# The Implicit Bias Transformation Program - Report

Siobahn Hotaling

Harvard Extension School

PSYC E-599 Section 1

Final Capstone Report – May 12, 2020

**Introduction**

One of the biggest challenges that we face in today's globalized world is how we – as individuals and as groups – treat one another. It is a natural human tendency to categorize, which leads to create certain biases or prejudices (Allport, 1954). However, when left unchecked, these beliefs can cause outcomes that range from unconscious microaggressions to targeted violence. If humanity is to survive this next stage of our collective history, we need to address these intolerances directly and develop tools that can help us evolve and thrive together.

Because the origins of these categorizations come early in life in the form of personal experiences and cultural indoctrination, they are often inaccessible to us on a conscious level. Ultimately, this means that our personal biases – whether we are aware of them or not - are rooted in an automatic, subconscious and involuntary process of cognition, a process borne from our most primitive internal mechanisms designed to ensure survival (Cottrell & Neuberg, 2005). Once a bias is created – whether it be through individual experience or social conditioning - we are automatically triggered when we experience the group in question. Perhaps even more importantly, this response exists despite any conscious, non-prejudiced beliefs we may hold (Devine, 1989).

While there has been a great deal of focus on bias in human behavioral research in recent years, the bulk of diversity trainings and interventions created to date have fallen short of creating any significant shifts. Methods such as education, awareness-building exercises, and repeated intergroup contact only work with the explicit aspects of prejudice, and tend to target groups rather than individuals; none have achieved lasting effects (Paluck & Green, 2009, Forscher et al, 2019). The best results to date have only shown improvements in behavioral

management; while individuals may learn to manage the expression of their implicit biases externally, they still experience (and are affected by) them internally (Paluck & Green, 2009).

These results are likely due to the fact that our biases have an inherent emotional component to them, yet the majority of these programs only address explicit, post-emotional behaviors. The constructs of emotions and biased responses actually share the same type of automated mechanism (Damasio, 2010; Cottrell & Neuberg, 2005); therefore, for any type of intervention to be truly successful in helping individuals remove their bias, the implicit and emotional aspects need to be addressed.

The Implicit Bias Transformation Program (IBTP) will address the limitations of the interventions that have preceded it by taking a more personalized and internally-focused approach, utilizing a methodology that works with the implicit aspect of individual beliefs through uncovering and addressing the origins and emotions surrounding these biases. The program's goal is to help motivated individuals ultimately transform their implicit biases - using combination of conceptual change instruction, metacognitive exercises and emotional explorations to specifically target the automated aspect of their biased beliefs.

In contrast to previous interventions, the program will utilize learning constructs that aim to create fundamental perceptual shifts in participants. This will be achieved first by utilizing the conceptual change framework outlined by Strike and Posner (1985) to create a foundation of understanding of the nature of bias. Metacognition – the act of reflecting and understanding one's own thoughts - is also a critical aspect of the IBTP; participants will be frequently asked to self-reflect and evaluate their thoughts and feelings throughout the process, as not only a measurement of their progress but also as a conduit of change. Finally, the program will not only assist participants in uncovering their own biases in a safe and welcoming environment, but

will also add a new – and necessary – angle of exploration. With a direct focus on the implicit aspect of bias, the IBTP will utilize the understanding of emotions, combined with the tool of metacognition, to help individuals facilitate the examination and resolution of their unexamined beliefs.

Participants – made up of adult volunteers who have expressed a desire to be more tolerant, and thus willing to examine their own limitations - will first be given a test to identify and measure their personal implicit biases. They will then go through a brief education to help them understand the nature of stereotypes and bias and how they are created. The next aspects of the program will be individually customized for each participant. Their attitudes and the nature of their bias will dictate the length and type of interventions that are administered in personal sessions with skilled facilitators, who will employ emotional processing therapy methods to assist participants in examining the implicit emotional aspects of their bias. These explorations will ultimately help them resolve their previously unexamined automatic responses, thereby permanently removing the implicit bias and giving the individual the emotional freedom to be more tolerant, humane, and empathetic towards outgroup persons.

The approach this program takes is more personalized – and therefore, more time-and-resource intensive - than other diversity training programs, because we believe that such an individualized design is necessary in order to create significant, permanent change in the participants. Ultimately, the goal of the IBTP is to become a highly effective and widely-utilized model that can address implicit bias of all kinds and truly transform the feelings, thoughts, and beliefs of those that participate, offering humanity a new tool to manage the challenges of our increasingly globalized world.

**Program Summary**

While the ultimate objective of the IBTP is to help participants transform their specific biases through a targeted exploration process, it has built into it a number of smaller objectives. Each of these objectives are implemented in order to methodically build a foundation for participants, in order to set the optimal conditions for the success of the overall objective. The act of facing our biases is a not an easy one, and therefore it is important to consider all of the challenges that participants will encounter in the process. To this end, participants will first undergo a series of educational and experiential trainings designed to prepare them – cognitively and emotionally - for this exploration.

To start, the IBTP utilizes conceptual change methods to instruct participants on the nature of implicit bias, with the goal of helping them understand their bias and gain a more compassionate perspective; this is an important prerequisite for the exploration process that will occur later in the program. Participants are then taught how to trigger the processes of metacognition and Damasio's "as-if" body loop (Damasio, 2010), tools that will be employed later in the program for ongoing self-assessments.

The participants will then go through another group-based training about shame and self-compassion. The emotion of shame often comes up when individuals are coming to terms with their biases (Devine & Monteith, 1993), and the IBTP seeks to proactively address this tendency by addressing it directly. Participants are given an overview of the definition and main elements that comprise the model of self-compassion; they are then taken through a series of exercises that utilize self-compassion to address any shameful emotions they might be experiencing.

Participants will then undergo their first self-assessment to determine if they have sufficiently worked through their shame; if they have not, they can repeat this section as many

times as needed until they are ready to move on. This non-linear approach is employed throughout the IBTP and speaks to the program's design, which is intended to ensure that each participant has made the necessary perceptual and emotional shifts needed to be successful in future explorations before moving forward.

Once participants have sufficiently removed their shame, they are now ready to directly examine the specific bias that they identified at the beginning of the program. These sessions are private, one-on-one sessions facilitated by professionals trained in emotional processing therapy and adjacent therapeutic models. Facilitators will use a process involving prompted metacognition, the "as-if" body loop, and emotional processing therapy to help participants examine all of the inconsistent beliefs and negative emotions that surround their bias. This section will be considered complete when a participant feels that they have transformed their thoughts and emotions surrounding their bias, and can span several separate exploration sessions if necessary. At the end of each session, the participant will undergo a self-assessment to determine their progress.

Once the participant believes that they have adequately neutralized their bias, they will move onto the final section of the program, the real-world post-test. Here, the participant will proactively put themselves in a situation that will cause them to interact with the group they have been biased towards, and experience their response in the situation. If they encounter any residual emotions, they can return to the program and engage in further explorations until they have reached their goal – to transform and neutralize their bias on a fundamental and implicit level.

**Program Structure and Understanding Goals**

The Implicit Bias Transformation Program is made up of a pre-testing section followed by four distinct modules, each with its own set of goals and objectives. It is designed in a fashion that allows participants to revisit modules and processes as necessary; rather than a linear, sequential approach, it is entirely objective-based. At the end of each module, participants undergo a guided self-assessment of their emotional and cognitive experience; these assessments determine the next step each individual will take, whether it be repeating the previous module or moving on to the next.

The overall understanding goals of the IBTP are as follows:

- Participants will understand how to transform their attitudes, thoughts, and emotions - ultimately, their overall perception - surrounding a specific social/ethnic/racial group.
- Participants will understand how transforming their bias allows them to live a more genuinely tolerant and accepting life, one where they are internally aligned with their values.

To reach these goals, however, participants need to achieve several secondary understanding goals in each section that will be the scaffolding for the overall transformation. The program is structured in such a way that the successful achievement of each subgoal is necessary for participants to move on to the next section; below is a description of each section in detail, along with the subgoals that are contained within each.

**Pre-Testing Section**

The program begins with a pre-testing section that is done prior to a participant's attendance at the first educational section. This section features a self-assessment form, where the participant reflects on their own biases, followed by an online-based administration of Greenwald, Banaji, and Nosek's (1995) Implicit Association Test (IAT). Once this is complete, they will move on to engage in the educational component of the IBTP.

**Section 1: What Is Implicit Bias?**

- Subgoal 1a: Through an instruction-based conceptual change framework, participants will understand the nature of implicit bias, why it exists in all humans, and how it exists within themselves.
- Subgoal 1b: Participants will understand the concept of metacognition and how to employ it in order to evaluate their own thoughts and perspectives.
- Subgoal 1c: Participants will understand the concept of the "as-if" body loop and how to employ it in order to initiate and evaluate their emotional responses to simulated events.

This initial educational section will be done in a group format; here, the facilitator will use Strike and Posner's (1995) model of conceptual change - the act of fundamentally changing an individual's understanding of a concept through either personal experience or methodical, measured instruction – to transform the way participants view and understand implicit bias. They will learn what implicit bias is, how it is natural, and how it has served humans until now. The goal is to not only educate but also give participants a new, more compassionate perspective on bias so they can feel more comfortable examining their own.

In this section, facilitators will also introduce the process of metacognition – thinking about one's own thoughts (Kuhn, 2000). While metacognition can be applied in a number of cognitive areas, this orientation will specifically focus on the concepts of developing awareness of personal beliefs and knowledge, as well as becoming aware of how these beliefs and knowledge were formed. This aspect of metacognition is critical to the mental processes that will be later required of participants as they explore their biases – which are built on personal beliefs and informed by prior knowledge.

Another process that will be introduced in this section is Damasio's "as-if" body loop (Damasio, 2010). In his model, Damasio describes the usual method of generating emotions as a "body loop", where a source event triggers a series of responses in the body that then lead to emotional feelings and the thoughts that accompany them. The "as-if" loop, then, is a simulation of an actual event that can be trigger simply by imagining it in our minds; we do this every time we anticipate something happening in the future or remember an emotional experience of the past. In the IBTP, however, participants will be learning how to trigger this loop in a more proactive and conscious way. This process, along with metacognition, is critical to the participants' future explorations of their bias, as it allows them to access the emotions and physiology surrounding their beliefs as they pursue understanding.

These tools will be used throughout the program by the participants in their ongoing self-assessments, ones in which they will be asked to undergo a self-reflective exercise where they reflect on their thoughts and emotions surrounding their bias. It is critical that participants gain not just an understanding of these processes on a cognitive level, but also to have the opportunity to further their understanding through performance (Perkins, 1998). While both of these processes are ones that most adult humans utilize subconsciously to varying extents, this section

is designed to give participants the framework to trigger them consciously as needed. To this end, participants will go through a series of exercises to practice these processes in order to further their understanding through performance before moving forward to the next stages of the program, to ensure that participants can employ them at will in future self-assessment exercises.

**Section 2: Learning Self-Compassion for Your Biased Self**

- Subgoal 2a: Participants will understand the nature of the emotion of shame and how it works within themselves
- Subgoal 2b: Participants will understand what self-compassion is and how it works
- Subgoal 2c: Participants will understand how to utilize self-compassion to change their perspective on shameful experiences

This section will also take place in a group setting. Facilitators will first do a brief presentation on the nature of shame, with a focus on what shame means, how it exists in us, and why. They will emphasize the positive reasons why we feel shame, as well as its potential downsides. An important point for facilitators to convey is how the action tendency of shame is to run from the thing they feel ashamed of. This illustrates to the participants why it is so important that they remove their shameful feelings in order to effectively remove their bias.

Facilitators will then introduce the concept of self-compassion to the group and how it works. They will describe the three main elements - self-kindness, common humanity, and mindfulness - and give examples of each (Neff, 2003). They will present some of the misconceptions of self-compassion (for example, that it is a form of self-pity, or a sign of weakness) as well as giving hard evidence of the benefits of self-compassion in various areas. Specific to this program, there have been a number of studies done to measure the effects that

self-compassion has on mitigating shame in various contexts; nearly all show that self-compassion work lessens shame and improves feelings of self-worth (Neff, 2011; Johnson & O'Brien, 2013; Karris & Caldwell, 2015; Martinčeková, & Enright, 2018; Zhang et al, 2018).

The facilitator will then describe how self-compassion can help reduce or remove shame, and how self-compassion exercises will be used in this program to specifically help address their shame surrounding their own personal bias. Once the group based discussion is complete, the facilitator will lead participants in an adapted version of a personal writing-based exercise from The Mindful Self-Compassion Workbook (Neff & Germer, 2019) called "Working with Negative Core Beliefs". This exercise helps participants practice getting in touch with negative beliefs they may have about themselves surrounding their bias, and then walks them through a way to apply the three main elements of self-compassion. This exercise, while useful in its own right, is also a primer for the next step, where they will be examining shame.

For the next part of this section, participants will schedule individual one-on-one meetings with a facilitator to work on their specific feelings of shame that they identified at the end of Section 1. In these one-on-one sessions, participants will go through a self-reflective exercise from The Mindful Self-Compassion Workbook (Neff & Germer, 2019) called "Working with Shame"; in this exercise, the facilitator will ask the participant to first utilize Damasio's "as-if" body loop to re-experience bringing up the event that triggers shame (in this case, it would be the recognition of their bias). They are then additionally guided by the facilitator to utilize metacognition to examine their thoughts around the event, specifically the negative beliefs they have about themselves. The practice of self-compassion is then brought in as a tool to help the individuals lessen the shame and have compassion for the part of themselves that is biased. This

session utilizes a balance between emotional experiencing, metacognition, and self-compassion to help the participant examine, understand, and reconcile their thoughts and emotions.

**Section 3: Personal Exploration of Bias**

- Subgoal 3a: Participants will understand how their personal biases were causally developed through their life experiences
- Subgoal 3b: Participants will understand how their emotions have created and maintained their bias
- Subgoal 3c: Participants will understand how changing their emotions can change their beliefs
- Subgoal 3c: Participants will understand how to transform their emotions and beliefs in order to view the biased group in a neutral/positive light

After participants successfully transform their perspective of their bias, their next step is to engage in a personal, one-on-one exploration of their bias with a skilled facilitator well-versed in emotional process therapy (EPT). The first part of this session requires the participant to engage in metacognition, asking them to reflect and assess what their specific beliefs are surrounding the bias, as well as where they believe they learned these beliefs (or, as discussed in Kuhn, 2000, "awareness of the sources of one's knowledge").

Recent research has shown that metacognitive awareness is required in order to create true conceptual change in one's belief system (Klaczynski, 2006). This is due to our tendency to become more certain in our beliefs over time, applying cognitive dissonance and selective focus in order to fortify and strengthen our convictions (Kuhn, 2000; Klaczynski, 2006). A guided and targeted exploration of the origins of these beliefs, and the participant's gradual understanding of how these experiences led to their current belief structures, can begin to unravel any

inconsistencies if administered correctly. Because cognitive dissonance is a strong mental defense system, however, it is important that the facilitators pace participants carefully to ensure that they feel safe and ready to challenge their own thought processes.

To assist in combating this tendency, the facilitator will be continually prompting the participant to engage the "as-if" body loop as a means of self-reflection and awareness of their emotional state. If the participant is not in touch with and aware of their emotional state, it is easier for them to engage in a more disassociated form of continued cognitive dissonance. Additionally, the foundation of emotional processing theory states that the modification of one's beliefs cannot occur unless the emotional component has been activated (Gillian & Foa, 2011). Therefore, it is critical for both the understanding goals as well as the overall outcome that the participant be consistently engaging in emotional awareness exercises. In this way, the mental exercise of metacognition and the emotional “as-if” body loop actually work together to help the participant remain connected to their emotions and aware of the thoughts surrounding their beliefs, setting the optimal conditions for self-reflection and eventual change.

Once the source events or experiences have been identified, the facilitator will guide the participant in an exploration to establish understanding of the emotional component of the biased belief, by helping them identify and explore the emotions surrounding it. The theory behind EPT leans heavily on the existing learning frameworks of classical fear conditioning, as described by Foa & Kozak (1986). It is therefore necessary to activate the fear in order to uncover the structure of the emotion, so that the inconsistent thoughts and/or beliefs contained within can be exposed. It was once believed that fear extinction was merely an eradication of the existing fear response through repetition; however, it is now theorized that true fear extinction requires the

formation of new stimulus-response patterns to the trigger, and this concept is one of the central tenets of EPT (Gillihan & Foa, 2011).

At this point in the process, the facilitator will ask the participant to utilize the previously learned "as-if" body loop technique to replay an experience where their bias was triggered in order to activate this fear. Once this has been done, the facilitator will utilize aspects of the conceptual change framework to help the participant explore the thoughts and feelings surrounding this fear response. As per the conceptual change model, the facilitator will first seek to highlight any existing dissatisfaction with the existing biased belief. This will be accomplished by examining any counter or contradictory data the participant might have, as well as any inconsistent or non-integrous belief structures they may not have yet considered. For example, if a person is biased towards Asians because they were told as a child that all Asians are violent, they might be asked to think of any Asian they have met who was not violent, or maybe even an Asian friend they have had. Perhaps they have read many things about Asians that don't match with their biased belief. Highlighting the factual incongruences of their bias can help foster this dissatisfaction and set the foundation for a more motivated exploration. Considering the automatic nature of implicit bias, it is likely that the individual has never truly examined the integrity or the origin of their beliefs until now, and so this exploration can be very enlightening for the participant.

The facilitator will also help the participant understand and embrace the reasons that they might have adopted the biased belief due to the emotional component as well as other factors (social pressure, lack of accurate information, etc). The facilitator will then further assist the participant in creating plausibility for adopting a non-biased perspective, through examining accurate information as well as the irrationality of the emotionally-driven aspects of the belief.

The facilitator will continue to prompt the participant to revisit the stimulus throughout, exploring until the response is neutralized. This will allow the participant to effectively re-encode the stimuli with different thoughts and feelings, following the model of fear extinction through replacement.

The ultimate goal of this section is to reach a point where the participant can imagine, on both a mental and emotional level, encountering the biased group while maintaining a neutral or positive state. Once this is complete, the participant will then take the final step in the program, which is to actually engage in a real-world experience with the biased group.

**Section 4: Real-World Testing**

This step is required in order to test if there is any other emotional response remaining that was not experienced in the "as-if" body loop; as Damasio emphasizes in his work, this body loop is merely a "simulation" of an actual event – of the real body loop that happens when our emotions are triggered – and therefore may not account for feelings that would present in real time (Damasio, 2010). Here, the participant will utilize the previously-learned skill of conscious metacognition to be aware of any physiological changes or emotional responses that occur in this real-world event. If the participant observes any significant response, this would then indicate to the participant that further exploration would need to be done. Participants would then re-engage in another facilitator-led exploration (Section 3) in order to address any emotional or belief-driven challenges that remain, re-testing themselves again upon completion.

To close the loop, the participants will re-take the Implicit Association Test. Ideally, they will experience a shift in their implicit responses to the previously-biased group that will reflect the perceptual and emotional changes created as a result of the IBTP.

**Assessment**

This program begins with an assessment protocol that will establish a baseline measurement of the participant's biases. This measurement utilizes both a self-assessment tool - a questionnaire asking the participant what they believe their biases to be, and specifically what biases they would like to remove - and an external measurement tool, namely Greenwald, Banaji, and Nosek's (1995) Implicit Association Test (IAT). The purpose of the questionnaire is to gain insight into the participant's level of self-awareness as it pertains to their bias, as well as determining which bias they want to address in this program.

Additionally, the answers from the questionnaire determine which version of the IAT the participant should then take; the IAT is designed with different versions that target specific biases. For example, there are versions that assess gender bias, racial bias, religious bias towards various groups, etc. This test is specifically designed to measure the more implicit, i.e. unconscious, aspects of an individual's bias in terms of the strength of the bias, a measurement that may differ from the participant's own self-assessment. This difference is to be expected, as this signifies the fundamental distinction between the more conscious, explicit aspects of bias and the more automated, implicit ones (Banaji & Greenwald, 1994). This initial assessment also serves as the baseline for each participant; when they have completed the program, they will take the IAT again. The results of the post-program test will be an external measure of the emotional and perceptual changes participants will experience throughout the program.

The nature of the overall goals of the program - to help participants transform their implicit bias - requires that the participants do a great deal of the assessment of their progress themselves. While the effects of implicit bias might be observed through external behaviors, the ongoing process of measuring the transformation of one's feelings and inherent beliefs requires internal benchmarks to be put in place. Built into the structure of the program, therefore, are intermittent points of self-reflection and self-assessment, where the participant utilizes metacognition and Damasio's "as-if" body loop to assess where they are at on an emotional and intellectual level about themselves and their biased beliefs. Conceptually, self-assessment protocols are also appropriate in context of this program, which fundamentally requires participants to be personally motivated to remove their biases. Self-assessment can promote a sense of personal agency and self-efficacy, which can then act as a powerful motivator for participants in completing their goals (Panadero, Jonsson, and Botella, 2017; Bandura, 2001).

The program is set up for the participant to undergo a self-assessment exercise at the end of each section to determine if they have met the understanding subgoals required. This measurement will determine whether they need to repeat the section or if they are ready to move on to the next one. These self-assessments are prompted by the facilitator, who directs the participant to self-reflect on how they are feeling emotionally and what they are thinking. Once the individual has done an internal self-assessment, they will then describe their reflections to the facilitator, who will record the results in the online portal. These outcomes are recorded as a means of tracking the participant's progress, as well as documenting any outstanding challenges that will be addressed in the next session. While the facilitator is present to observe the participant's responses and offer objective feedback (for example, if the participant states that they are no longer feeling angry, yet appear angry in their physicality or their language), self-

assessment is the primary tool of measurement of a participant's progress towards the overall understanding goal, as well as the subgoals.

Because self-assessment is so central to knowing where a participant is in their progress, as well as determining their next steps, it is a critical part of the program design that the participants are trained in the type of self-assessment strategy they will be expected to employ throughout; this facility comprises one of the early subgoals of understanding. In the first instructional session (Section 1), facilitators will first give participants a conceptual framework for both metacognition and the "as-if" body loop for emotional engagement. However, while the acts of metacognition and emotional "simulation" (Damasio, 2010) are ones that most adults engage in automatically, it may be difficult for participants to initially trigger these internal processes on cue. Therefore, it is also important for the facilitators to then take participants through several exercises where they can practice consciously engaging these mechanisms, and give them feedback as needed in order to assist them in gaining a basic facility with these tools. In this way, the facilitator's main role here is to provide scaffolding for the participants as they build their self-assessment skills.

To this end, it is also important to consider the limitations of self-assessment, as we are often unreliable perceivers of ourselves due to issues of self-image as well as a general lack of awareness (Baumeister, 2005; Dunning, 2005; Leary, 2004; Mabe & West, 1982; Podsakoff & Organ, 1986, as cited in Taylor, 2014). Therefore, the observations of the facilitator are not insignificant; because emotional awareness is so critical to the overarching understanding goal of transforming one's implicit bias, it is important to ensure that the participants are skilled in not only connecting to their current emotional state, but also in being able to assess it in relationship to previous states.

If this ability is not present, the facilitator's feedback can be valuable in revealing to them where they lack this awareness. For example, if a facilitator notices an emotional incongruence between a participant's answers and their expression (as in the example described above), the individual might still be experiencing areas of emotional discomfort that they are unable or unwilling to address. This would lead the facilitator to assess what the participant might require next; this could come in the form of guiding the participant towards repeating one of the previous sections or practices (such as the self-compassion exercise, for example) to address this resistance, or in taking them through an exercise of practicing emotional connection with a more benign example, one that can be used as a stepping stone towards addressing the more difficult emotion.

The final assessment is one in which the participant actually engages with the biased group in a real-life situation; this assessment is a true measurement of whether the participant has achieved the overall understanding goal - to transform their implicit bias. This assessment is also the first one that happens outside of the simulated experience that the participant has been utilizing over the course of the program, and therefore it may also reveal additional aspects of the bias that have previously remained dormant, due to the limitations of the "as-if" body loop: "As-if patterns cannot possibly feel like the body-looped feeling states because they are simulations; not the genuine article, and also because it is probably more difficult for the weaker as-if pat-terns to compete with the ongoing body patterns than for the regular body loop versions to do so" (Damasio, 2010). If it is found that the participant does indeed retain previously unaddressed aspects of their bias, they will be guided to re-engage with a section of the program that can assist in examining these aspects. Once the participant believes they are ready - i.e. when they have assessed that the emotion seems effectively neutralized - they can repeat this real-world

assessment. It is expected that this process may need to be repeated several times, and the program is designed to allow participants to do so as often as necessary to achieve their goal.

**Learners and Learning Challenges**

Recent research has emphasized that only low-prejudice people are likely to ever transform their prejudices; in order to change these automatic stereotypes, Devine and colleagues argue, we must not only be aware that they exist, but also have the motivation and desire to transform them (Devine & Monteith, 1993; Plant & Devine, 2009, as cited in Devine et al, 2012). Therefore, a prerequisite for all participants of the IBTP is that they are all self-selected volunteers, individuals who possess the desire to become more tolerant and a willingness to do the work in order to remove their personal biases.

As they engage with the program, this particular cohort will face a variety of learning challenges. One of the primary challenges participants will face relates to the negative emotions that will arise as they face aspects of their bias, emotions that can hinder them from achieving their goals in the program. It is well documented that many individuals – specifically, those considered to be "low-prejudice" – often have a negative response to the information of their bias (Casad et al, 2013; Devine & Monteith, 1993; Stevens & Abernethy, 2018). This negative emotion tends to manifest as shame; unfortunately, the emotion of shame is not a common motivator for people. Instead, it generally inspires them to want to deny or avoid the thing that causes them to feel shame (Stevens & Abernethy, 2018) – the antithesis of motivation. Given that the target population for this intervention would inherently fall into the "low-prejudice" category, it is essential to create contingencies for shameful reactions that might occur.

Therefore, the IBTP seeks to proactively address this challenge in two ways: Through conceptual change and self-compassion.

In Section 1, the facilitator will methodically set the conditions for conceptual change based on Strike and Posner's (1985) model in order to assist participants in changing their perception on bias in general. The goal in this section is to help participants be kinder and more understanding with themselves as they discover their own biases, and to realize that their implicit responses and feelings are inevitable aspects of being human instead of a verdict on their humanity. This perceptual shift is essential for participants to be able to move forward.

The first condition of conceptual change, dissatisfaction with existing conditions, will be addressed by directly engaging with the participants about their feelings surrounding their bias. Without a framework of understanding of how bias is natural - rather than something that is to be judged or criticized - many participants will experience a certain level of discomfort from the discovery/reaffirmation of their bias after completing the IAT. The facilitator's objective here is first to openly address this discomfort, and then begin to lay the framework for the second condition, minimal understanding.

Minimal understanding will be presented by the facilitators first by giving a thorough description of the cognitive process of categorical thinking and why it is necessary for humans. Next, they will demonstrate how that natural process creates bias through a description of Tajfel's social identity theory (1979), its steps of categorization, identification, and comparison, and how this creates the in-group/out-group paradigm—a construct developed as an adaptive strategy for survival. Facilitators will then elaborate on why this way of thinking was evolutionarily advantageous for early humans. The goal of presenting this information is to create a logical construct that the participants can not only easily digest but also will be more

likely to accept, due to the well-researched amount of information that has its underpinnings in scientific principles.

In order to satisfy the third condition, plausibility, the facilitator will draw analogies between historical humans and their need for in/out group thinking and how this behavior play out in today's world. Examples will be given to illustrate this point - such as how people became more biased towards Muslims after 9/11, the Israeli-Palestinian conflict, etc., as well as more benign examples, such as how this mechanism plays out with fans of sports teams. Facilitators will also talk about the need for humans to belong, to feel like they are part of a group, and how this tendency also lends itself towards biases (Tajfel, 1979). The facilitator will also lead a discussion with the participants about this information, asking them to come up with examples of their own of where they have seen this sort of implicit, natural tendency in themselves or others around them. This discussion will help participants demonstrate how much they have integrated the material presented so far.

The final condition, fruitfulness, will be satisfied by a final discussion between the facilitator and the participants about how the material specifically relates to them. Specifically, the facilitator's objective is to guide a discussion to help participants see the benefits of this new perspective in terms of being able to address their specific biases. This understanding will give participants the incentive they need to accept and integrate this information, therefore creating the conceptual change they need to move on in the program.

For many participants, this new perspective will alleviate the shame or guilt that surrounds the uncovering of their implicit bias. However, it is likely that others may still have lingering feelings that could hinder further explorations, and it is important that this program has another set of contingency tools in place to overcome this learning challenge. To this end, for

those that need it, the processes in Section 2 offer a series of tools based on the principle of self-compassion to help them embrace the parts of them that have created the bias.

Self-criticism and shame activate the same threat-detection system that we employ when we encounter danger, triggering our fight-or-flight mechanisms and gearing ourselves up to defend against our attacker (which, in the case of self-criticism, is ourselves) (Neff, 2019). This body response then puts us in an angry or fearful emotional state, neither of which is conducive to learning and introspection; the action tendencies of these emotions are to narrow our focus and attend to danger, not to broaden and build our knowledge (Fredrickson, 2001). Self-compassion, on the other hand, activates our care system, which counteracts the threat response and allows the individual to be able to re-approach themselves as well as be in a more receptive state for learning.

Recent research has found that self-compassion exercises consistently increase motivation for self-improvement and learning, even in areas where the individual feels they have committed a moral transgression (Brienes & Chen, 2012). This program utilizes specific exercises from The Self-Compassion Workbook (Neff, 2019) in order to assist participants in removing their shame – so that they can move forward to explore their biases with (relative) ease.

Another learning challenge is in the use of metacognition as a means of self-assessment, which is a foundational component of the program. As described in Perkins, Simmons, & Tishman (1990), one difficulty presented by utilizing metacognition is the cognitive load that it adds to a task. Especially considering that a key component of the participant's self-assessment is an exercise in self-reflection - one that requires them to be aware of not only their cognitive

process but their emotional state - it is important to lessen the cognitive load as much as possible in order to make room for a genuine and emotionally connected experience.

One way that this program seeks to lessen the cognitive load is by incorporating automation (Perkins, Simmons & Tishman, 1990). This is achieved in Section 1, where the concept of metacognition is introduced; here, the facilitator takes participants through a series of short metacognitive exercises to first introduce them to the idea of consciously triggering this process and then help them tokenize it. In this way, the goal is that participants will become adept enough at triggering metacognition in their self-assessments that it will become more automatic.

**Critique**

The design of the IBTP is intentionally time-and-resource intensive in order to achieve maximum success in removing bias. However, this intensity might prove to be problematic in terms of the practicality of actual real-world implementation. First, organizations or companies might balk at the cost and time required, and would need to have a deep and authentic value of tolerance in order to be willing to invest in creating such an opportunity. Additionally, due to the complexities of the model, it may be difficult to find or train facilitators that have the necessary skills required to adequately help participants process their biases. The current model design implies that the program will be carried out using a single facilitator, but it may be unrealistic to expect a single individual to be able to sufficiently train participants in all of the concepts necessary – implicit bias, metacognition, the "as-if" body loop, shame, and self-compassion – as well as have the therapeutic background to engage in emotional processing therapy methods.

One potential solution for this might be to create a team of facilitators, each with their own specialty, who will work together with each group of participants.

Further, these same complexities require the facilitators to keep track of many moving parts within a specific cohort, as each individual will likely be at different stages of the program as they move through it. While hinted at in the prototype, the program currently does not have a well-designed system in place to assist facilitators in this way, and would be essential to any actual implementation. These logistics would need to be coordinated by an administrator; these logistics have not as yet been clearly developed and the role of the administrator would need to be more clearly defined.

Along these lines, it is possible that the timeline and content in the model is too ambitious for most participants; each section, as currently designed, requires them to undergo significant perceptual shifts in not only understanding external concepts, but also themselves, in a single short session. Participants might realistically need more time in between sessions, or multiple practice sessions after each block of learning, in order to process and integrate the material in the way that the understanding goals require. Future design considerations might therefore include post-section practice sessions or workshops that participants can attend with a facilitator.

Another limitation of the existing model is the potential for attrition of participants. Because the design is as such that participants may be required to repeat sections several times, is easy to imagine that they might become frustrated if they find themselves unable to shift their perspectives or feelings sufficiently. The current design does not have contingencies for this, but it is an important consideration for future iterations. Such contingencies might include a targeted session to address the specific frustrations, or for the participant to work with a different facilitator altogether, one who might be able to assist them in a different way. Even the best

facilitators and therapists don't work for everyone, and it is important to consider that in order to best support the participants in achieving their goals.

Participants might also become discouraged if they feel that they have conquered their biases in their "as-if" simulations, yet continue to experience emotional reactions in their real-world testing. The IBTP does not currently address how to manage any ongoing discrepancies in this arena, and would require further research into a solution for how to do so.

**Conclusion**

The Implicit Bias Transformation Program takes a revolutionary approach to addressing implicit bias by working with individuals in a targeted, methodical sequence of interventions designed to facilitate fundamental change. By combining the tools of conceptual change, metacognition, self-compassion, and emotional processing therapy, the IBTP comprehensively and compassionately guides its participants through the difficult task of uncovering, exploring, and ultimately resolving their biases – creating a team of genuinely tolerant, evolved, and compassionate humans who can incrementally make the world a better place through their new understanding.

References

Allport, G. (1954). *The nature of prejudice* (Unabridged, 25th anniversary ed.). Addison-Wesley Pub.

Banaji, M. R., & Greenwald, A. G. (1994). Implicit stereotyping and unconscious prejudice. In M. P. Zanna & J. M. Olson (Eds.), *The psychology of prejudice, The Ontario Symposium* (Vol. 7, pp. 55-76). Erlbaum.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52. 1-26

Breines, J. G., & Chen, S. (2012). Self-Compassion Increases Self-Improvement Motivation. *Personality and Social Psychology Bulletin*, 38(9), 1133–1143.

Casad, B., Flores, A., & Didway, J. (2013). Using the Implicit Association Test as an Unconsciousness Raising Tool in Psychology. *Teaching of Psychology*, 40(2), 118-123.

Cottrell, C., & Neuberg, S. (2005). Different Emotional Reactions to Different Groups: A Sociofunctional Threat-Based Approach to "Prejudice". *Journal of Personality and Social Psychology*, 88(5), 770-789.

Damasio, A. (2010). *The self comes to mind: Constructing the conscious brain* (Chapter 5, pp. 115-138). Vintage Books.

Devine, P. (1989). Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology*, 56(1), 5-18.

Devine, P., & Monteith, M. (1993). Chapter 14 - The Role of Discrepancy-Associated Affect in Prejudice Reduction. In Mackie, D., & Hamilton, D. (Eds.), *Affect, Cognition and Stereotyping* (pp. 317-344). Elsevier Science.

Devine, P., Forscher, P., Austin, A., & Cox, W. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal Of Experimental Social Psychology*, 48(6), 1267-1278.

Forscher, P., Lai, C., Axt, J., Ebersole, C., Herman, M., Devine, P., & Nosek, B. (2019). A Meta-Analysis of Procedures to Change Implicit Measures. *Journal of Personality and Social Psychology*, 117(3), 522-559.

Fredrickson, B. (2001). The Role of Positive Emotions in Positive Psychology. *American Psychologist,* 56(3), 218-226.

Gillihan, S., & Foa, E. (2011). Fear Extinction and Emotional Processing Theory. In Schachtman, T., & Reilly, S. (Eds.), *Associative Learning and Conditioning Theory* (Chapter 2). Oxford University Press.

Greenwald, A., Benaji, M., and Kosek, B. (1995). *The Implicit Association Test* [Measurement instrument]. Retrieved from https://implicit.harvard.edu/implicit/takeatest.html.

Johnson, E., and O'Brien, K. (2013). Self-Compassion Soothes the Savage EGO-Threat System: Effects on Negative Affect, Shame, Rumination, and Depressive Symptoms. *Journal of Social and Clinical Psychology*: Vol. 32, No. 9, pp. 939-963.

Karris, M., & Caldwell, B. (2015). Integrating Emotionally Focused Therapy, Self-Compassion, and Compassion-Focused Therapy to Assist Shame-Prone Couples Who Have Experienced Trauma. *The Family Journal*, 23(4), 346-357.

Klaczynski, P. (2006). Learning, Belief Biases, and Metacognition. *Journal of Cognition and Development*, 7(3), 295-300.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science*, 9(5), 178-181.

Martinčeková, L., & Enright, R. (2018). The effects of self-forgiveness and shame-proneness on procrastination: Exploring the mediating role of affect. *Current Psychology*, 1-10.

Neff, K. (2003). Self-Compassion: An Alternative Conceptualization of a Healthy Attitude Toward Oneself. *Self and Identity*, 2(2), 85-101.

Neff, K. (2011). Self-Compassion, Self-Esteem, and Well-Being: Self-Compassion, Self-Esteem, and Well-Being. *Social and Personality Psychology Compass*, 5(1), 1-12.

Neff, K., & Germer, C. (2019). *The Mindful Self-Compassion Workbook : A Proven Way to Accept Yourself, Build Inner Strength, and Thrive*. Guilford Publications.

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98.

Paluck, E., & Green, D. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology*, 60(1), 339-367.

Perkins, D., Simmons, R. & Tishman, S. (1990). Teaching cognitive and metacognitive strategies. *Journal of Structural Learning*, 10(4), 285-303.

Perkins, D. (1998). What is understanding? In M.S. Wiske (Ed.), *Teaching for Understanding: Linking research with practice* (pp. 39-57). San Francisco, CA: Jossey-Bass.

Stevens, F., & Abernethy, A. (2018). Neuroscience and Racism: The Power of Groups for Overcoming Implicit Bias. *International Journal of Group Psychotherapy*, 68(4), 561-584.

Strike, K.A. & Posner, G.J. (1985). A conceptual change view of learning and understanding. In L.H.T West and A.L. Pines (Eds.) *Cognitive Structures and Conceptual Change* (pp. 211-231). Academic Press, Inc.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56-65.

Zhang, H., Carr, E., Garcia-Williams, R., Siegelman, A., Berke, G., Niles-Carnes, A., . . . Kaslow, L. (2018). Shame and Depressive Symptoms: Self-compassion and Contingent Self-worth as Mediators? Journal of Clinical Psychology in Medical Settings, 25(4), 408-419.