows support for equal rights and vice versa). Because of these edges, the nodes in the network not only are descriptions of current states of the nodes but also determine states of other nodes. In addition to the influence between nodes, which are modeled by the edges in attitude networks, nodes are also influenced by external factors that are not covered by the attitude network. These external factors can be either personal characteristics (e.g., an individual, who is generally predisposed to fearful reactions, might have a strong disposition to react fearfully to members of a stigmatized group) or factors external to the person (e.g., frequent exposure to media coverage on crimes committed by members of the stigmatized group might increase the likelihood that a person shows fearful reactions to the stigmatized group). This external influence is modeled by allowing for different dispositions of the nodes and thus determines the likelihood of a given node's state independent of the other nodes. Whether these dispositions translate into states of the nodes depends also on connected nodes (e.g., fearful reactions toward members of a stigmatized group might be inhibited by believing in equal rights).

A central assumption of our network theory of attitudes is that psychological analogues of inverse temperature increase both the interactions between attitude nodes and the influence of the dispositions on the nodes' states and thereby reduce attitudinal entropy (Dalege et al., 2018). These psychological analogues are, for example, reflecting on the attitude and attention and thought directed at the attitude object. Because of the increased interactions between attitude nodes and influence of dispositions, the overall behavior of the attitude network becomes more consistent, stable, and extreme (i.e., low in entropy). When an individual reflects on, for example, her attitude toward the stigmatized group, feelings of fear and beliefs in equal rights are pressured to cohere, so that beliefs in equal rights might overrule feelings of fear, and an overall positive attitude—represented as the sum of the states of the nodes in the attitude network—toward the minority group emerges. On the other hand, when an individual does not think much about the stigmatized group, the nodes in attitude networks behave relatively independently from each other, so that feelings of fear and beliefs in equal rights do not inhibit each other. This limited dependence of nodes on their dispositions and other nodes results in more random behavior of attitude nodes.

Based on these assumptions, our network theory of attitudes is able to reproduce several established phenomena in the attitude literature (Dalege et al., 2018), such as the mere thought effect (i.e., that merely thinking leads to polarization of attitudes; Tesser, 1978; Tesser & Conlee, 1975) and the classic finding in the persuasion literature that uninvolved individuals rely on heuristic cues while involved individuals process arguments systematically (Chaiken, Liberman, & Eagly, 1989; Petty & Cacioppo, 1986). Most relevant to the current article, our network theory of attitudes is also able to integrate the finding that scores on indirect measures of attitudes show high stability and strong prediction of behavior at the group level, while they show low stability and weak prediction of behavior at the individual level (Payne, Vuletich, & Lundberg, 2017).

IMPLICIT MEASURES AND ATTITUDINAL ENTROPY REDUCTION

In our earlier work, we focused on general properties of implicit measures that follow from our network theory of attitudes (Dalege et al., 2018). Based on the assumption that reflecting on attitudes reduces attitudinal entropy, it follows that measurement of attitudes influences attitudinal entropy. It is important to note here that the natural state of attitude networks is that attitudes are in high entropy states. Because direct measures ask individuals to reflect on their attitudes, direct measures tap attitudes in lower entropy states than indirect measures, all else being equal. On the other hand, personally important attitudes, about which individuals think frequently (Krosnick, Boninger, Chuang, Berent, & Carnot, 1993), are in chronically low entropy states. As a consequence, measuring such attitudes directly or indirectly has less effect on the entropy of the attitude—If an individual, for example, just thought about a stigmatized group before the attitude measurement, measuring this individual's attitude indirectly or directly will not result in a strong difference in entropy reduction.

From this reasoning, the hypothesis follows that characteristics of the attitude influence the psychometric properties of (indirect) measurement. Specifically, how much individuals think about an attitude object should be positively related to the internal consistency of indirectly measured attitudes (Dalege et al., 2018). This hypothesis was recently supported by an analysis of internal consistencies of 190 different IATs (Van Dessel, De Houwer, Hughes, & Hussey, 2018). While not providing definite proof of the measurement implications of our network theory of attitudes, this analysis serves as an illustration that internal consistencies of indirectly measured attitudes systematically vary with substantive properties of the assessed attitude. Our theory implies that the attitudes which showed high internal consistency were measured less implicitly than the attitudes that showed low internal consistency. Conversely, our theory implies that directly measured attitudes showing low internal consistency are likely to assess attitudes in a more implicit way (e.g., a direct measure might result in an implicit measurement outcome if the individual was distracted during answering a questionnaire). We thus assume that indirect measures *only* succeed in measuring attitudes implicitly if they assess attitudes in high entropy states.

A corollary of this assumption is that some attitudes *cannot* be measured implicitly. For example, measuring highly important attitudes that are in chronically low entropy states indirectly or directly should converge to the same result. This corollary is supported by the finding that the relation between indirectly and directly measured attitudes is stronger for personally important attitudes than for personally unimportant attitudes (Karpinski, Steinman, & Hilton, 2005).

While our earlier analysis of implicit measures led to a sobering conclusion—implicit measures are necessarily noisy—in the current article we focus on a more positive aspect of implicit measures from the perspective of our network theory of attitudes: By assessing attitudes in noisier states, implicit measures are able to assess the dispositions of conflicting attitude nodes (i.e., attitude nodes with dispositions that are not in line with the dominant nodes' dispositions) in a more fine-grained and accurate manner. To illustrate this point, we use an attitude network toward a stigmatized group as an example. Let us assume that the affective nodes have negative dispositions (e.g., individuals are disposed to negative feelings toward a stigmatized group) and that cognitive nodes have positive dispositions (e.g., individuals are disposed to believing in equal rights). Additionally, we assume that the affective nodes are of a lower number than cognitive nodes, which is based on the finding that individuals generally list less affective properties than cognitive properties in open-ended questionnaires assessing attitude components (Esses & Maio, 2002; Haddock & Zanna, 1998). We focus on the situation in which the implicit measure taps mostly affective evaluative reactions, as implicit measures generally do (e.g., Cunningham & Zelazo, 2007; Fazio & Olson, 2003; Gawronski & Bodenhausen, 2006; Hofmann et al., 2005; Smith & Nosek, 2011).[1]

1. This does not hold for all indirect measures, as some indirect measures, such as the stereotype misperception task (Krieglmeyer & Sherman, 2012) and IATs assessing gender stereotypes (Nosek, Banaji, & Greenwald, 2002), are designed to tap cognitive reactions.

How do these assumptions interact with the principle that explicit measures result in higher entropy reduction than implicit measures? When we measure an attitude network in a high entropy state, the different nodes will be noisier but they will also be less dependent on other nodes. Because of this, implicit measures will be more likely to tap the conflicting nodes' (e.g., affective nodes') dispositions. In contrast, when we measure the attitude network in a low entropy state, the measurement will be less noisy but we will also be less likely to measure nodes independently from each other. Because of this, the dominant nodes' (e.g., cognitive nodes') positive dispositions will overrule the negative conflicting nodes' (e.g., affective nodes') dispositions, because all nodes have bidirectional links to each other.[2] Scores on an explicit measure aimed at conflicting evaluative reactions are thus determined by both dispositions of these conflicting evaluative reactions *and* connections to the dominant evaluative reactions. In the simulation reported next, we provide a formal illustration of this mechanism.

SIMULATION

The aim of this simulation is to provide a formal illustration of the implied dynamics of our network theory of attitudes for implicit versus explicit measures. To do so, we set up a simulation using a simple network with 12 nodes that are all positively connected. Four of these nodes represent the conflicting nodes in the attitude network and have negative dispositions. The other eight nodes represent the dominant nodes in the attitude network and have positive dispositions. We then measured the network under varying amounts of reflection on the attitude (by varying attitudinal entropy reduction). The variations in attitudinal entropy reduction represent that indirect measurement was successful in measuring attitudes implicitly. The R-code for the simulation is available at https://osf.io/svgcb/?view_only=5441f0354bcf495282224fe7d6cf62b4.

SIMULATED DYNAMICS

Our network theory of attitudes uses the Ising (1925) model as an idealized model of attitude dynamics. To use the Ising model, we make the simplifying assumption that attitude nodes are of binary nature. For example, a node representing fearful reactions to members of a stigmatized group can be either in the -1 state (one currently has fearful reactions) or in the +1 state (one has currently no fearful reactions). To simulate dynamics on such Ising networks, we use Glauber dynamics (Glauber, 1963). This is a way to implement the Ising model's central postulate that systems strive toward energy minimization. In attitude networks, this energy minimization represents consistency maximization.

Glauber dynamics work in the following way. First, a node in the network is randomly chosen. Second, the energy $E(x_i)$ of this node is calculated. Third, the

2. Note that this does not imply that individuals cannot report ambivalent attitudes. Our network theory of attitudes, however, implies that the more individuals reflect on their attitude, the less likely it will be that individuals accurately report their ambivalent attitudes.

energy when the node was flipped to its opposite state $E(-x_i)$ is calculated. Fourth, the node flips its state with a probability dependent on the difference between $E(x_i)$ and $E(-x_i)$. If the flipped state has lower (higher) energy, the node is likely (unlikely) to flip. The way this energy difference translates into a probability is moderated by the parameter β that represents inverse temperature in the original Ising model. In attitude networks, β represents attitudinal entropy reduction. When individuals, for example, reflect on the attitude, the behavior of the attitude network becomes more organized, because nodes are more likely to move to low energy states and by this the attitude becomes more consistent and stable.

The energy of a given node is determined by the following equation:

$$E\left(x_i\right) = -\tau_i x_i - \sum_j \omega_{ij} x_i x_j, \tag{1}$$

where x_i represents the randomly chosen node. τ_i represents the node's disposition to be in the positive or negative state. ω_{ij} represents the connection weight between x_i and one of its connected nodes x_j. Energy becomes lower when the state of the chosen node conforms to its disposition and when the node is in the same (different) state as the nodes to which it is positively (negatively) connected. Energy is thus lowest when the node is consistent with its disposition and with the other nodes it is connected to.

The probability that the node flips is then given by:

$$\Pr\left(x_i \rightarrow -x_i\right) = 1 / \left(1 + e^{(-\beta(E(x_i) - E(-x_i)))}\right), \tag{2}$$

where $E(x_i)$ represents the energy of the chosen node and $E(-x_i)$ represents the energy when the node was flipped. β represents attitudinal entropy reduction and moderates how strongly the nodes' behavior depends on the network parameters (i.e., weights and dispositions). If β goes to infinity, the network's behavior becomes completely deterministic—nodes always change to (or remain in) the state with lower energy. If β is zero, the network's behavior becomes completely random—nodes always change their state with a .5 probability, and the energy of the states has no influence on the nodes' behavior. If β lies between zero and infinity, nodes are more likely to change to (or remain in) the state with lower energy—with the behavior of the network becoming increasingly deterministic with increasing β.

SIMULATION SETTINGS

We used a fully connected 12-node network for the simulation with all edge weights, ω_{ij}, set to .1. The first four nodes represent the negatively disposed conflicting nodes, with dispositions τ_i sampled from a uniform distribution between –.6 and –.3, which represents moderate to strong dispositions. The last eight nodes represent the positively disposed dominant nodes, with dispositions τ_i sampled from a uniform distribution between .1 and .4, which represents moderate dispositions. For each simulated individual, one value for the dispositions of the conflicting nodes and one value for the dispositions of the dominant nodes were drawn,

but for the sake of simplicity a given individual had the same disposition for each conflicting node and the same disposition for each dominant node. Every individual thus had a negative disposition for the conflicting nodes and a positive disposition for the dominant nodes, but the strength of these dispositions varied between individuals. Note that this setup results in the majority of individuals having a positive disposition on average for the whole attitude network.

β was varied between 0 and 2, representing that individuals, for example, vary in how much they reflect on their attitude. β was varied in steps of .1, resulting in 21 different values for this parameter. The variations in β represent different types of measurement. Low values (approximately lower than 1) represent implicit measurement, moderate values (approximately between 1 and 1.5) represent explicit measurement, and high values (approximately higher than 1.5) represent situations in which individuals are motivated to elaborate during the measurement (e.g., individuals are asked to think about the attitude object before they answer the question; cf., Tesser, 1978). For each value of β, we simulated 5,000 individuals, resulting in a total of 105,000 simulated individuals.

We simulated individuals by first randomly drawing the dispositions for the conflicting and dominant nodes, respectively, and the value for β. Then based on these settings, 500 iterations of Glauber dynamics were simulated per individual (the values of the disposition and β remained constant throughout these 500 iterations). The first 250 iterations were used to settle the network and the second 250 iterations represented the measurement. We averaged the scores on each node at each iteration for each individual. These scores represent, for example, different trials in an IAT. For each individual, we then calculated the mean of these averages as their attitude score. To investigate the stability of the attitude scores, we correlated the scores at the 250th and 500th iteration. We used this as a measure of (test-retest) reliability, because it represents the most straightforward and easiest to interpret measure of reliability. Other measures of reliability are less optimal, because they rely on more assumptions (e.g., Cronbach's alpha relies on the assumption that scores are based on a one-factor model with equal factor loadings; for example, McNeish, 2018).

RESULTS

With the simulation, we aimed to answer three questions. First, how do the conflicting negatively disposed nodes behave under different amounts of attitudinal entropy reduction (e.g., how much individuals are asked to reflect on their attitude during the measurement)? Answering this question provides insight into how implicit versus explicit measures affect the likelihood that we measure the disposition of conflicting nodes accurately. We define accuracy of measurement as the alignment of nodes with their dispositions. Second, how do the dominant positively disposed nodes behave under different levels of attitudinal entropy reduction? Answering this question provides insight into how implicit versus explicit measures affect the likelihood that we measure the dispositions of dominant nodes accurately. Third, how is the stability of both sets of nodes affected by variations in

attitudinal entropy reduction? Answering this question provides insight into how we can maximize both accuracy and stability of attitude measures.

To answer the first question of how the conflicting nodes behave under different levels of attitudinal entropy reduction, we investigated the mean of these nodes averaged per β parameter. As can be seen in Figure 2a, the average of these nodes takes the form of a U-shape. Under very low attitudinal entropy reduction, the nodes behave almost completely randomly and we are mostly measuring noise.[3] When the β parameter increases to moderate values, the nodes behave in accordance with their dispositions—the mean of the individuals is negative. With further increasing β parameter, the nodes again move further away from their dispositions, because they are pressured to align with the positively disposed nodes. Note, however, that the conflicting nodes move to a neutral position, illustrating that their states at this point are results of their negative dispositions and their connections to the positively disposed dominant nodes. The increasing error bars also illustrate that these nodes move to increasingly extreme positions—in general, they move to the positive position congruent with the dominant nodes, but in some instances, they move to the negative position congruent with their own dispositions.

To answer the second question of how the dominant nodes behave under different levels of attitudinal entropy reduction, we investigated the mean of these nodes averaged per parameter. As can be seen in Figure 2b, the average of these nodes is linearly increasing to the point when the β parameter reaches 1.5. This implies that with increasing attitudinal entropy reduction up to this point, we measure these nodes with increasing accuracy. When the β parameter further increases (which would likely be the case when individuals are prompted to elaborate on their attitude), the nodes move again move away from their dispositions. This is due to the increasing pressure of the nodes to align (as can be seen by the increasing variance), so that in some instances all nodes become negative. As can be seen in Figure 2c, the average of the whole network shows similar behavior as the dominant nodes—implying that the positively disposed dominant nodes overrule the negatively disposed conflicting nodes up to the point where the attitude moves increasingly to both extremes. This increasing bimodality of the attitude illustrates that in some instances the conflicting nodes overrule the dominant nodes. Such a situation could, for example, arise when individuals rationalize their conflicting feelings toward a minority group by changing their dominant beliefs (cf., Crandall, Bahns, Warner, & Schaller, 2011).

To answer the third question of how the stability of the attitude nodes is affected by attitudinal entropy reduction, we investigated the correlation between the different sets of attitude nodes at the 250th and 500th iteration as a measure of test-retest reliability. As can be seen in Figures 2d–f, the test-retest reliability of all attitude nodes increases with increasing β parameter. This implies that measuring the conflicting nodes both accurately and with high reliability is not possible. The point where we can measure these nodes most accurately lies at 0.9 for the β parameter (a point likely to represent implicit measurement). However, stability

3. Note that the small error bars at low levels of the attitudinal entropy reduction arise because all nodes constantly flip, which results in a stable neutral sum score of the nodes.

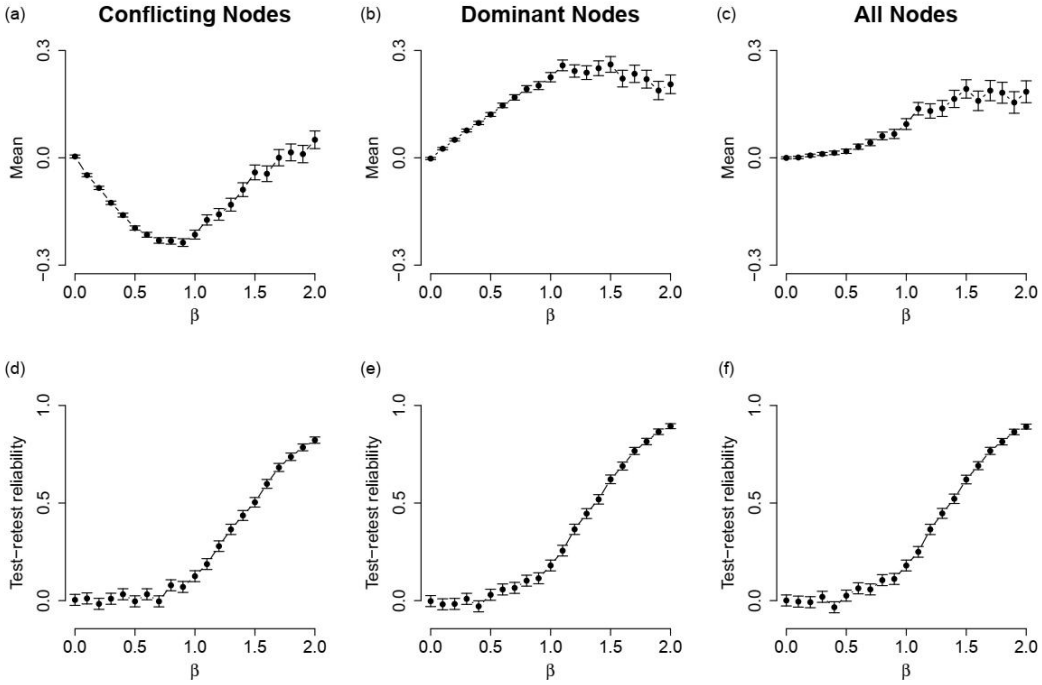

FIGURE 2. Results of the simulation. (a) shows the mean of the conflicting nodes for the different β parameters representing different levels of attitudinal entropy reduction; (b) shows the mean of the dominant nodes; and (c) shows the mean of all nodes. (d) shows the test-retest reliability of the mean of the conflicting nodes between the 250th and the 500th iteration for the different β parameters; (e) shows the test-retest reliability of the mean of the dominant nodes; and (f) shows the test-retest reliability of the mean of all nodes. Error bars represent +/- two standard errors around the mean or test-retest reliability.

of the nodes at this point is rather low (the test-retest reliability is only .07 when the β parameter is at 0.9).[4] In contrast, accuracy and reliability for the measurement of dominant nodes aligns rather well. These nodes are measured most accurately when the β parameter is at 1.5 (a point likely to represent explicit measures) and the test-retest reliability at this point is .62. While further increases lead to less accurate measurement of the dominant nodes but higher stability, the decrease in accuracy is far from the decrease in accuracy for the conflicting nodes.

DISCUSSION

In this article, we modeled implicit versus explicit measures using our network theory of attitudes (Dalege et al., 2016, 2018). This modeling of implicit versus

4. One might object that the correlation is unrealistically low, because test-retest reliability is higher for implicit measures (LeBel & Paunonen, 2011). However, test-retest reliability in the implicit measures literature is typically assessed by correlating two scores that are each averaged over many trials. The way we assessed stability can be best compared to a situation in which a researcher would correlate two trials of an implicit measure.

explicit measures is based on the fact that indirect measures assess attitudes without asking individuals to reflect on their attitude with the aim to assess attitudes implicitly. Linking this fact to our network theory of attitudes implies that indirect measures tap attitudes in lower entropy states than direct measures, all else being equal. Indirect measures succeed in measuring attitudes implicitly if they tap the attitude in low entropy states. Results of our simulation show that limiting the amount of attitudinal entropy reduction can lead to more accurate measurement of the nodes of interest independent of their relationship to other nodes (such as when one wants to measure attitude elements that have conflicting dispositions with the dominant attitude elements), but also to lower reliability. Implicit measures therefore present a principally noisier measure than explicit measures (cf., Schimmack, 2019), but in some instances this noise allows for a more accurate measurement.

IMPLICATIONS

Our network theory of attitudes has several implications for implicit measures. A straightforward implication of our theory is that the difference between implicit and explicit measures is not a qualitative difference (cf., De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). Instead, measurement of attitudes ranges from highly implicit when attitudinal entropy reduction is low (e.g., the attitude object is evaluated without reflecting on it) to highly explicit (e.g., the attitude object is evaluated while individuals elaborate on their attitude).

A corollary of treating the distinction between implicit and explicit measures as a continuous one is that the same measurement instrument might in some situations be regarded as an implicit measure, while in other situations it might be regarded as a more explicit measure. For example, if a study first asks participants to think about their attitudes toward a given issue and then assesses the participants' attitudes with an IAT, the IAT in this instance is probably a more explicit measure than when the IAT is administered without having participants think about their attitudes beforehand. The time scale of how fast attitude entropy would rise again after individuals have thought about the attitude object is an important question for future research. A questionnaire under time pressure on the other hand might be interpreted as an implicit measure. Given that our simulation shows that there is an optimal point of the implicitness of a measure at which the accuracy of the measurement (for conflicting attitude nodes) is maximized while reliability is in an acceptable range, an important task for future research is to investigate where this point lies. A possibility to do this would be to manipulate how much individuals reflect on their attitudes during a measure. This could, for example, be accomplished by varying the presentation of primes in an evaluative priming task. Our network theory of attitudes predicts that with longer presentation of the primes, reliability of the tasks would increase but also that the accuracy of measuring conflicting attitude nodes would decrease at some point.

A related implication is that the relatively low internal consistency and test-retest reliability of implicit measures (e.g., Bar-Anan & Nosek, 2014; Gawronski et al.,

2017; Hofmann et al., 2005; LeBel & Paunonen, 2011) is (to some extent) inherent to the construct implicit measures assess and cannot be solved by improving the measurement instruments.[5] This implication has three important methodological consequences for implicit measures of attitudes. First, relations between implicitly measured attitudes and *any* other variable are necessarily relatively weak, because of their high variability.[6] This consequence aligns well with a recent meta-analysis showing that indirectly measured attitudes show rather weak relations to behavior (Oswald et al., 2013).[7] Our network theory of attitudes therefore implies that only studies with large sample sizes have sufficient power to detect effects of implicitly measured attitudes on other variables.

Second, several studies assessing implicitly measured attitudes correct for low reliability (e.g., Cunningham, Preacher, & Banaji, 2001; Hofmann et al., 2005; Nosek & Smyth, 2007). From the perspective of our network theory of attitudes this practice leads to artificially high effects of implicitly measured attitudes on other variables. The reason for this is that correcting for low reliability rests on the assumption that the construct could in principle be measured with perfect reliability (Borsboom & Mellenbergh, 2002). This assumption, however, is unlikely to be tenable in the case of implicitly measured attitudes.

Third, our network theory of attitudes also implies that implicit measures are better suited to assess group means than individual differences (cf., Payne et al., 2017). This is also illustrated by our earlier modeling work on implicit measures (Dalege et al., 2018). We modeled the finding that scores on indirect measures show high stability at the group level versus low stability at the individual level (e.g., Baron & Banaji, 2006; Gawronski et al., 2017) by simulating a large group of individuals in which the majority has a positively disposed attitude network (Dalege et al., 2018). Note that this simulation assumed that the indirect measures resulted in implicit measurement outcomes. The β parameter, representing attitudinal entropy reduction, was set to a low value. We found that such a situation results in high variability at the individual level. The group mean, however, was remarkably stable. The finding that scores on indirect measures show strong relations to behavior at the group level versus weak relations at the individual level (Hehman, Flake, & Calanchini, 2018; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013) was modeled in a similar way. In this simulation, we simulated several groups varying in the dispositions of their attitude networks, and we added a node representing behavior (Dalege et al., 2018). The β parameter was again set to a low value.

5. One might object that this is an overgeneralization because we only used one measure of reliability in our simulation. However, all measures of reliability (e.g., Cronbach's alpha, split-half reliability, coefficient omega) rely on the correlational structure between items. Therefore our test-retest reliability measure using correlations between iterations should have a linear relation with other measures of reliability.

6. Note that scores on indirect measures might sometimes show rather strong relations to other measures. However, our network theory of attitudes implies that such strong effects arise because other factors, such as personal importance of the attitude, caused the indirect measure to assess attitudes explicitly.

7. Note that our network theory of attitudes also implies that under low attitudinal entropy reduction, behavior is mostly determined by strong dispositions regardless of other nodes in the attitude network (e.g., a strong habit).

We found that the individual-level correlation between the behavior node and the other nodes was rather weak, while the correlation at the group level was strong. The underlying process of these findings is that, due to low attitudinal entropy reduction, the networks at the individual level show highly variable behavior. At the group level, these variations are averaged out, so that stable effects emerge.

Another implication of our network theory of attitudes is that the question of what kinds of measures are more likely to tap the "true" attitude is misguided. Different forms of measurement set in motion different processes, and therefore the more relevant question becomes which processes the researcher wants to measure. Similar to a recent account of implicit measures as simulations of everyday processes (De Houwer, 2019), our network theory of attitudes implies that the measurement of attitudes mirrors different situations. If one wants to simulate the dynamics of attitudes in high entropy states (e.g., in situations in which individuals react spontaneously; Fazio, 2007; Gawronski & Bodenhausen, 2006), indirect measures are a valid way to assess attitudes. On the other hand, if one is interested in situations in which attitudes are in relatively low entropy states (e.g., in a conversation about the attitude object), direct measures of attitudes are a more valid way to assess attitudes. In general, our network theory of attitudes implies that measurement of attitudes is not simply a way to read out individuals' attitudes, but also sets processes in motion that influence attitudinal processes. This implication echoes work on measurement effects on attitudes (e.g., Strack & Martin, 1987; Thurstone, 1927; Tourangeau, Rips, & Rasinski, 2000; Wang & Busemeyer, 2013).

CONCLUSION

In this article, we applied our network theory of attitudes to implicit measures and by this introduced a formal model of implicit measures of attitudes. This model rests on the principle that implicit measures assess attitudes in high entropy states, leading to low consistency and low stability of attitudes. A central implication of our network theory of attitudes is that implicit measures are sometimes more accurate than explicit measures *because* they are noisier. We hope that by creating a closer connection between substantive theorizing and psychometrics, our network theory of attitudes contributes to a better understanding of the construct that implicit measures assess.

REFERENCES

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin*, *38*, 1194–1208.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*, 668–688.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, *17*, 53–58.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*, 505–514.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–252). New York: Guilford.

Crandall, C. S., Bahns, A. J., Warner, R., & Schaller, M. (2011). Stereotypes as justifications of prejudice. *Personality and Social Psychology Bulletin*, *37*, 1488–1498.

Cummins, J., Hussey, I., & Hughes, S. (2019, May 21). The AMPeror's new clothes: Performance on the affect misattribution procedure is mainly driven by awareness of influence of the primes. https://doi.org/10.31234/osf.io/d5zn8

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*, 163–170.

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, *11*, 97–104.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) Model. *Psychological Review*, *123*, 2–22.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018). The Attitudinal Entropy (AE) framework as a general theory of individual attitudes. *Psychological Inquiry*, *29*, 175–193.

De Houwer, J. (2006). What are implicit measures and why are we using them? In R.W. Wiers & A.W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage.

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science*, *4*, 835–840.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*, 347.

Esses, V. M., & Maio, G. R. (2002). Expanding the assessment of attitude components and structure: The benefits of openedned measures. *European Review of Social Psychology*, *12*, 71–101.

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social Cognition*, *25*, 603–637.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition. research: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology*, *27*, 307–316.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*, 300–312.

Glauber, R. J. (1963). Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, *4*, 294–307.

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*, 1464–1480.

Haddock, G., & Zanna, M. P. (1998). On the use of open-ended measures to assess atitudinal components. *British Journal of Social Psychology*, *37*, 129–149.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology, 16*, 769–794.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*, 393–401.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369–1385.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus [Contribution to the theory of ferromagnetism]. *Zeitschrift Für Physik*, *31*, 253–258.

Karpinski, A., Steinman, R. B., & Hilton, J. L. (2005). Attitude importance as a moderator of the relationship between implicit and explicit attitude measures. *Personality and Social Psychology Bulletin, 31*, 949–662.

Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, *103*, 205–224.

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology*, *65*, 1132–1151

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*, 570–583.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*, 44–59.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology, 54*, 19–24.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277–293.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*, 233–248.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.

Rosenberg, M. J., Hovland, C. I., McGuire, W. J., Abelson, R. P., & Brehm, J. W. (1960). *Attitude organization and change: An analysis of consistency among attitude components.* New Haven, CT: Yale University Press.

Schimmack, U. (2019). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*. Advance online publication.

Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology, 42*, 300–313.

Strack, F., & Martin, L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123–148). New York: Springer-Verlag.

Tesser, A. (1978). Self-generated attitude change. *Advances in Experimental Social Psychology*, *11*, 289–338.

Tesser, A., & Conlee, M. C. (1975). Some effects of time and thought on attitude polarization. *Journal of Personality and Social Psychology*, *31*, 262–270.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Van Dessel, P., De Houwer, J., Hughes, S., & Hussey, I. (2018). An analysis of the scientific status and limitations of the attitudinal entropy framework and an initial test of some of its empirical predictions. *Psychological Inquiry*, *29*, 213–217.

Vianello, M., & Bar-Anan, Y. (2020). Can the Implicit Association Test measure automatic judgment? The validation continues. *Perspectives on Psychological Science.* Advance online publication.

Wang, Z., & Busemeyer, J. R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. *Topics in Cognitive Science*, *5*, 689–710.