Exploring the Relationship Between Affect and Bayesian Inference

Ethan Gray (GRYETH001) and Gemma Strohbach (STRGEM001)
Supervised by Dr. Donné van der Westhuizen and Prof. Mark Solms

Word count: 7997

PLAGIARISM DECLARATION

1. We know that plagiarism is wrong. Plagiarism is to use another's work and to pretend that it is one's own.

2. We have used the *American Psychological Association (APA)* convention for citation and referencing. Each significant contribution to, and quotation in, this report from the work, or works, of other people has been attributed, and has been cited and referenced.

3. This essay report is our own work.

4. We have not allowed, and will not allow, anyone to copy our work with the intention of passing if off as his or her own work.

5. We acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is our own work.

SIGNATURES: Ethan Gray and Gemma Strohbach

Abstract

The role of emotion in belief generation and updating has received insufficient attention in theories of Bayesian inference. Recent literature in the field suggests that the processing of incoming information and its integration into existing beliefs is crucially influenced by one's current emotional state. However, there exists a significant gap in the empirical understanding of how this process occurs. Thus, this study sought to investigate the influence of affect and alexithymia on Bayesian belief updating using the novel Bias Against Disconfirmatory Evidence (BADE) task. Alexithymia was investigated in addition to affect, as it is characterized by blunted emotional awareness and, thus, may influence how affect is processed in updating beliefs. An online survey was administered to 86 participants to measure current emotional state, alexithymia, and integration of evidence into existing beliefs using the BADE task. No significant relationship was observed between prior probability of emotional hypotheses and alexithymia ($p = .11$) or positive and negative affect ($p = .26$). Negative affect and alexithymia could not significantly predict integration of confirmatory evidence ($p = .15, .27$) or disconfirmatory evidence ($p = .16, .26$). A weak positive relationship was observed between positive affect and integration of disconfirmatory evidence ($r = .26$, $p = .02$). These results may indicate that the BADE task is not suitable to quantify the impact of affect on belief updating and thus, cannot disprove that affect and alexithymia may significantly influence the generation and updating of Bayesian beliefs. Implications of this are discussed, including recommendations for future research.

Keywords: affect, alexithymia, Bayesian inference, Bias Against Disconfirmatory Evidence task, emotion

Bayesian predictive coding is understood as the primary neurocomputational mechanism that enables perception and inference about the external environment (Clark, 2013; Friston & Kiebel, 2009; Hohwy, 2013; Hohwy et al., 2008). Within this framework, the brain is understood as a hypothesis-generating machine that continuously forms predictions, referred to as priors or beliefs, to actively anticipate what sensory input should be if certain environmental events occur (Hohwy, 2013). The discrepancy between prior predictions and actual sensory input is quantified as *prediction error* and is processed to facilitate effective updating of priors to better match sensory input (Mumford, 1992). This allows neural processing to become more efficient, as the brain only needs to process unpredicted information. Selecting the most accurate hypothesis out of an array of competing alternatives is therefore a key challenge for the brain's perceptual systems. In Bayesian terms, the likely accuracy of priors is referred to as their *posterior probability*, and is determined by both the degree to which they match sensory input (*likelihood*) and their probability, based on knowledge of the environment, irrespective of current sensory input (*prior probability*) (Friston, 2002). Priors with the highest posterior probability can minimise prediction error by successfully inferring environmental causes of alterations in the sensory apparatus of the brain and, thus, determine the perceptual content of the system (Friston, 2002; Kersten et al., 2004).

Within this framework, the brain refines its predictive models by hierarchically updating prior beliefs in response to prediction errors when presented with contradictory sensory input. (Mumford, 1992). Crucially, the ability of prediction errors to update priors depends on the inferred confidence placed in sensory input, called *precision* (Friston, 2008). New information with low precision is regarded as noise and requires perceptual inference to be guided disproportionately by high-precision priors (Hesselman et al., 2010). For example, diminishing light levels at dusk render sensory information less reliable and thus, priors are more heavily weighted in determining the perceptual scene in this context. Conversely, attention can be recruited to increase precision assigned to sensory input, consequently reducing the precision of priors, allowing priors to be updated in response to prediction errors (Feldman & Friston, 2010; Hesselman et al., 2010). Priors are thus updated until their posterior probability is deemed high enough to accurately explain sensory input (Friston & Kiebel, 2009).

According to Friston (2010), this inferential process is centrally driven by the minimization of 'free energy': statistical uncertainty that arises when an organism must infer

external states from sensory states, reflecting the mismatch between predictions and incoming signals (Friston, 2010; Joffily & Coricelli, 2013; Seth & Friston, 2016). Because unexpected changes in external environments threaten the stability of an organism's internal state, the free energy principle states that all self-organising biological systems aim to minimize free energy by effectively modelling anticipated changes (Friston, 2010; Friston, et al., 2012). This results in a necessary imperative of organisms to act upon their external environment through 'active inference' to ensure that sensory input better matches predictions (Solms & Friston, 2018). Within Bayesian theory, free energy is correlated with the sum of squared prediction errors weighted by their estimated precision (Solms & Friston, 2018). Thus, by optimizing the precision assigned to prediction errors during the inferential process, the brain can either update its priors to better match incoming signals or disregard incoming signals as noise, both of which result in minimization of free energy (Solms & Friston, 2018).

However, the adaptive value of a selected strategy for minimising free-energy depends upon current context (Solms & Friston, 2018; Friston, 2010). For instance, psychological health requires one to reduce sensory precision in order to tune out and habituate to noise. However, if new information is inaccurately disregarded as noise, this may have deleterious consequences for survival. Thus, optimal assignment of precision to prediction errors is crucial for adaptive functioning. In novel contexts (i.e., under states of high uncertainty) precision must be assigned to prediction errors based on expected reliability (Badcock et al., 2019). Consequently, Solms and Friston (2018) argue that changes in expected free energy must be subjectively experienced as *valanced affect* whereby deviations away from homeostatic setpoints, associated with increased free energy, are felt as negative affect, while returning towards these set points elicits positive affect (Damasio, 2010; Pezzulo, et al., 2015).

Thus, affect corresponds with the measurement of free energy to guide optimization of precision assigned to priors and sensory input in perpetual and active inference (Solms & Friston, 2018). Specifically, under conditions of uncertainty, negative affect will increase as prior precision decreases, prompting the system to increase precision assigned to sensory input (Barrett & Satpute, 2013; Clark, 2013; Joffily & Coricelli, 2013). Once homeostatic equilibrium is restored, positive affect is experienced so that adaptive response may be reinforced for subsequent precision estimates (Solms & Friston, 2018).

Thus, emotional feelings fundamentally arise in response to changes in implicit expectations about the reliability of generative models versus sensory input. This theory is aligned with other authors who consider affect to be a primordial instinctive system that drives learning of adaptive responses (Panksepp, 1992; Solms & Turbull, 2018; Barrett, 2017). This relationship between precision optimization and affective functioning has also been presented as a Bayesian explanation of hypervigilance triggered by uncontrollable environmental conditions (Edwards et al., 2012). Stress-related feelings have been conceptualized by Peters et al. (2017) in terms of an enduring "state of uncertainty about what needs to be done to safeguard physical, mental or social well-being" (p.184). Under stress, precision is therefore thought to be heavily weighted towards sensory evidence, explaining the state of hypervigilance.

Consequently, the functioning of the brain's affective system should significantly influence how incoming information is processed to form and update beliefs about the external environment. Affect holds immense adaptive value as it reinforces interactions with the environment that resolve homeostatic and social needs, resulting in the learning of behaviours in the context of distinct environments that promote survival (Panksepp, 2000; Solms & Turnbull, 2018). Similarly, Panksepp (1992) argues that basic emotions may be understood as unique action-orientated brain states that instinctively impel organisms towards the satiation of homeostatic and social needs. All affects are, thus, valanced experiences that determine salience and guide action and decision-making in contexts of uncertainty (Solms & Turnbull, 2018; Solms & Friston, 2018). Affects are therefore the brain's strategy for resolving uncertainty related to predicted consequences in the decision-making process in ambiguous settings where acquired knowledge offers no advantage. In the Bayesian framework, this implies that affect plays a key role in the selection of appropriate hypotheses in uncertain situations by allowing an individual to feel through the optimization of precision assigned to competing priors and incoming sensory input. In other words, the experience of valenced affect determines which model is preferentially selected and the weight that a selected predictive model carries in proportion to incoming sensory evidence.

While emotions are considered an intrinsic and conditioned adaptive response to resolve uncertainty generated by ambiguous environments, optimal functioning of the Bayesian brain requires that predictions are flexibly revised when faced with conflicting information (Fotopolou, 2013). More specifically, automatic emotional responses may not always be optimal approaches

for reducing free energy (Solms & Friston, 2018). Precision assigned to predictive models needs to be flexible so that expectations can align to reality. Consequently, recent research has attempted to conceptualize various forms of psychopathology in terms of aberrant precision processing and updating of prior beliefs (Smith et al., 2020). For example, delusions in psychotic disorders may be understood as abnormal inference processes that result from either errors in sensory processing or abnormal emphases on prior expectations (Pfuhl, 2017) which prevent the adjustment of false beliefs and ability to update one's Bayesian model (Buck et al., 2012; Coltheart, 2007; Moritz & Woodward, 2006). Similarly, depressive disorders have been characterized by persistent negative beliefs about oneself and one's environment that distort how incoming information is processed and how precision is assigned in relation to current priors (Chekroud, 2015; Barrett et al., 2016). This emerging literature suggests that emotional affect may influence the assignment of precision to priors and in cases of psychopathology, impair optimal updating of beliefs during Bayesian inference.

Despite evidence for the role of affect in influencing the processing of sensory information (Solms & Turnbull, 2018; Flykt, 2005; Frischen et al., 2008; Öhman et al., 2001) and motivational behaviour (Brown & Pluck, 2000; Lang, 1995), there exists a significant gap in empirical understanding of exactly how individuals use affect to select hypotheses and revise beliefs within a Bayesian framework. In fact, research has largely presented emotional reasoning as a psychopathological heuristic (Arntz et al., 1995; Gilbert, 1998), despite the enormous role of emotion in adaptive learning. More specifically, there has been an absence of empirical investigation into how subjectively experienced emotion influence the process of optimizing precision assigned to prediction errors.

Of particular importance in this regard is *alexithymia*, a non-clinical trait characterized by a disturbance in emotional awareness, manifesting as difficulties in describing and identifying one's emotional states and differentiating internal feelings from bodily states of emotional arousal (Nemiah et al., 1976; Nicholson et al., 2018). Alexithymia has been observed among a significant proportion of patients suffering from various psychopathologies such as depressive disorders, substance use disorders, and eating disorders, amongst others (Taylor, 2000). The distorted emotional regulation observed within alexithymia also results in maladaptive behavioural responses to emotional stimuli and thus, impairs the function of emotions to facilitate conditioned adaptive responses (Pandey et al., 2011). Furthermore, since Solms and

Friston (2018) argue that the optimization of precision assigned to prediction errors is subjectively felt to enable emotional learning in new environments, the blunted emotional awareness that characterizes alexithymia potentially offers significant insight into understanding the role of affect in decision-making and the Bayesian updating of beliefs. Notably, alexithymia reflects a broad deficit in both affective and non-affective interoception (Brewer et al., 2016) suggesting it may interfere with the optimal updating of precision in emotional settings. Thus, further investigation is required into how alexithymia may influence the optimization of precision assigned to prediction errors within Bayesian inference.

If the Bayesian account of affect by Solms and Friston (2018) is correct, affective states should significantly influence prior precision assigned to competing hypotheses under conditions of uncertainty and the consequent updating of these precision assignments in response to incoming information. This study employed the *bias against disconfirmatory evidence* (BADE; Woodward et al., 2006) task to present participants with competing interpretations of progressively less ambiguous scenarios and examine how affective variables contributed towards the inference process.

Based on the notion that emotions function to guide precision assignment under conditions of uncertainty, in entirely ambiguous scenarios, participants with healthy emotional awareness (as indicated by lower alexithymia scores) were expected to assign greater prior precision to hypotheses that carry emotional valence. Conversely, participants who scored higher on alexithymia were not expected to preferentially prioritize emotional hypotheses in ambiguous scenarios due to purportedly decreased emotional awareness. It was further expected that prior probability estimates for emotional hypotheses would vary as a function of individuals' baseline emotional state. Specifically, because negative affect drives the system towards the reassignment of precision under states of uncertainty, participants with higher negative affect scores should place less confidence in prior beliefs, reflected in lower prior precision assigned to emotional hypotheses. Conversely, because positive affect signals that one's model of the ambiguous external environment is successfully minimizing free energy, participants with higher positive affect scores were expected to assign precision to prior beliefs more liberally, reflected in higher prior precision assigned to emotional hypotheses.

It was further hypothesized that participants higher on the alexithymia spectrum would be less likely to revise their beliefs when presented with information conflicting prior beliefs. This

is because the blunted emotional awareness associated with alexithymia was expected to impair the measurement of free energy generated in response to prediction errors; therefore limiting motivation to reassign higher precision to incoming information and update beliefs. Similarly, because positive affect reflects the successful minimization of free energy, it was expected that participants with higher positive affect scores would be less likely to update prior beliefs even when presented with information that conflicted those beliefs. Conversely, it was expected that participants with higher negative affect scores would allocate greater attention and precision to incoming sensory evidence, particularly that which conflicts prior beliefs and thus, would exhibit an exaggerated updating of beliefs in response to conflicting information. (Crucially, however, because adaptive emotional functioning requires flexibility in precision assignment, a healthy sample may optimize the precision assigned to prior beliefs in ways that most accurately correspond to changing sensory evidence. Thus, these effects may be insignificant in participants without pathologically distorted emotional state.) Thus, this study aims to explore how affective states and awareness of these states influence the Bayesian updating of beliefs under conditions of decreasing uncertainty.

Research Aim and Question

This study aimed to explore how positive and negative affective states and alexithymia trait scores influence prior precision assigned to competing hypotheses under conditions of uncertainty and the updating of beliefs in response to disambiguating information in the BADE task.

Hypotheses

1. There will be a significant positive correlation between negative affect scores and alexithymia trait scores.
2. Precision assigned to competing hypotheses in the BADE task will differ significantly on account of the emotional valence of the interpretation and the level of ambiguity of the scenario.
3. There will be a significant positive correlation between the decrease in precision of disconfirmed hypotheses and the increase in precision of confirmed hypotheses in the BADE task.
4. There will be a significant negative correlation between prior probability estimates of emotional interpretations in the BADE task and alexithymia trait scores.

5. Prior probability estimates of emotional interpretations in the BADE task will be significantly higher amongst participants with higher positive affect scores and significantly lower amongst participants with higher negative affect scores.
6. There will be a significant negative correlation between the magnitude of belief updating and alexithymia trait scores.
7. The magnitude of belief updating will be significantly lower amongst participants with higher positive affect scores and significantly higher amongst participants with higher negative affect scores.

Methods

Setting

This study was conducted within South Africa, limiting participants to students at the University of Cape Town (UCT) and other South African residents. Data was collected online using an electronic survey.

Research design

A correlational design was employed to investigate the relationships between primary measures of affect and performance on the BADE task. Participants completed the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), 20-item Toronto Alexithymia Scale (TAS-20; Bagby et al., 1994) and the BADE task (Sanford et al., 2014). The PANAS allowed for estimation of self-reported negative affect and positive affect while TAS-20 quantified alexithymia trait levels which all served as continuous independent variables for statistical analyses. Conversely, the BADE task allowed for a quantification of prior precision assigned to competing hypotheses and the change in precision estimates as more contextual information was presented (ICE scores and IDE scores). These continuous BADE task variables served as both outcome and moderating variables in analysis. The individual and interacting relationships between each of these affective measures and BADE task outcomes was explored through various simple and multiple regression analyses.

Participants

The final sample for this study consisted of 86 participants aged between 18 and 57 (M = 24.5, SD = 9.96) of whom 65 were female (75.6%) and 21 were male (24.4%). Participants were required to be South African residents aged 18 or older. Participants were primarily recruited via non-probability convenience sampling through two research recruitment emails (Appendix A) on

the UCT server sent to psychology students and the broader student body. 74 participants (86%) were UCT students recruited through research advertisements. The remaining 12 participants (14%) were non-UCT students that were recruited through social media platforms, Facebook and Instagram, to expand the sample beyond UCT students and meet minimum sample size. Minimum sample size was set at 75 (n=75) based on correlational sample size calculation with probability of type I and type II errors set at 0.05 (α=0.05, β=0.05) and minimum significant correlation coefficient set at 0.4 (r=0.4) (Hulley, 2013). Beyond age and geographic constraints, exclusion criteria included having a history of a psychiatric or neuropsychological disorder which could distort affect and cognition. Furthermore, participants were required to be fluent in English to ensure sufficient comprehension of BADE task items.

Measures

Affect

PANAS. The Positive Affect and Negative Affect Schedule (PANAS) designed by Watson et al. (1988) is the most broadly used scale for assessing positive and negative affect (Díaz-García et al., 2020). Multiple studies on the structure of affect have shown positive and negative affect to be two dominant facets that are independent of one another (Watson et al., 1988). Thus, the PANAS includes two 10-item scales with one pertaining to negative affect and the other to positive affect. Participants use a 5-point Likert scale to rate the degree to which they are currently experiencing specific positive and negative emotions (Mulder, 2018). The sum of ratings for the 10 positive or negative items provides a score for positive and negative affect respectively. Scores on both positive and negative scales can range between 10 and 50, with higher scores representing higher levels of the respective affect (Watson et al., 1988). The scales have been shown to be uncorrelated, internally consistent, and stable (Watson & Clark, 1994). Watson et al. (1998) also presented discriminant and convergent validity for the scales. Subsequently, clinical and non-clinical studies have found the PANAS to be a reliable and valid tool (Merz et al., 2013).

Alexithymia scale (TAS-20). The 20-item Toronto Alexithymia scale (TAS-20) developed by Bagby et al. (1994) is the most commonly-used measure for quantifying degree of alexithymia across numerous research and clinical settings (Bagby et al., 2020). The TAS-20 presents participants with 20 statements describing behaviour or cognitions typically observed in those with high levels of alexithymia. Participants are then required to rate the degree that each

statement describes them on a 5-point Likert scale. Participants' ratings for each of the statements are summed to provide an overall score between 20 and 100, whereby higher scores reflect higher degrees of alexithymia (Bagby et al., 1994).

The TAS-20 has strong construct validity with factor analysis indicating that difficulty identifying feelings, difficulty describing feelings to others and externally-orientated thinking fundamentally underlie TAS-20 scores (Bagby et al., 1994). Correlation patterns of TAS-20 scores with scores for scales measuring similar constructs supports the convergent and discriminant validity of the TAS-20 (Bagby et al., 2020). Furthermore, the TAS-20 has strong internal reliability with Cronbach alpha values above .8 across numerous samples (Bagby et al., 1994). Retest reliability of the TAS-20 is also significant, supporting the TAS-20 as a stable measure of alexithymia traits (Bagby et al., 2020).

Bayesian updating of beliefs.

BADE task. The BADE task was developed by Woodward et al. (2006) to investigate biases against disconfirmatory evidence within clinical patients with schizophrenia (Sanford et al., 2014). Within a Bayesian framework, these biases refer to the failure to revise prior beliefs despite access to incoming information which contradicts these beliefs (Woodward et al., 2006). The task has been performed and refined several times to ensure greater reliability and validity within clinical and non-clinical samples (Woodward et al., 2006). Within this study, a modified version of the most recent and comprehensive version of the BADE task will be used. The task contains 24 written ambiguous scenarios, and four possible interpretations of each scenario. Two of the interpretations are referred to as 'lures', one emotional and one neutral, and initially seem the most plausible but are disconfirmed once all information is provided. The final two interpretations are the 'true' interpretation which emerges as most plausible once all information is provided, and the 'absurd' interpretation, which remains implausible throughout the task. Participants are required to rate the plausibility of each interpretation along a continuous 10-point scale whereby a rating of zero indicates that the interpretation is implausible and a rating of 10 indicates that the interpretation is extremely plausible. The BADE task requires participants to sequentially rate the plausibility of each interpretation after being provided with three pieces of information about the scenario. Thus, information about the scenario reflects incoming sensory information that renders the scenario progressively less ambiguous (Woodward et al., 2006). Plausibility ratings of each interpretation reflect the probability assigned to competing priors

with a dominant prior being established as the interpretation with the highest plausibility rating at each stage.

The change in plausibility ratings of each interpretation allows for a quantification of the integration of incoming evidence and the precision assigned to this evidence. The absolute change in lure interpretation plausibility ratings between when the first piece of information is presented (emotional lure 1 and neutral lure 1) and the last piece of information is presented (emotional lure 3 and neutral lure 3) reflects a quantification of integration of disconfirmatory evidence (IDE) (Woodward et al., 2006). Because the lure interpretations become less plausible as more information is provided, a significant decrease in lure interpretation plausibility ratings represents greater IDE by assigning prediction errors greater precision (Woodward et al., 2006; Friston, 2008). Conversely, a relatively small decrease in plausibility assigned to lure interpretations reflects a lower level of precision being assigned to prediction errors and thus, the disregarding of disconfirming new information as noise. The difference between final lure plausibility ratings and initial lure plausibility ratings will be calculated for emotional and neutral lures separately for each BADE task item, providing a measurement of the integration of emotional and neutral disconfirmatory evidence respectively. The absolute change in plausibility ratings for each of these lures is then summed across all items to provide an overall estimate of participants' IDE. Conversely, integration of confirmatory evidence (ICE) is quantified as the change in true interpretation plausibility ratings between when the first piece of information is presented (True 1) and the last piece of information is presented (True 3) (Woodward et al., 2006). Because the true interpretations become more plausible as more information is provided, a significant increase in true interpretation plausibility ratings represents greater ICE and greater precision assigned to incoming sensory evidence (Woodward et al., 2006). The change in true interpretation plausibility rating for each item is then summed to provide an overall ICE score for each participant. Thus, by measuring the IDE and ICE, the BADE task allows for the assignment of precision to prior beliefs and incoming information in response to prediction errors to be quantified.

The BADE task was further modified by requiring participants to provide a plausibility estimate for each interpretation before being presented with any information about the scenario, allowing for the prior probability of each interpretation to be measured (Pfuhl, 2017). A higher plausibility estimate of an interpretation in the absence of incoming information reflects high

prior probability assigned to the interpretation (Pfuhl, 2017). The prior probability rating for each of the four types of interpretations was then summed across the 24 items to quantify prior probability that each participant assigned to each type of interpretation. Since the posterior probability of a prior is determined both by its prior probability and degree that the prior matches sensory input (Friston, 2002), the prior probability of interpretations may moderate the degree to which priors update in response to incoming information.

Procedure

An online survey comprised of four sections was constructed using Google Forms (Appendix C). The first section collected basic demographic data including age, gender, and whether participants were UCT students. The following two sections included the PANAS and the TAS-20. The final section included the BADE task which presented participants with instructions for the task and a practice trial item, followed by the 24 experimental trial items that generated scores for the BADE task outcomes of interest. Two distractor items without a clear correct interpretation were interspersed with the experimental items to prevent unified response patterns. Additionally, interpretations for each item were presented in a randomized order at each stage so that the true interpretation could only be deduced from the information provided.

Participants were primarily recruited via a research recruitment email to the UCT student body and, secondarily, through social media platforms, Facebook and Instagram, where potential participants were provided with details of the study, its inclusion criteria and a link to the survey (Appendix A). Willing participants were required to confirm that they met the necessary inclusion criteria for the study and provide consent by digitally agreeing to conditions on an informed consent form (Appendix B). Then, participants could proceed to the survey. When completed survey responses surpassed 80, the survey was closed and the raw data downloaded as a single Excel spreadsheet. This data was then cleaned in Excel to calculate PANAS scores, TAS-20 scores and the BADE task results.

Ethical considerations

Each participant was required to willingly volunteer to participate in the study by visiting the link provided through the recruitment email and media post. Prior to consenting, participants were informed of the nature of the study and their engagement in it (Appendix B). The study posed no significant risks to participants, with the only potential harms being the sacrifice of 45 minutes for participation and minor distress in self-reporting affective states. To ensure

autonomy, participants were free to withdraw from the study at any point, including withdrawing their results from the study after data collection. No monetary reward was offered for participation, thus avoiding coercion, and distorted validity of the results through incentivisation. Participants' identities remained confidential throughout the study. Ethical clearance was granted by UCT.

Statistical analyses

Statistical analysis of cleaned survey data was conducted using the RStudio statistical software (version 1.2.5). For all statistical analyses, significance level was set at $p < .05$. Initial descriptive analysis of the data was performed to explore the distributions, mean values and standard deviations of the demographic data, affective measures, BADE task outcomes and prior precision estimates for each interpretation type. The distribution of negative affect scores was positively skewed and thus, a natural logarithmic transformation was applied to improve the normality of the distribution. Any data value >3 standard deviations (SD) above or below the mean for any of the variables was flagged as a potential outlier and excluded from statistical analyses involving the relevant variable. One participant was removed from the dataset during initial data cleaning on account of a complete absence of variability in response patterns on the PANAS, TAS-20, and BADE task, providing a maximum rating for every item in the survey.

Basic reliability analyses were conducted on the BADE task outcomes to explore separate inter-item response correlations for priors of each interpretation type, ICE and IDE across all 24 BADE task items. Inter-item correlations were significant for all of these BADE task variables ($\alpha > .9$) indicating that all items were consistent in relation to other items and could be reliably retained.

Hypothesis 1

The relationship between alexithymia and negative affect was investigated using log-transformed negative affect scores to ensure normal distribution of the data. Two participants were provisionally removed as their negative affect score was >3 SD above the mean. The correlation between these variables was calculated to assess direction and strength of the relationship. Subsequently, a simple linear model was constructed to predict alexithymia trait scores from negative affect scores.

Hypothesis 2

A 3x3 factorial ANOVA was conducted to investigate if there was an interaction between the type of interpretation (emotional lure, neutral lure and true) and level of ambiguity about the scenario (prior rating, rating 1 and rating 3) on the plausibility rating assigned to interpretations. Two participants were provisionally removed as their total plausibility rating scores for true 3 were >3 SD below the mean. In post-hoc tests, familywise error rate was accounted for using the Bonferroni correction (Abdi, 2010). Diagnostic tests highlighted one participant that significantly distorted the model and violated assumptions of linearity and homogeneity of residuals. Thus, this participant was removed and the ANOVA re-run.

Hypothesis 3

The strength and direction of the relationship between IDE and ICE was quantified through correlation between these variables. Following this, a simple linear model was constructed that predicted ICE from IDE. To control for variation in participants' baseline plausibility estimates, a separate linear model was constructed to predict ICE from true 1 ratings. IDE was then added to this model as a second predictor and the subsequent R2 change noted. Diagnostic tests of the model indicated two participants that had significant influence on the model and violated assumptions of linearity and normal distribution of residuals that were removed. The model exhibited no threat of multicollinearity of predictor variables. Similarly, a separate linear model was constructed that predicted IDE from neutral 1 ratings (emotional 1 ratings were removed due to multicollinearity with VIF > 10). ICE was then added to this model as a second predictor and the subsequent R2 change noted. Diagnostic tests of the model indicated four outliers that had significant influence on the model and violated assumptions of linearity and homoscedasticity to be removed.

Hypothesis 4

The strength and direction of the relationship between prior plausibility estimates of emotional lures and alexithymia trait scores was quantified through correlation coefficients between these variables. Thereafter, a simple linear model was constructed that predicted ICE from IDE.

Hypothesis 5

The strength and direction of the relationship between prior plausibility estimates of emotional lures and positive and negative affect was quantified through correlation coefficients between emotional prior plausibility estimates and each of these affective variables. A multiple

linear model was constructed that predicted prior plausibility estimates of emotional lures from both positive and negative affect scores. Two participants were provisionally removed as their negative affect score was >3 SD above the mean. Diagnostic tests of the model highlighted four further data points that were significantly distorting the model and violating assumptions of linearity and homoscedasticity to be removed.

Hypothesis 6

The strength and direction of the relationship between alexithymia trait scores and ICE and IDE was quantified through correlation coefficients between alexithymia trait scores and these variables. Two simple linear models were constructed that aimed to predict ICE and IDE from alexithymia trait scores. Diagnostic tests indicated three data points in each model that were significantly distorting the model to be removed.

Hypothesis 7

The strength and direction of the relationships between positive and negative affect and ICE and IDE were quantified through correlation coefficients between each of these affective variables and each of these evidence integration scores. Two multiple linear models were constructed that aimed to predict ICE and IDE from positive affect scores and log negative affect scores. Two participants were provisionally removed as their negative affect score was >3 SD above the mean. For the both models, three further data points emerged that were significantly distorting the model and were, therefore, removed. Because negative affect was insignificant as a predictor of IDE in the multiple linear model, a second simple linear model was constructed that predicted IDE from positive affect only using all data points. However, diagnostic tests indicated three data points that were distorting the model that were then removed.

Results

Descriptive statistics

Descriptive statistics, including means (*M*) and standard deviations (*SD*) for the primary affective variables and primary BADE task variables are presented in Table 1.

Table 1

Descriptive Statistics

Variables	*n*	*M*	*SD*
Positive affect score	86	30.79	7.5
Log negative affect score	84	1.22	.15
Alexithymia trait score	86	46.05	12.16
Prior plausibility ratings			
Emotional lure priors	86	147.87	53.78
Neutral lure priors	86	149.34	52.89
True priors	86	145.1	53.49
IDE	86	188.03	82.96
Emotional lure change	86	-89.80	42.04
Neutral lure change	86	-98.23	42.91
ICE	86	60.03	32.22

Hypothesis 1

A moderate positive correlation was observed between negative affect and alexithymia trait scores (r = .31). Alexithymia trait scores could be significantly predicted from negative affect scores ($F(1,82) = 11.21$, p = .00). However, negative affect could only account for 12% of the variability in alexithymia trait scores (R^2 = .12) suggesting it was a weak predictor of alexithymia level.

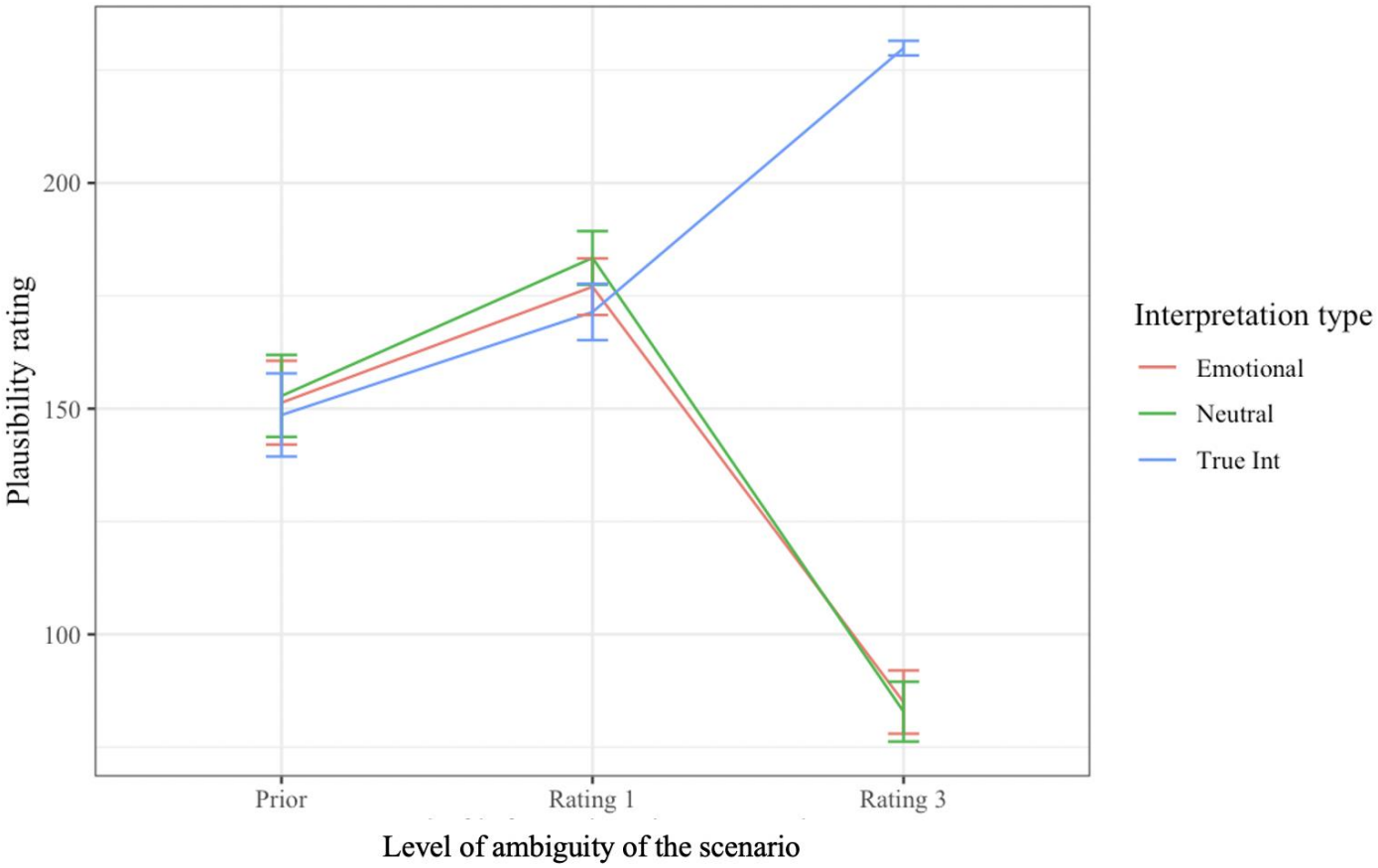
Hypothesis 2

A 3x3 factorial ANOVA found a significant interaction between the type of interpretation and ambiguity of the scenario on the plausibility rating assigned to interpretations in the BADE task

($F(4,738) = 135.73$, p <. 00). Notably, the cell means plot in Figure 1 indicates that this interaction is ordinal in nature.

Figure 1

Cell Means Plot of 3x3 Factorial ANOVA Output



Note. I bars are 95% confidence intervals.

Pairwise simple effects analysis indicates that plausibility ratings only significantly differ between interpretation types at rating 3 with plausibility ratings for true interpretations (M = 229.89, SD = 8.95) significantly higher than than plausibility ratings for emotional lure interpretations (M = 84.99, SD = 38.81), t(164) = 23.51, p < .00 and neutral lure interpretations (M = 82.87, SD = 36.68), t(164) = 23.86, p < .00.

Furthermore, simple effects analysis indicated that for each type of interpretation, plausibility ratings differed significantly as the scenario became less ambigious. For emotional lure interpretations, plausibility ratings significantly increased from prior rating (M = 151.33, SD = 51.38) to rating 1 (M = 177.02, SD = 34.71), t(164) = 4.17, p <.00 and then significantly

decreased between rating 1 and rating 3 (t(164) = -14.94, p <.00). Similarly, plausibility ratings amongst neutral lure interpretations significantly increased from prior rating (*M* = 152.83, *SD* = 50.32) to rating 1 (*M* = 183.4, *SD* = 32.88), t(164) = 4.96, p <.00 and then significantly decreased between rating 1 and rating 3 (t(164) = -16.31, p <.00). Finally, for true interpretations, plausibility ratings also significantly increased between prior ratings (*M* = 148.6, *SD* = 51.01) and rating 1 (*M* = 171.42, *SD* = 34.56), t(164) = 3.7, p <.00. However, plausibility ratings of true interpretations further significantly increased between rating 1 and rating 3 (t(164) = 9.49, p <.00).

Hypothesis 3

A moderate negative correlation was observed between IDE and ICE (*r* = -.42). Although ICE could be significantly predicted from IDE (*F*(1,82) = 17.17, p = .00), IDE could only account for 18% of the variation in ICE (R^2 = .16) indicating that IDE is a weak predictor of ICE.

In controlling for baseline plausibility estimates, ICE could be significantly predicted from True 1 ratings (*F*(1,82) = 769.46, R2 = .90, p = .00). Adding IDE as a second predictor only explained an additional 2% of the variation in ICE (ΔR^2 = .02, p = .00). Similarly, IDE could be significantly predicted from Neutral 1 ratings (*F*(1,82) = 58.13, , R^2 = .41, p = .00). ICE was added to this model but was found to be insignificant as a predictor of IDE when controlling for neutral 1 ratings (t(166) = 1.61, p = .11).

Hypothesis 4

A weak positive correlation was observed between alexithymia trait scores and emotional prior ratings (*r* = .17) although emotional prior ratings could not be significantly predicted from alexithymia trait scores (*F*(1,82) = 2.56, R^2 = .03, p = .11).

Hypothesis 5

A weak negative correlation was observed between positive affect scores and emotional prior ratings (*r* = -.17) while a notably weaker positive correlation was observed between negative affect and emotional prior ratings (*r* = .07). Prior plausibility estimates of emotional lures could not be significantly predicted from positive affect scores (t(160) = -1.53, p = .13) or negative affect scores (t(160) = .7, p = .48) (*F*(2,77) = 1.38, Adj.R^2 = .01, p = .26).

Hypothesis 6

A weak negative correlation was observed between alexithymia trait scores and both ICE (r = -.12) and IDE (r = -.13). However, alexithymia trait scores could not significantly predict ICE ($F(1,81) = 1.24$, $R^2 = .02$, p = .27) or IDE ($F(1,81) = 1.50$, $R^2 = .02$, p = .22).

Hypothesis 7

A weak negative correlation was observed between ICE and negative affect (r = -.21) while a notably weaker negative correlation was observed between ICE and positive affect (r = -.08). Neither negative affect scores (t(162) = 1.45, p = .15) nor positive affect scores (t(162) = .72, p = .47) could significantly predict ICE ($F(2,78) = 1.37$, $Adj.R^2 = .01$, p = .26).

A weak negative correlation was observed between IDE and negative affect (r = -.14) while a weak positive correlation was observed between IDE and positive affect (r = .26). IDE could be significantly predicted from positive affect scores alone ($F(1,81) = 5.78$, p = .02) although positive affect could only explain 7% of the variation in IDE suggesting it is a weak predictor of IDE ($R^2 = .07$). Negative affect was insignificant as a predictor of IDE (t(162) = -1.41, p = .16) when modelling IDE from both positive and negative affect scores ($F(2,78) = 3.97$, $Adj.R^2 = .09$, p = .02).

Discussion

In exploring how affect influences prior precision assigned to competing hypotheses under conditions of uncertainty, emotional prior plausibility ratings were predicted to be negatively correlated with alexithymia trait scores and negative affect, and positively correlated with positive affect. In exploring how affect influences the updating of beliefs in response to disambiguating information, the magnitude of change in precision estimates was expected to be positively correlated with negative affect and negatively correlated with alexithymia and positive affect.

In support of the first hypothesis, a significant but weak, positive relationship was found between negative affect and alexithymia. This corroborates previous studies that observed negative affect to be positively correlated with alexithymia using the TAS-20 (Khosravani et al., 2020; Parker & Taylor, 1997). However, alexithymia and negative affect are considered independent affective domains that can operate in a mutually reinforcing relationship (Lumley et al., 1994; Bilotta et al., 2016). The weak relationship may be partially attributed to the low and positively skewed distribution of negative affect scores which was unexpected considering the

impact of the COVID-19 pandemic on affective wellbeing. However, this may have been due to social desirability bias against disclosing the true extent of negative feelings.

Hypothesis 2 was partially supported as plausibility ratings were shown to differ significantly as the scenario became less ambiguous for all interpretation types. This indicates that for emotional, neutral and true interpretations, plausibility ratings and thus, beliefs were significantly updated in response to new information. Within a Bayesian framework, this suggests that incoming information was assigned significant precision so as to adjust the posterior probability assigned to competing hypotheses. Furthermore, because plausibility ratings of emotional and neutral lures significantly decreased between rating 1 and rating 3 while plausibility of true interpretations significantly increased between rating 1 and rating 3, this indicates that participants were able to successfully integrate both confirmatory and disconfirmatory evidence so that true interpretations successfully emerged as the dominant hypothesis when given access to all information.

Crucially, however, there was no significant difference in plausibility ratings between the types of interpretations at prior rating and rating 1 which fails to support hypothesis 2. This suggests that the lure interpretations were not initially more plausible than true interpretations which conflicts with previously observed response patterns on BADE task (Bronstein & Cannon, 2017) and suggests potentially impaired reliability of responses. This may also explain the unexpected negative relationship between ICE and IDE which conflicts hypothesis 3. More specifically, because rating 1 did not significantly differ between lure and true interpretations, the potential magnitude with which lure plausibility ratings could decrease from rating 1 was inversely correlated to the potential magnitude with which true plausibility ratings could increase. Thus, participants who initially assigned higher plausibility estimates to interpretations exhibited a greater decrease in lure ratings and smaller increase in true ratings in response to disambiguating evidence, while participants who initially assigned lower plausibility estimates to interpretations exhibited a smaller decrease in lure ratings and greater increase in true ratings in response to disambiguating evidence. This is corroborated by the observation that ICE and IDE were unable to explain significant variance in one another when controlling for baseline plausibility estimates. Because baseline plausibility estimates significantly explained the majority of the variance in evidence integration, variation in ICE and IDE was likely primarily due to variation in the liberalism of initial plausibility assignment.

Furthermore, no significant relationship was observed between prior ratings assigned to emotional lures and alexithymia, negative affect or positive affect, failing to support hypotheses 4 and 5. This finding suggests that affect and affective awareness do not significantly influence the assignment of precision to competing hypotheses under conditions of uncertainty, which conflicts with the conception of affect as an optimizer of precision with regards to free energy minimization proposed by Solms and Friston (2018). However, because initial plausibility ratings did not significantly differ between types of interpretations, emotional lures may not have been interpreted as more emotionally charged than neutral and true interpretations by participants. Consequently, interpretations may have presented no significant emotional valence to interact with participants' affective state and awareness in prior precision assignment. Furthermore, because prior emotional interpretations were presented to participants in the BADE task as written explanations of hypothetical ambiguous scenarios, it is possible that these interpretations were processed as sensory information rather than internally-generated prior hypotheses. Thus, assigning plausibility to these interpretations may represent a distinct process from Bayesian prior precision estimation under novel conditions, explaining the unexpected result.

Additionally, neither negative affect nor alexithymia were found to be significantly predictive of ICE or IDE, which contradicted with hypotheses 6 and 7. This may be partly due to demand characteristics bias, whereby participants' assigned plausibility to interpretations on account of how they believed they were expected to respond, masking the true impact of affective variables on belief updating. Alternatively, because high alexithymia levels or negative affect are not pathological in themselves, these variables may exhibit no significant distortion on the flexible and adaptive integration of incoming information into existing beliefs. Because the BADE task has primarily been employed to identify biases in belief updating in those with delusional cognitive distortions (Woodward et al., 2006), it may not be sensitive enough to quantify variation in adaptive belief updating due to these affective measures.

However, the observed insignificant relationship between affective measures and evidence integration must necessarily be interpreted in relation to the significant impact of initial plausibility estimates on ICE and IDE. Because a significant majority of variation in ICE and IDE was explained by initial true and lure plausibility estimates respectively, this result may simply indicate that negative affect and alexithymia trait scores were unrelated to initial

plausibility estimates which aligns with the insignificant findings of hypotheses 4 and 5. Therefore, the absence of observed relationship between ICE and IDE and negative affect and alexithymia cannot exclude the possibility that affect and alexithymia may influence the Bayesian updating of beliefs.

In contradiction with hypothesis 7, a weak, positive relationship was observed between positive affect and IDE. Notably, because positive affect was not significantly related to prior ratings of emotional lures, it is unlikely that this relationship was mediated by initial plausibility estimates. Thus, positive affect appears to significantly account for 7% of the unique variance in IDE. Because experienced positive affect is associated with sucessful optimization of precision assigned to prediction errors and the reduction of free energy within Bayesian inference (Pezzulo, et al., 2015; Solms & Friston, 2018), a more positive affective state may create greater flexibility and freedom in relation to changing precision assignment, allowing for more significant updating of hypotheses in response to disconfirmatory information. Alternatively, greater residual positive affect may mean that greater precision must be assigned to prediction errors and, thus, hypotheses revised with greater magnitude in order for the system to register additional positive affect that arises from effective precision adjustment.

Limitations and directions for future research

This study has been the first to explore the role of affect and alexithymia in the Bayesian updating of beliefs using the BADE task. However, the general absence of significant relationships between affective measures and BADE task outcomes and overriding influence of initial plausibility estimates in determining magnitude of evidence integration suggests that the BADE task may be an unreliable tool to investigate this study's objective. Because the BADE task has primarily been used to detect biased belief updating in those with psychiatric distortions, the task may not be sensitive enough to identify nuanced differences in evidence integration that arise due to non-pathological variation in positive and negative affect and alexithymia. As a result, future studies of a similar nature may benefit from including participants who have been diagnosed with clinical affective disorders, rather than limiting the sample to non-clinical participants. Similarly, although sample size was sufficient, a larger and more diverse sample with respect to age, gender, and cultural or socioeconomic variables may have produced more reliable results.

Further concern arises about the validity of the BADE task as a means to accurately quantify variation in Bayesian inference. While the BADE task was designed to detect biases in the assignment and updating of plausibility of competing hypotheses, the task may assess a form of meta-cognitive belief updating that is at least partially distinct from the automatic inference which Bayesian models seek to explain. Thus, BADE task outcomes may be confounded by attitudes towards specific items, the resemblance of items to salient personal experiences or linguistic misunderstanding of interpretations or disambiguating information. Further threats to validity arose from the self-reported nature of responses and thus, potential variation in what participants may regard as extremely plausible or implausible. Thus, future research on Bayesian belief updating using the BADE task should administer an additional verified measure of Bayesian inference, such as binocular rivalry (Hohwy et al., 2008) or the 'beads task' (Ross et al., 2015) to verify the validity of the BADE task as a measure of Bayesian inference. Additionally, the task may be improved by categorizing the valence of emotional lures to enable the influence of positive and negative affective states on the precision assigned to affectively congruent and incongruent hypotheses to be explored. Affect may also be periodically reassessed over the course of the task to explore how change in affect relates to changing precision of hypotheses for more valid mimicking of the Bayesian process.

Some participants expressed confusion about the instructions of the task, suggesting it may have been more reliably administered within a controlled laboratory setting with more explicit instructions and examples, so that misinterpretation of the task and ambiguity of self-report measures could be minimized. Because the lack of significantly higher initial plausibility ratings for lure interpretations than true interpretations may signify impaired reliability of BADE responses, repeating the task in a more controlled laboratory setting would be recommended. Additionally, an experimental setting allows for the manipulation of affective outcomes between groups of participants so that a control group may be established for more effective statistical comparison, and individual variation in affective measures to be overcome. Through these improvements, the BADE task may offer a valid and reliable means to explore the influence of affect on Bayesian inference.

Conclusion

Despite the scarcity of conclusive insights into the relationship between affect and Bayesian updating of beliefs, this study holds distinct value in guiding future research within this

field. This was the first time that the BADE task was used to remotely explore variability in Bayesian belief updating within a healthy population and thus, presented methodological challenges. While participants successfully integrated incoming evidence into existing hypotheses, the influence of affective variables on BADE task responses was largely inconclusive. As a result, the task may not be sensitive enough to account for the influence of affect on assignment and adjustment of precision to competing hypotheses under conditions of decreasing uncertainty. Furthermore, limitations in the validity and reliability of the task as a means to quantify Bayesian belief updating rendered the task ineffective and inconclusive to successfully execute the objectives of this study. Therefore, future research on the relationship between affect and belief updating should explore the validity of the BADE task in relation to other measures of Bayesian inference within a controlled laboratory setting. Additionally, future experimental research should focus on the influence of positive affect on integration of disconfirmatory evidence to verify the significance of this observed relationship.

References

Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design, 1*(8), 1-8. http://dx.doi.org/10.4135/9781412961288.n178

Arntz, A., Rauner, M., & Van den Hout, M. (1995). "If I feel anxious, there must be danger": Ex-consequentia reasoning in inferring danger in anxiety disorders. *Behaviour Research and Therapy, 33*(8), 917-925. https://doi.org/10.1016/0005-7967(95)00032-s

Badcock, P. B., Friston, K. J., Ramstead, M. J., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, & Behavioral Neuroscience, 19*(6), 1319-1351. https://doi.org/10.3758/s13415-019-00721-3

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38*(1), 23-32. https://doi.org/10.1016/0022-3999(94)90005-1

Bagby, R. M., Parker, J. D., & Taylor, G. J. (2020). Twenty-five years with the 20-item Toronto Alexithymia Scale. *Journal of Psychosomatic Research, 131*, 109940. https://doi.org/10.1016/j.jpsychores.2020.109940

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*(1), 1-23. https://doi.org/10.1093/scan/nsx060

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology, 23*(3), 361-372. https://doi.org/10.1016/j.conb.2012.12.012

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1708), 20160011. https://doi.org/10.1098/rstb.2016.0011

Bilotta, E., Giacomantonio, M., Leone, L., Mancini, F., & Coriale, G. (2016). Being alexithymic: Necessity or convenience. Negative emotionality × avoidant coping interactions and alexithymia. *Psychology and Psychotherapy: Theory, Research and Practice, 89*(3), 261-275. https://doi.org/10.1111/papt.12079

Brewer, R., Cook, R., & Bird, G. (2016). Alexithymia: A general deficit of interoception. *Royal Society Open Science, 3*(10), 150664. https://doi.org/10.1098/rsos.150664

Bronstein, M. V., & Cannon, T. D. (2017). Bias against disconfirmatory evidence in a large nonclinical sample: Associations with schizotypy and delusional beliefs. *Journal of Experimental Psychopathology*, *8*(3), 288-302. https://doi.org/10.5127/jep.057516

Brown, R. G., & Pluck, G. (2000). Negative symptoms: The 'pathology' of motivation and goal-directed behaviour. *Trends in Neurosciences*, *23*(9), 412-417. https://doi.org/10.1016/s0166-2236(00)01626-x

Buck, K. D., Warman, D. M., Huddy, V., & Lysaker, P. H. (2012). The relationship of metacognition with jumping to conclusions among persons with schizophrenia spectrum disorders. *Psychopathology*, *45*(5), 271-275. https://doi.org/10.1159/000330892

Chekroud, A. M. (2015). Unifying treatments for depression: An application of the Free Energy Principle. *Frontiers in Psychology*, *6*, 153. https://doi.org/10.3389/fpsyg.2015.00153

Clark, A. (2013). Whatever next? Neural prediction, situated agents, and the future of cognitive science. *Behavioural and Brain Sciences, 181-204*, *36*(3). https://doi.org/10.1017/S0140525X12000477

Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief. *Quarterly Journal of Experimental Psychology*, *60*(8), 1041-1062. https://doi.org/10.1080/17470210701338071

Damasio, A., (2010). *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon.

Díaz-García, A., González-Robles, A., Mor, S., Mira, A., Quero, S., García-Palacios, A., … Botella, C. (2020). Positive and negative affect schedule (PANAS): Psychometric properties of the online Spanish version in a clinical sample with emotional disorders. *BMC Psychiatry 20*(1), 56. https://doi.org/10.1186/s12888-020-2472-1

Edwards, M. J., Adams, R. A., Brown, H., Parees, I., & Friston, K. J. (2012). A Bayesian account of 'hysteria'. *Brain*, *135*(11), 3495-3512. https://doi.org/10.1093/brain/aws129

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. https://doi.org/10.3389/fnhum.2010.00215

Flykt, A. (2005). Visual search with biological threat stimuli: Accuracy, reaction times, and heart rate changes. *Emotion*, *5*(3), 349. https://doi.org/10.1037/1528-3542.5.3.349

Fotopoulou, A. (2013). Beyond the reward principle: Consciousness as precision seeking. *Neuropsychoanalysis*, *15*(1), 33-38. https://doi.org/10.1080/15294145.2013.10773715

Frischen, A., Eastwood, J. D., & Smilek, D. (2008). Visual search for faces with emotional expressions. *Psychological Bulletin*, *134*(5), 662-676. https://doi.org/10.1037/0033-2909.134.5.662

Friston, K. (2002). Functional integration and inference in the brain. *Progress in Neurobiology, 68,* 113–143. https://doi.org/10.1016/s0301-0082(02)00076-x

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11): e1000211. https://doi.org/10.1371/journal.pcbi.1000211

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11,* 127–138. https://doi.org/10.1038/nrn2787

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3,* 1-7. https://doi.org/10.3389/fpsyg.2012.00130

Gilbert, P. (1998). The evolved basis and adaptive functions of cognitive distortions. *British Journal of Medical Psychology, 71*(4), 447-463. https://doi.org/10.1111/j.2044-8341.1998.tb01002.x

Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive coding Or evidence accumulation? False inference and neuronal fluctuations. *PLoS One, 5*(3), e9926. https://doi.org/10.1371/journal.pone.0009926

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687-701. https://doi.org/10.1016/j.cognition.2008.05.010

Hulley, S. B. (2013). Appendix 6C. In S. B. Hulley, S. R. Cummings, W. S. Browner, D. Grady & T. B. Newman (Eds.), *Designing clinical research: An epidemiologic approach* (4th ed., pp. 79). Lippincott Williams & Wilkins.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, *9*(6), e1003094. https://doi.org/doi:10.1371/journal.pcbi.1003094

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference.

Annual Review of Psychology, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Khosravani, V., Ardestani, S., Alvani, A., & Amirinezhad, A. (2020). Alexithymia, empathy, negative affect and physical symptoms in patients with asthma. *Clinical Psychology & Psychotherapy, 27*(5), 736-748. https://doi.org/10.1002/cpp.2458

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*(5), 372-385. https://doi.org/10.1037/0003-066x.50.5.372

Lumley, M. A., Downey, K., Stettner, L., Wehmer, F., & Pomerleau, O. F. (1994). Alexithymia and Negative Affect: Relationship to Cigarette Smoking, Nicotine Dependence, and Smoking Cessation. *Psychotherapy and Psychomatics, 61,* 156-162. https://doi.org/10.1159/000288884

Merz, E. L., Malcarne, V. L., Roesch, S. C., Ko, C. M., Emerson, M., Roma, V. G., & Sadler, G. R. (2013). Psychometric properties of Positive and Negative Affect Schedule (PANAS) original and short forms in an African American community sample. *Journal of Affective Disorders, 151*(3), 942-949. https://doi.org/10.1016/j.jad.2013.08.011

Moritz, S., & Woodward, T. S. (2006). A generalized bias against disconfirmatory evidence in schizophrenia. *Psychiatry Research, 142*(2-3), 157-165. https://doi.org/10.1016/j.psychres.2005.08.016

Mulder, P. (2018, May 25). PANAS Scale [E-learning platform]. Toolshero. https://www.toolshero.com/psychology/panas-scale/

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241-251. https://doi.org/10.1007/bf00198477

Nemiah, J. C., Freyberger, H., & Sifneos, P. E. (1976). Alexithymia: A view of the psychosomatic process. In O. W. Hill (Ed.), *Modern trends in psychosomatic medicine* (Vol. 3; pp. 430-439). Butterworths.

Nicholson, T. M., Williams, D. M., Grainger, C., Christensen, J. F., Calvo-Merino, B., & Gaigg, S. B. (2018). Interoceptive impairments do not lie at the heart of autism or alexithymia. *Journal of Abnormal Psychology*, *127*(6), 612. https://doi.org/10.1037/abn0000370

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, *80*(3), 381-396. https://doi.org/10.1037/0022-3514.80.3.381

Pandey, R., Saxena, P., & Dubey, A. (2011). Emotion regulation difficulties in alexithymia and mental health. *Europe's Journal of Psychology*, *7*(4), 604-623. https://doi.org/10.5964/ejop.v7i4.155

Panksepp J. (1992). A critical role for "affective neuroscience" in resolving what is basic about basic emotions. *Psychological Review, 99*(3), 554–560. https://doi.org/10.1037/0033-295x.99.3.554

Panksepp, J. (2000). The neurodynamics of emotions, an evolutionary-neurodevelopmental view. In M. D. Lewis & I. Granic (Eds.), *Emotion, self-organization, and development* (pp. 236–264). Cambridge University Press.

Parker J. D. A., & Taylor G. J. (1997). Relations between alexithymia, personality, and affects in G. J. Taylor, R. M. Bagby & J. D. A. Parker (Eds.), *Disorders of affect regulation: Alexithymia in medical and psychiatric illness* (pp. 67-92). Cambridge University Press. https://doi.org/10.1017/CBO9780511526831.007

Peters, A., McEwen, B. S., & Friston, K. (2017). Uncertainty and stress: Why it causes diseases and how it is mastered by the brain. *Progress in Neurobiology*, *156*, 164-188. https://doi.org/10.1016/j.pneurobio.2017.05.004

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17-35. https://doi.org/10.1016/j.pneurobio.2015.09.001

Pfuhl, G. (2017). A Bayesian perspective on delusions: Suggestions for modifying two reasoning tasks. *Journal of Behavior Therapy and Experimental Psychiatry*, *56*, 4-11. https://doi.org/10.1016/j.jbtep.2016.08.006

Ross, R. M., McKay, R., Coltheart, M., & Langdon, R. (2015). Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. *Schizophrenia Bulletin*, *41*(5), 1183-1191. https://doi.org/10.1093/schbul/sbu187

Sanford, N., Veckenstedt, R., Moritz, S., Balzan, R. P., & Woodward, T. S. (2014). Impaired integration of disambiguating evidence in delusional schizophrenia patients. *Psychological Medicine*, *44*(13), 2729. https://doi.org/10.1017/s0033291714000397

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160007. https://doi.org/10.1098/rstb.2016.0007

Smith, R., Kuplicki, R., Feinstein, J., Forthman, K. L., Stewart, J. L., Paulus, M. P., Tulsa 1000 investigators, & Khalsa, S. S. (2020). A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS Computational Biology*, *16*(12), e1008484. https://doi.org/10.1371/journal.pcbi.1008484

Solms, M., & Friston, K. (2018). How and why consciousness arises: Some considerations from physics and physiology. *Journal of Consciousness Studies*, *25*(5-6), 202-238. https://doi.org/10.3389/fpsyg.2018.02714

Solms, M., & Turnbull, O. (2018). *The brain and the inner world: An introduction to the neuroscience of subjective experience.* Routledge. https://doi.org/10.4324/9780429481239

Taylor, G. J. (2000). Recent developments in alexithymia theory and research. *The Canadian Journal of Psychiatry*, *45*(2), 134-142. https://doi.org/10.1177/070674370004500203

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the positive and negative affect schedule-expanded form.* Unpublished manuscript, University of Iowa. https://doi:10.17077/48vt-m4t2

Woodward, T. S., Moritz, S., Cuttler, C., & Whitman, J. C. (2006). The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, *28*(4), 605-617. https://doi.org/10.1080/13803390590949511

Appendix A

Recruitment Advertisement

Dear student/potential participant,

You are invited to participate in a research project being conducted within the Psychology Department at the University of Cape Town. This research aims to explore the relationship between emotion and the updating of beliefs using an evidence-integration task . The study has been granted ethical clearance by the University of Cape Town's Research Ethics committee. The information generated through this study will be used for the purpose of compiling an academic thesis. Primary researchers are Psychology honours students, Ethan Gray and Gemma Strohbach, under the supervision of Dr Donné van de Westhuizen and Professor Mark Solms.

Participation is completely voluntary. If you choose to participate in this study, you will be required to complete an online survey that will ask you to disclose basic demographic information, subjectively assess your mood and complete a task exploring how you utilize new incoming information to update your beliefs . This survey should not take you more than 45 minutes to complete and poses no risks of harm to you. After beginning the survey, you are free to withdraw from participation at any point .

All personal information details in addition to your results on the task will remain completely confidential . Your data will be assigned a random number which will be used as a reference as opposed to your name. Any data collected will only be viewed by the research team, and will not be distributed to the public or included within the academic paper. In order to participate, you need to:

- Be at least 18 years of age
- Be fluent in English
- Currently reside in South Africa
- Not have been diagnosed with any psychiatric or neurological disorder including, but not limited to, depressive disorders, anxiety disorders or concussion.

If you are interested in participating and you meet all of the above criteria, please click the following link. This link will direct you to a consent form which must be completed prior to beginning the survey.

https://bit.ly/AffectUpdateBeliefs

Thank you for your participation and assistance.

Kind regards,

Gemma Strohbach (STRGEM001@myuct.ac.za) and Ethan Gray (GRYETH001@myuct.ac.za)

Appendix B

Informed Consent Form

UNIVERSITY OF CAPE TOWN

DEPARTMENT OF PSYCHOLOGY

The relationship between emotion and updating of beliefs

1. Invitation and Purpose

Thank you for your interest in participating in this study about emotion and beliefs. We are Honours students from the Psychology department at the University of Cape Town. The information gained in the study will be used for educational purposes.

2. Exclusion criteria

You may only take part in this study if you are 18 years or older, are currently residing in South Africa, are fluent in English and do not have a history of any psychiatric or neurological disorder that may influence your mood or cognitive reasoning.

3. Procedures

The session should take about 45 minutes, in which you will be asked to complete the Positive and Negative Affect Schedule, The 20-item Toronto Alexithymia Scale and the 'Bias-Against Disconfirmatory Evidence' (BADE) task .

4. Voluntary Participation and Confidentiality

Participating in this study is voluntary. You are free to stop completing the questionnaire without any consequences. Any information you share will be kept strictly confidential. Your identity will remain completely confidential throughout the research process. You have the right to request that any information you have shared to be removed from the study.

5. Benefits

The benefits of this study include assisting in developing greater understanding of how affect functions and interacts with cognitive beliefs mechanisms[CW14] .

6. Risks, Discomforts & Inconveniences

This study poses a low risk of harm to you. You might be inconvenienced by having to [CW15] give up 45 minutes of your time.

7. Questions

You are encouraged to direct any questions that you may have about the study towards the following researchers:

Primary researchers:

Ethan Gray
Contact no.: 083 286 2372
Email: GRYETH001@myuct.ac.za

Gemma Strohbach
Email: STRGEM001@myuct.ac.za
Contact no.: 074 793 9075

Research supervisors:

Dr Donné van der Westhuizen
Email: donvanwest@gmail.com

Prof. Mark Solms
Email: mark.solms@uct.ac.za

Questions about your rights as a study participant, comments or complaints about the study also may be presented to the Research Ethics Committee, Department of Psychology, UCT[CW16] .

☐ **I have been informed about this research study and understand its purpose. I agree to take part in this research as a participant and consent to my scores being used in this study. I know that I am free to withdraw from this study at any time, and that doing so will not disadvantage me in any way.**

Agree and continue to the survey

Appendix C

Online Survey

Survey link: https://bit.ly/AffectUpdateBeliefs

Digital informed consent form

The Relationship Between Affect and Updating Beliefs

Dear potential participant,

Thank you for expressing interest in participating in this study. Please read through the following information before providing consent to participate.

1. Invitation and Purpose
You have shown interest in participating in this study that explores the relationship between affect and updating of beliefs within the BADE task. We are Honours students from the Psychology department at the University of Cape Town. The information gained in the study will be used for educational purposes.

2. Exclusion criteria
You may only take part in this study if you are 18 years old or older, are currently residing in South Africa, are fluent in English and do not have a history of any psychiatric or neurological disorder including, but not limited to, depressive disorders, anxiety disorders, attention disorders or concussion.

3. Procedures
The session should take about 45 minutes, in which you will be asked to complete two questionnaires pertaining to your current emotions, and one task which will require you to rate the plausibility of four statements in relation to a hypothetical scenario.

4. Voluntary Participation and Confidentiality
Participating in this study is voluntary. You are free to leave the session without any consequences. To ensure that any information you share is strictly confidential, a number will be assigned to your responses as opposed to your name. You have the right to request that any information you have shared to be removed from the study. If you are a psychology student at the University of Cape Town in need of SRPP points, your student number will be required.

5. Benefits
The benefits of this study include assisting in developing greater understanding of how emotions function and interact with cognitive beliefs mechanisms. If you are an undergraduate psychology student at UCT, you may receive 2 SRPP points to go towards the 2021 academic year.

Ethan Gray
Contact number: 083 286 2372
Email: GRYETH001@myuct.ac.za

Gemma Strohbach
Email: STRGEM001@myuct.ac.za
Contact number: 074 793 9075

Research supervisor:
Prof. Mark Solms
Email: mark.solms@uct.ac.za

Questions about your rights as a study participant, comments or complaints about the study also may be presented to the Research Ethics Committee, Department of Psychology, UCT.

* Required

Please confirm that you satisfy all the inclusion criteria by selecting the checkboxes that apply to you *

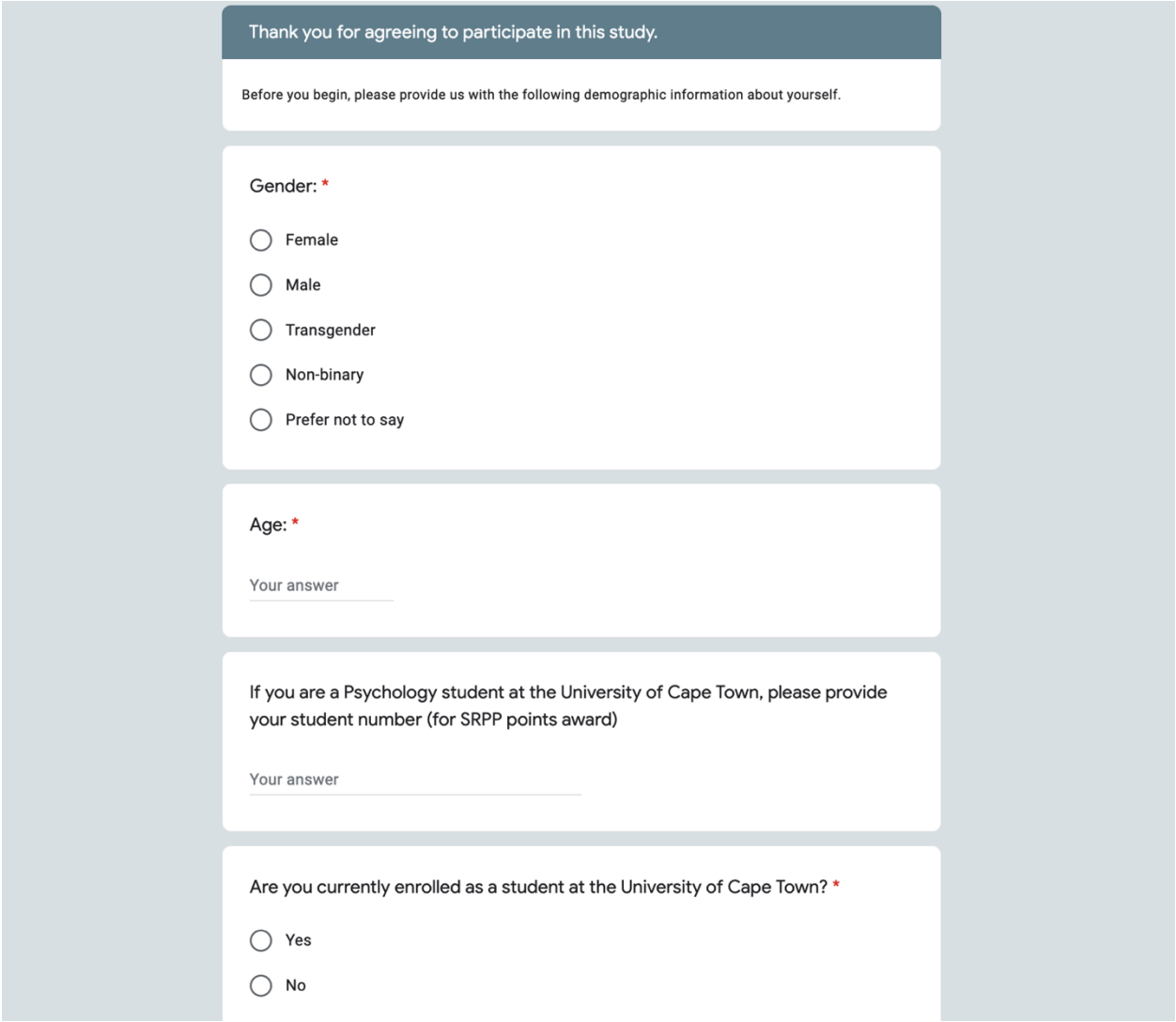
- [ ] I am 18 or older
- [ ] I am currently residing in South Africa
- [ ] I am fluent in English
- [ ] I have not been diagnosed with any form of depressive disorder, anxiety disorder, attention disorder or concussion

I have been informed about this research study and understand its purpose. I agree to take part in this research as a participant and consent to my scores being used in this study. I know that I am free to withdraw from this study at any time without it disadvantaging me in any way. *

- ◯ I agree

Next

Section 1: Demographic Information



Thank you for agreeing to participate in this study.

Before you begin, please provide us with the following demographic information about yourself.

Gender: *

- ◯ Female
- ◯ Male
- ◯ Transgender
- ◯ Non-binary
- ◯ Prefer not to say

Age: *

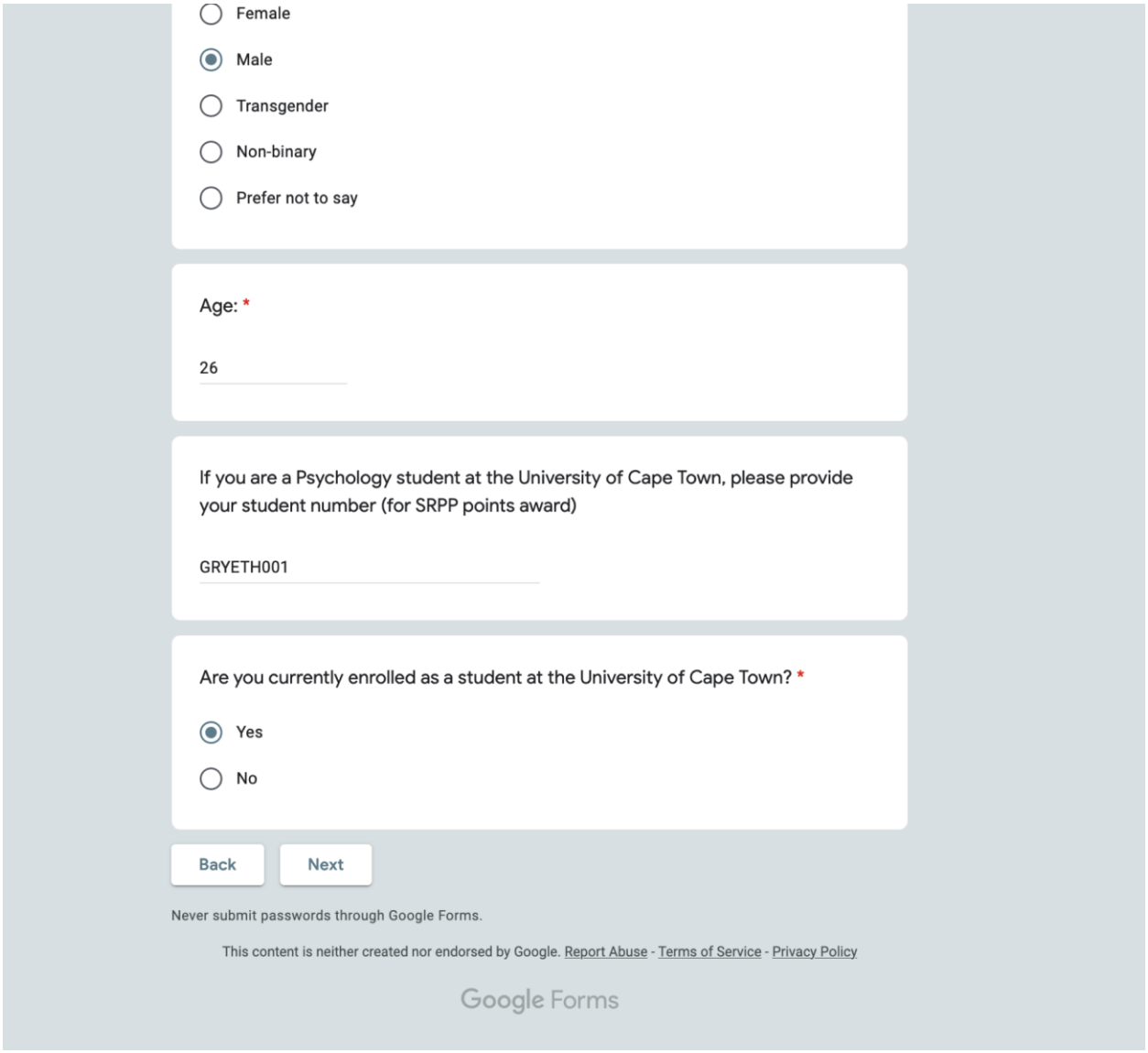
Your answer

If you are a Psychology student at the University of Cape Town, please provide your student number (for SRPP points award)

Your answer

Are you currently enrolled as a student at the University of Cape Town? *

- ◯ Yes
- ◯ No



Female

Male

Transgender

Non-binary

Prefer not to say

Age: *

26

If you are a Psychology student at the University of Cape Town, please provide your student number (for SRPP points award)

GRYETH001

Are you currently enrolled as a student at the University of Cape Town? *

Yes

No

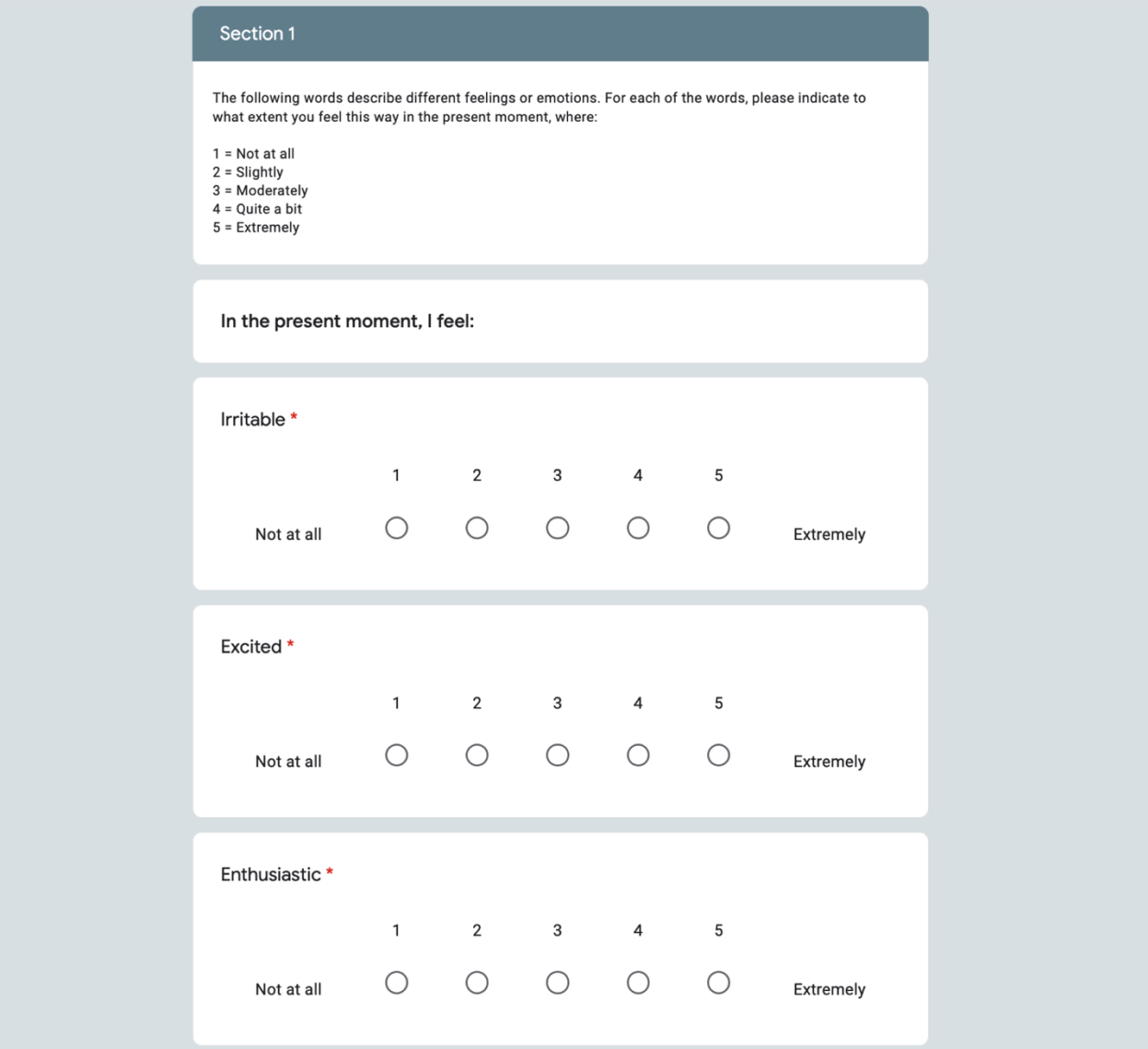
Back Next

Never submit passwords through Google Forms.

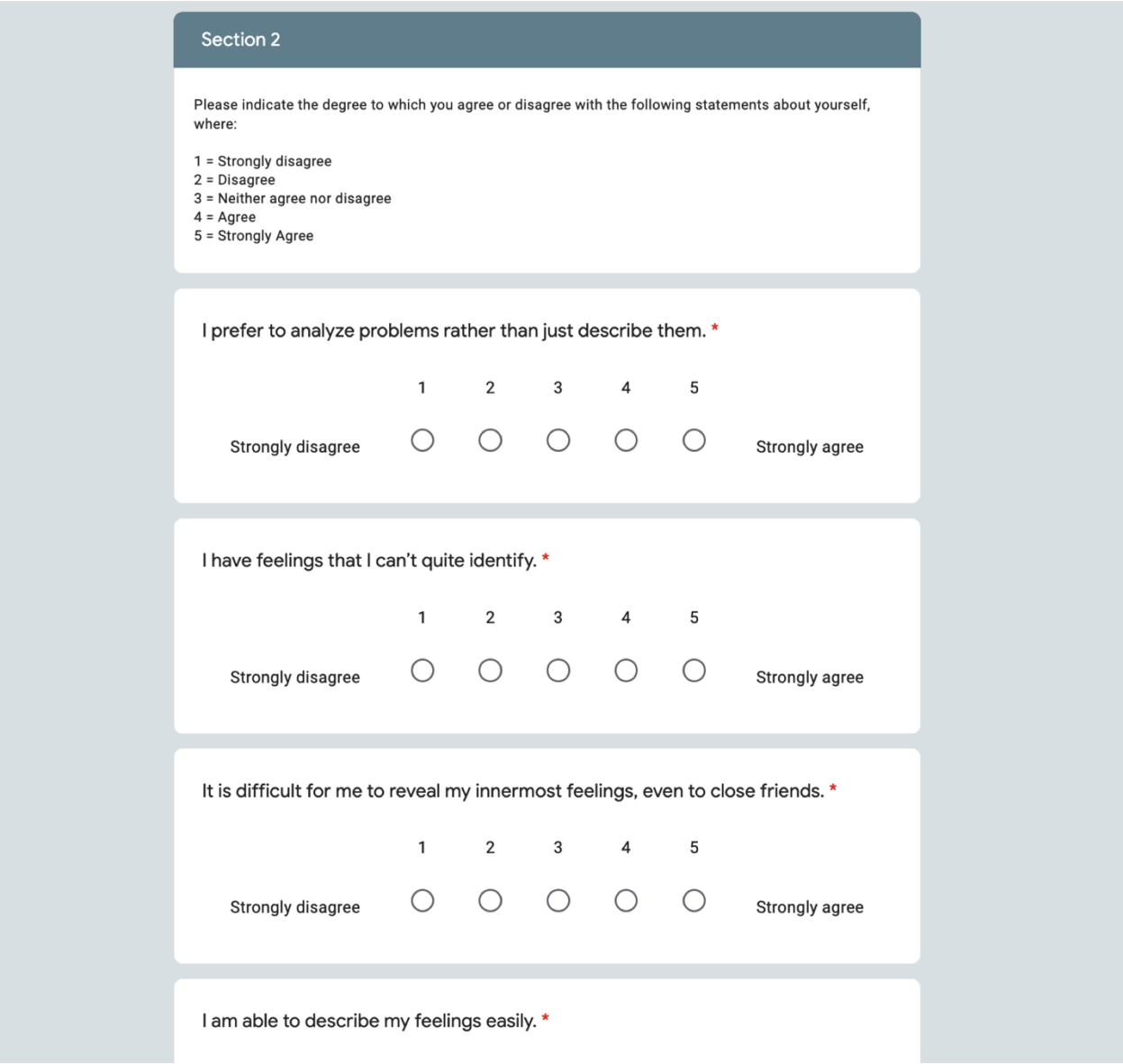
This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms

Section 2: PANAS



Section 3: TAS-20



Section 2

Please indicate the degree to which you agree or disagree with the following statements about yourself, where:

1 = Strongly disagree
2 = Disagree
3 = Neither agree nor disagree
4 = Agree
5 = Strongly Agree

I prefer to analyze problems rather than just describe them. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I have feelings that I can't quite identify. *

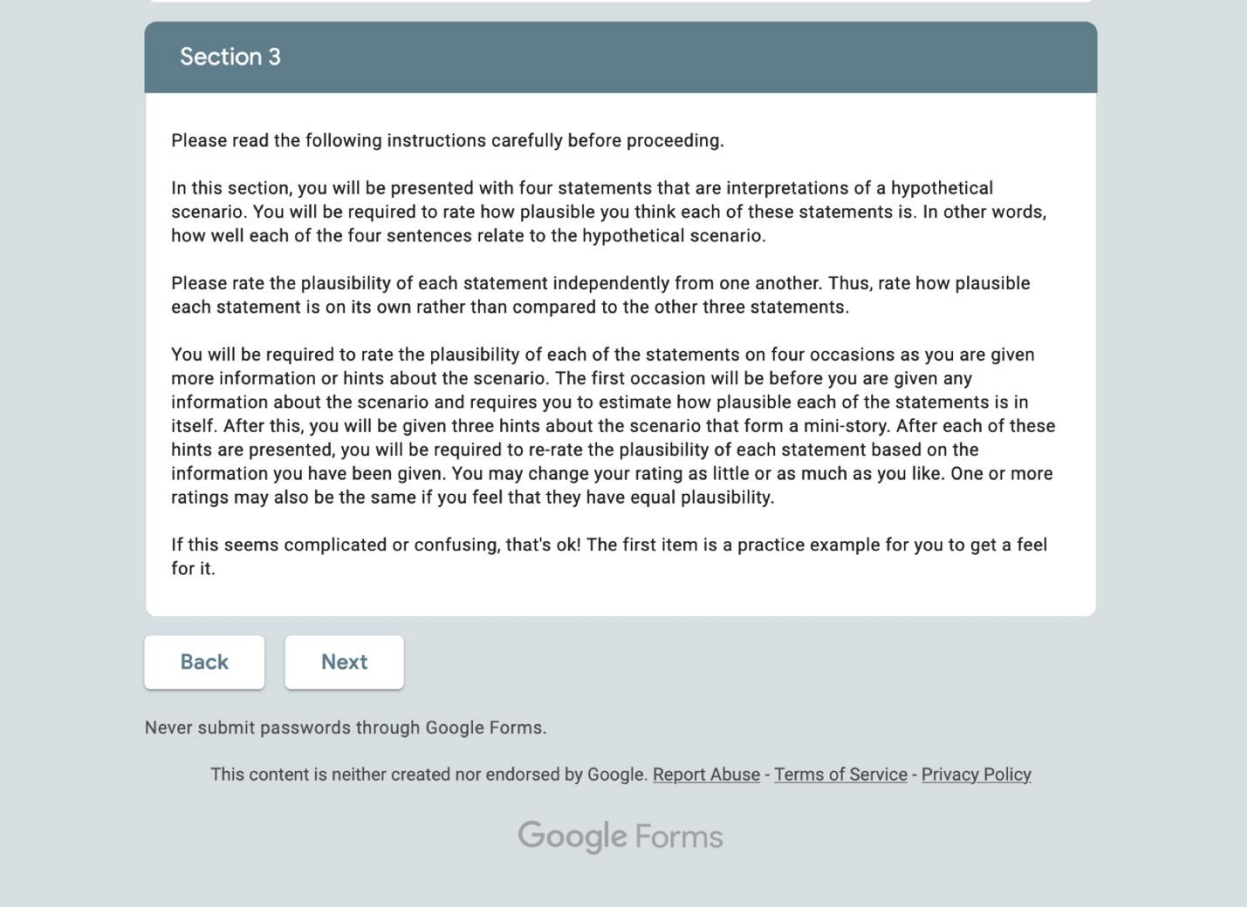
	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

It is difficult for me to reveal my innermost feelings, even to close friends. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I am able to describe my feelings easily. *

Section 4: BADE Task



Section 3: Practice items

Without any information about the scenario, please rate the plausibility of the following statements:

Note: A rating of 1 indicates that the statement is implausible while a rating of 10 indicates that the statement is extremely plausible.

Trevor is about to have a heart attack. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor is suffering from drug withdrawal. *

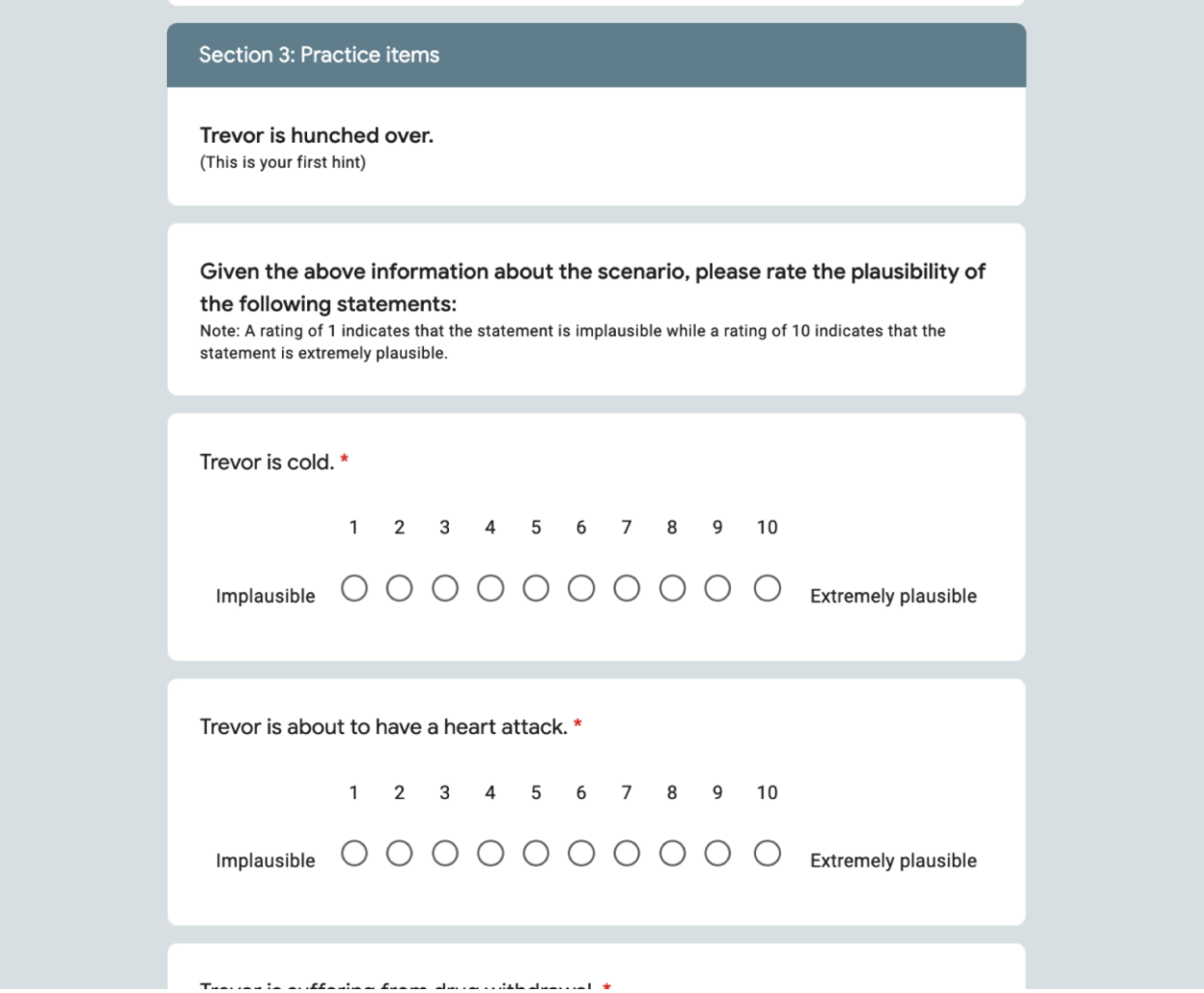
	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor is cold. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor in desperate need of some cash. *

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○



Section 3: Practice items

Trevor is hunched over.
(This is your first hint)

Given the above information about the scenario, please rate the plausibility of the following statements:
Note: A rating of 1 indicates that the statement is implausible while a rating of 10 indicates that the statement is extremely plausible.

Trevor is cold. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor is about to have a heart attack. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor is suffering from drug withdrawal. *

Section 3: Practice items

Trevor is hunched over.
(This is the first hint)

Trevor is acting strangely.
(This is the second hint)

Given the above information about the scenario, please rate the plausibility of the following statements:
Note: A rating of 1 indicates that the statement is implausible while a rating of 10 indicates that the statement is extremely plausible.

Trevor is cold. *

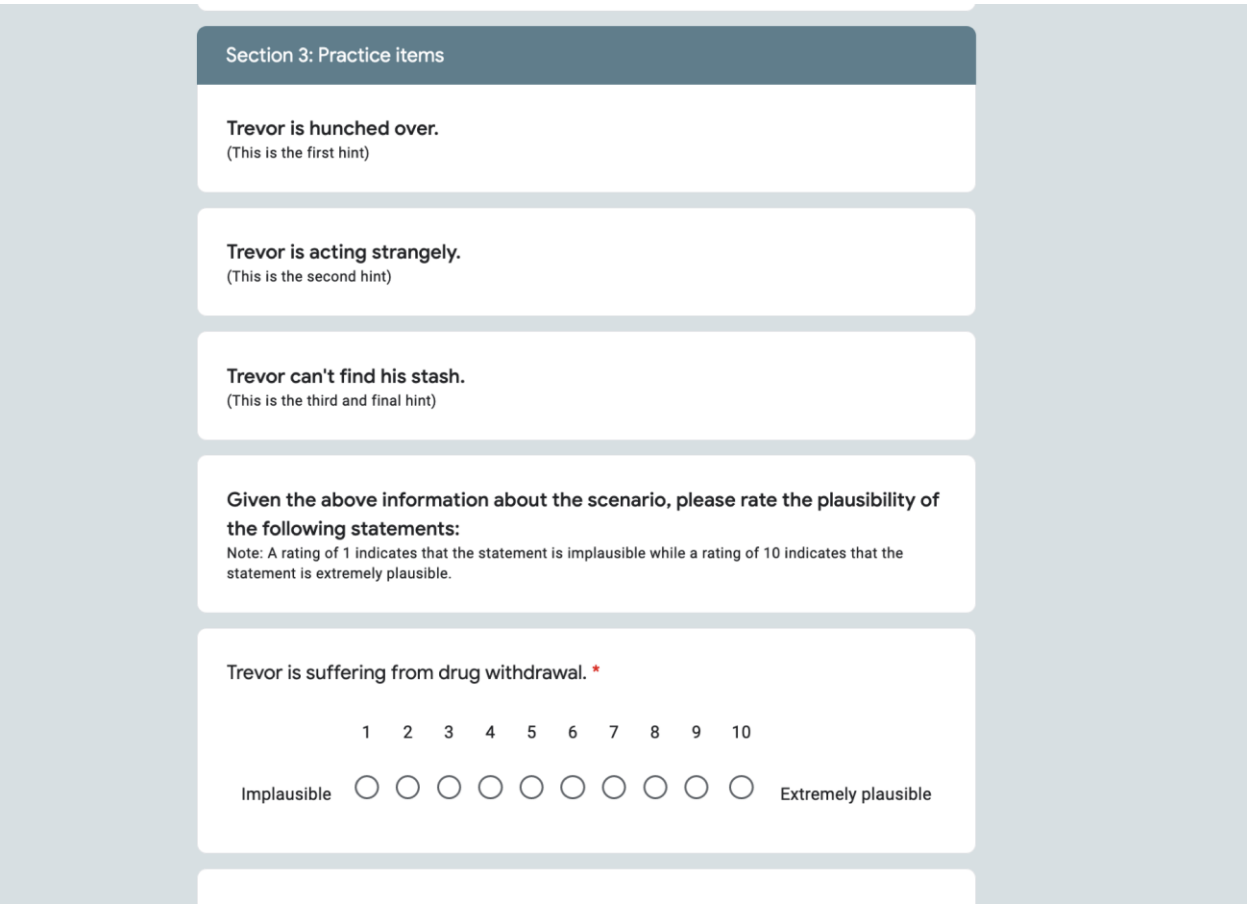
	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor is suffering from drug withdrawal. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Trevor is about to have a heart attack. *

1	2	3	4	5	6	7	8	9	10



Section 3: Practice items

Trevor is hunched over.
(This is the first hint)

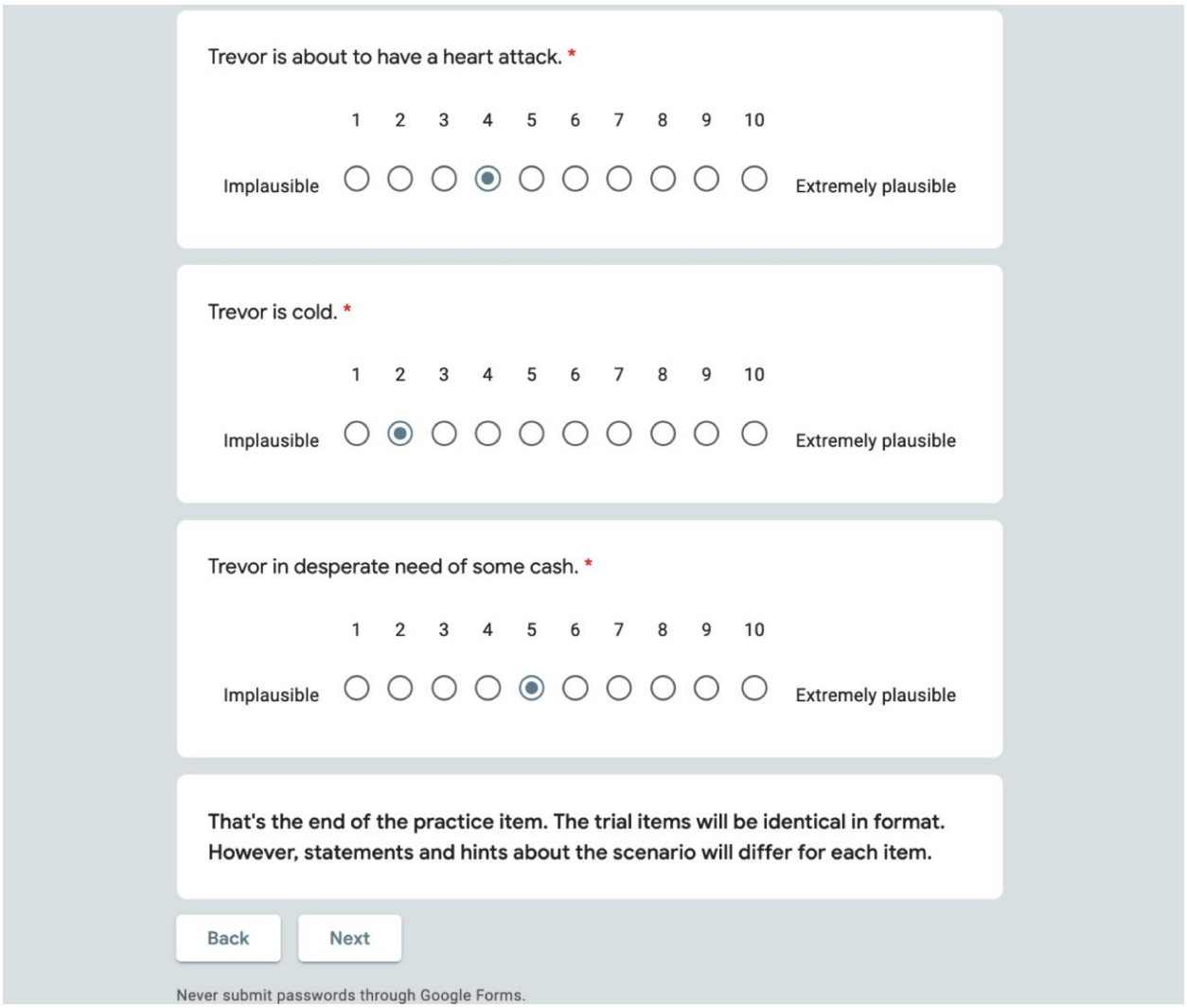
Trevor is acting strangely.
(This is the second hint)

Trevor can't find his stash.
(This is the third and final hint)

Given the above information about the scenario, please rate the plausibility of the following statements:
Note: A rating of 1 indicates that the statement is implausible while a rating of 10 indicates that the statement is extremely plausible.

Trevor is suffering from drug withdrawal. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible


Trevor is about to have a heart attack. *
1 2 3 4 5 6 7 8 9 10
Implausible
Extremely plausible
Trevor is cold. *
1 2 3 4 5 6 7 8 9 10
Implausible
Extremely plausible
Trevor in desperate need of some cash. *
1 2 3 4 5 6 7 8 9 10
Implausible
Extremely plausible
That's the end of the practice item. The trial items will be identical in format. However, statements and hints about the scenario will differ for each item.
Back
Next
Never submit passwords through Google Forms.

Without any information about the scenario, please rate the plausibility of the following statements:

Note: A rating of 1 indicates that the statement is implausible while a rating of 10 indicates that the statement is extremely plausible.

Amanda is homeless. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Amanda has an eating disorder. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Amanda is a runway model. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Amanda has lost her fake teeth. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

1. Amanda is very thin.

Given the above information about the scenario, please rate the plausibility of the following statements:

Note: A plausibility rating of 1 indicates that the statement is implausible while a plausibility rating of 10 indicates that the statement is extremely plausible.

Amanda has an eating disorder. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Amanda is a runway model. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Amanda is homeless. *

	1	2	3	4	5	6	7	8	9	10	
Implausible	○	○	○	○	○	○	○	○	○	○	Extremely plausible

Amanda has lost her fake teeth. *