Attempting to Analyze Perspective-Taking with a False Belief Vignette Using the Implicit Relational Assessment Procedure

Attempting to Analyze Perspective-Taking with a False Belief Vignette Using the Implicit Relational Assessment Procedure

Abstract

Cognitive perspective-taking research has primarily been conducted under the rubric of theory of mind (ToM), with the core skill believed to involve the correct attribution of mental states to oneself and others as a means of explaining and predicting behavior. Relational frame theory (RFT) has provided a behavioral account of performances on true and false belief protocols by appealing to the three perspective-taking (deictic) relations. The current research sought to investigate the relative strength of cognitive perspective-taking abilities within the context of a false belief vignette and related IRAP. Experiment 1 investigated the impact of block order presentation and vignette stimuli order on IRAP performances. That is, across four conditions, rule order presentations (i.e. Vignette Consistent versus Vignette Inconsistent) and vignette stimuli presentation were manipulated. Results indicated that vignette consistent responding was observed to varying degrees across conditions. To decrease this variability across conditions, Experiment 2 presented a vignette before each block of trials but again the IRAP showed only limited sensitivity to the vignette. The current findings and considerations for future research are discussed in terms of a recently published conceptual analysis of false belief by Kavanagh, et al. (2020).

Key words: Relational frame theory, perspective-taking, false belief, behavioral processes

Perspective-taking has long been considered pivotal for human socialization (Mead, 1934; Piaget, 1948) in terms of enabling an individual to overcome early egocentrism and to adjust their behavior according to the expectations of others. The ability to take another's perspective is crucial in competitive settings (Galinsky et al., 2008); the establishment and maintenance of healthy interpersonal relations (Hughes, & Leekam, 2004); and strengthening social bonds (Galinsky & Ku, 2004; Vescio, et al., 2003). As well as the mainstream approach to perspective-taking, largely represented by Theory of Mind (ToM: Sodian & Kristen-Antonow, 2015), behavioral researchers working under the rubric of Relational Frame Theory (RFT) have approached perspective-taking as involving responding in accordance with three deictic relations: the interpersonal (I/you); the spatial (here/there); and the temporal (now/then; see Hayes, et al., 2001). Most RFT studies on deictic relations, or perspective-taking that appears to involve these relations, have employed the Barnes-Holmes (2001) protocol (or some variant), originally developed to assess and establish these relations in young children (for a review see Kavanagh, et al., 2020; Montoya-Rodríguez, et al., 2017). More recently, researchers have begun to explore other methodologies, such as the implicit relational assessment procedure (IRAP), to study deictic relational responding (see Golijani-Moghaddam, et al., 2013; Vahey, et al., 2015, for recent reviews of the reliability and validity of the IRAP).

The IRAP allows researchers to juxtapose alternative relational responses and thus obtain a measure of the relative strength or probability of specific relational responding. The IRAP typically presents combinations of positive and negative label and target stimuli (e.g., the word "pleasant" with a picture of a flower) and requires participants to confirm or disconfirm the relational coherence between them (i.e., "true" on coherent trials and "false" on incoherent trials). Thus, IRAPs comprise four trial types (e.g., *Flower-Pleasant*, *Flower-Unpleasant*, *Insect-Pleasant*, and *Insect-Unpleasant*) that are typically analyzed independently in terms of the difference in response latencies between responding that is deemed consistent (coherent) versus inconsistent (incoherent) with a participant's verbal history. In general, response latencies are expected to be shorter during blocks of trials that require history-consistent versus history-inconsistent responding (e.g., all things being equal one might predict responding "true" more quickly than "false" on the *Flower-Pleasant* trial type).

Three published studies have thus far used the IRAP to investigate deictic relations, particularly in terms of comparing responding to self versus responding to others. Barbero-Rubio, et al. (2016) presented participants with what they referred to as a *perspective-taking IRAP* that contained each participant's own name (self) versus the name of the researcher (other) as label stimuli, and statements describing specific current

characteristics of the self (e.g., “is in front of the laptop”) versus other (e.g., “is standing up”) as targets, along with “Yes” and “No” as response options. In order to manipulate perspective-taking, explicit rules were provided prior to each block of trials. Specifically, in order to encourage responding from one’s own perspective, participants in some blocks were instructed: “For the next block of trials, you have to respond as if you were you and Adrian [researcher] were Adrian.” In contrast, in order to encourage responding from the perspective of another, participants in other blocks were instructed: “For the next block of trials, you have to respond as if you were Adrian and Adrian were you.” The four trial types were referred to as: *I-I* (participant name-participant characteristics); *Other-Other* (researcher name-researcher characteristics); *I-Other* (participant name-researcher characteristics); and *Other-I* (researcher name-participant characteristics). The differences between self- and other-perspective blocks for each trial type were in the predicted direction (i.e., shorter latencies during self-perspective blocks), and these differences were significant in terms of the normalized D_{IRAP}-scores. Overall, the researchers concluded that these significant D_{IRAP}-scores indicated that the participants had little flexibility in changing from their own perspective to another perspective.

In a systematic replication of the Barbero-Rubio et al. (2016) study, Kavanagh, et al. (2018) used a similar IRAP, but the study also included a control IRAP that did *not* require responding to self versus other. That is, instead of comparing self with other, the control IRAP compared responding between two separate others (i.e., the researcher and a picture of another unknown participant). In Experiment 1, the data from the IRAP showed significantly larger *DIRAP*-scores on the *I-I* trial type versus *Other-Other*, but there was no difference in the control IRAP between *Researcher-Researcher* and *Other-Other*. Whilst a range of methodological differences between the two studies preclude systematic comparisons, both studies did show evidence of differences in responding to self versus other, but no difference in responding to two others in the context of the control IRAP.

One possible concern that could be raised regarding the two studies involving the self- versus other-IRAPs described above is that differences that emerged between responding to self and other within the IRAP could be attributed to factors other than perspective-taking per se. For example, in the study by Kavanagh et al. (2018) a pattern known as the single trial type dominance effect (STTDE) emerged in Experiment 1. That is, the size of the D_{IRAP}-score for the *I-I* trial type was significantly larger than for the *Other-Other* trial type. Although this dominance effect could indicate a history of responding from one’s own perspective more frequently than from another perspective, it does not necessarily indicate differences in the relative ability to take the

perspective of self versus another (see Kavanagh et al. for a detailed discussion). Research by Finn, et al. (2018) has also reported a STTDE when shapes and colors were presented as categories within the IRAP. This pattern consisted of significant differences in magnitude between trial types that shared the response option "true" during history-consistent blocks of trials. Specifically, participants who completed a 'shapes and colors IRAP' consistently produced larger D_{IRAP}-scores on *Color-Color* than on *Shape-Shape* trial types. The authors suggested that the difference in the size of the effects for the *Color-Color* and *Shape-Shape* trial types (the STTDE) may be explained by the fact that in natural language color words occur with far greater frequency than do shape words. Therefore, it is assumed that these differences in frequency in natural language are likely to have produced differences in the functional properties of color words and shape words. This same logic could be applied to a single IRAP that requires responding to self versus other (i.e., the effect could be the result of responding to self more frequently than to other in natural language, rather than an ability to perspective-take).

A second potential concern that could be raised regarding both experiments pertains to the simple target phrases that specified characteristics of self and other (e.g. "is sitting down", "is in front of the computer"). As such, it could be argued that responding on the IRAP simply required deictic relational responding, but not perspective-taking. Indeed, perspective-taking would appear to require more complex target statements or relational networks that involve taking the perspective of self versus other. For example, such statements could take the form of "When event X happens, self or other thinks or feels Y." In principle, this sort of complex relational network requires that the participant responds to statements that coordinate with how the self responds to particular events, versus how they perceive others will respond to the same events (*basic* deictic relational responding does not necessarily involve "working out" how someone else will respond).

Based on this reasoning, research by Kavanagh et al. (2019) employed a novel version of the IRAP, known as the Natural Language-IRAP (NL-IRAP), which presented complex statements in a natural-language format. Across a sequence of six experiments, a 'self-focused IRAP' required participants to respond to both positive (e.g., "I'm proud when I succeed in my exams") and negative (e.g., "Getting a fine make me angry") statements about themselves, whilst an 'other-focused IRAP' required participants to respond to similar statements about others. Experiment 1 and 2 investigated perspective-taking with regard to an *unspecified* other. That is, IRAP statements referred to others in general (e.g., "It makes other people happy if they win the lottery"). Experiments 3-6 investigated perspective-taking with regard to a *specified* other. A specified other referred to a significant other that each participant identified before completing the IRAP. The name of this

significant other was then inserted into the IRAP statements (e.g., “It makes David happy if he wins the lottery”. Across experiments the specific relationship between the significant other and the participant was manipulated. The results from the first two experiments indicated that there were significant differences between the self- versus other-focused IRAPs, when the other remained unspecified. The remaining four experiments, however, indicated that when the other was specified there was limited evidence that performances on the two IRAPs differed significantly. Overall, the IRAP effects for the most part, were in the predicted direction. However, on balance, the results could be seen as somewhat disappointing because there was little evidence of perspective-taking when other was specified, at least in terms of different performances across the two IRAPs, or in correlations among the IRAPs and the self-report measures.

In reflecting upon the results obtained in Kavanagh et al. (2019), two key issues emerged that seemed important to address. The first relates to the stimuli used to specify the self and other within the IRAPs. Specifically, self-related terms involved using the participant’s name or words such as “I”, “my” or “me”, whilst other-related terms involved using another’s name or words such as “they” or “others”. The use of such stimuli might allow for some ambiguity in how these stimuli were interpreted by participants. For example, when the on-screen stimulus was “I” the assumption was that the participant would interpret this as referring to self, rather than to the computer or another person. In general, it appeared that this assumption was upheld, but of course room for ambiguity remained.

A second concern was that the complexity of the statements might have encouraged participants to find ways of simplifying the task, and thus undermined the complex relational responding that was aiming to be captured. For example, when the on-screen statement was “It makes other people happy if they win the lottery” the assumption was that the participant would read the complete statement. It may have been the case that rather than responding to the complete statement, participants were solely responding to the words “other” and “happy”. Simplifying the task in this way would undermine or reduce the complex deictic relational responding that the IRAPs were aiming to capture.

The challenge was to develop an IRAP that facilitated responding to complex relational networks while maintaining sensitivity to self versus other. The two current experiments addressed this issue in two ways: 1. Self- and other-pictures were employed in an IRAP to ensure that the functions of self and other relational networks were controlling participants responding; 2. Perspective-taking scenarios or vignettes, in the form of False Belief tasks (see below) were presented before IRAP blocks, rather than presenting complex perspective-

taking statements within each IRAP trial. The key question was, would the observed IRAP effects be consistent with the false belief vignette?

Experiment 1

The strategy we adopted involved avoiding the use of complex statements within the IRAP, but instead presenting vignettes that required perspective-taking just prior to the completion of an IRAP. In other words, the vignettes were designed to produce different patterns for self versus other trial types within a single IRAP. The vignette was based on one of the most widely used formats for exploring perspective-taking in the ToM literature, namely the Change in Location task. This task was designed to assess the attribution of false beliefs (see Perner et al., 1989). Specifically, the false belief vignette comprises a written paragraph that described a scenario involving the participant and the other person depicted in the other-face picture. In this scenario, the participant observed that items in a box switched locations when the other person left the room. A belief IRAP was then presented that required the participant to respond to what they thought was in the box and what they thought that the other person thought was in the box. Given that the items had been switched when the other person had the left the room, the self and other should differ in terms of what they believed to be in the box.

The key question was, would the observed IRAP effects be consistent with the false belief vignette? A control vignette was presented to half of the participants, in which there was no exchange of the items in the box and therefore no false belief attribution was required. At this stage, we were interested in determining if differential patterns of responding would be observed in the IRAP performances across the two conditions (false belief versus control). Participants were also asked to complete six questionnaires. Given the exploratory nature of the study, we made no formal predictions concerning the extent to which the IRAP in the two conditions would produce different outcomes or how performance on those IRAPs might correlate with responses to the questionnaires. Exploratory in this context refers to the fact that, as far as we were aware, no previously published study had attempted to examine the impact of a false belief vignette on an IRAP designed to assess perspective-taking.

Method

Design. There were three stages in Experiment 1: 1. Familiarisation IRAP; 2. Condition vignette and belief IRAP; 3. Questionnaires.

Participants. Seventy-five participants were recruited for Experiment 1, 63 females and 12 males. Participants ranged from 17-34 years (M = 21.27) and were recruited through random convenience sampling from the **XX** participant pool. Each participant was paid an hourly rate of 10 euro. The general strategy for recruiting numbers of participants was guided by the results of a recent meta-analysis of IRAP effects in the clinical domain, indicating that a minimum of 29 is required to achieve a power of 0.8 for first-order correlations (Vahey, et al., 2015). Because participants sometimes failed to reach various performance criteria for the IRAP (details provided subsequently), it was necessary to recruit more than 29 participants to yield an adequate dataset for analyses.

Materials and apparatus. Experiment 1 involved two computer-based tasks presented on standard computers, the familiarisation IRAP and the belief IRAP. The belief IRAP involved two pictures of faces selected by each participant, one picture presented the participant's face, while the other picture was the face of a stranger considered by the participant to be similar in looks to themselves (i.e. same gender, age, hair colour, skin colour and eye colour). Two short vignettes were also constructed for current purposes, with each pertaining to one of the two conditions (False Belief or Control). Two questionnaires, developed specifically for this experiment, assessed performance strategies and perceived physical similarities between the picture of self and the picture of other. The study also included six questionnaires: the Community Assessment of Psychic Experiences (CAPE); psychological flexibility (using the Psychological Flexibility Index, PFI); self-warmth (using a Self-warmth Thermometer); emotional attachments (using the Experiences in Close Relationships-Relationship Structures questionnaire, ECR-RS); and relationships with others (The Inclusion of Other in the Self, IOS; and the Experiencing of Self Questionnaire, ESQ). The PFI was a measure of psychological flexibility being developed by Bond and colleagues as an alternative to the AAQ. The Self-warmth Thermometer was included to determine whether performance in the self-IRAP correlated with self-warmth (Vahey, et al., 2009). The various attachment questionnaires were included because pre-existing difficulties in attachment relationships may manifest in difficulties in perspective-taking with regard to others (Bernstein, et al., 2015). All materials were presented in Dutch (translated into English when referred to in the text). The CAPE was the only questionnaire with a validated Dutch version. The instructions and items of the remaining measures were created using a backward forward translation procedure (World Health Organization, WHO, 2017). There are no clinical cut-offs for any of the measures.

Picture stimuli used in the IRAP. The face picture stimuli were collected for both IRAPs prior to the experiment. Participants were asked to bring to the experiment two pictures; a picture of themselves that they liked and a picture of an unknown other who they considered to be similar in looks to themselves (i.e. same gender, age, hair colour, skin colour and eye colour). These pictures were included in the self-picture and other-picture IRAPs, respectively.

Familiarisation IRAP. The familiarisation IRAP did not contain stimuli relevant to perspective-taking and was employed simply to familiarise participants with the procedure, because no practice blocks were presented in the subsequent belief IRAP. The IRAP was presented on standard personal computers. The IRAP software was used to present the instructions and stimuli and to record responses. The familiarisation IRAP presented two label words at the top of the screen: *Fruits* and *Vegetables*. Eight target words were individually presented in the centre of the screen; four were fruits (e.g. "Pear") and four were vegetables (e.g. "Broccoli"). The response options "Yes" and "No" were presented at the bottom left- and right-hand corners. The familiarisation IRAP comprised four trial types: *Fruit-Fruit, Vegetable-Vegetable, Fruit-Vegetable and Vegetable-Fruit*.

Belief IRAP. Each trial in the belief IRAP presented the picture of the participant or the picture of the other person as a label stimulus at the top of the screen. The target stimuli comprised 12 statements, with one presented on each trial. Six of the statements referred to beliefs about a scarf (e.g. "thinks there is a scarf*"*) and six referred to beliefs about gloves (e.g. "thinks there is a glove*"*), see Table 1. The response options "Yes" and "No" were presented at the bottom left- and right-hand corners. The four trial types were denoted as: *Self-Glove* (participant's picture-statement about a glove); *Self-Scarf* (participant's picture-statement about a scarf); *Other-Glove* (picture of other-statement about a glove); and *Other-Scarf* (picture of other-statement about a scarf); see Figure 1.

INSERT TABLE 1 AND FIGURE 1 HERE

Condition Vignettes. The two vignettes were presented on-screen in a word document. Each comprised a written paragraph that described one of two scenarios involving the participant and the other person depicted

in the other-face picture, followed by questions that served to check that participants had read and understood the vignette.

False Belief Vignette. The false belief vignette was based on a ToM Change in Location task designed to assess the attribution of false beliefs (see Perner, et al., 1989). The English translation of the vignette presented in Dutch is as follows:

> *The following sentences describe a scenario that involves you and the person in the second picture that you brought with you today. I want you to imagine that you and this person are in a room. In front of you both is a box. When you both open the lid of the box, you both see together that there is a scarf in the box and then you place the lid back on the box again. At this point, the other person leaves the room. When they are no longer in the room (but you still are) the scarf is removed from the box and is replaced with a glove. The lid is then put back on the box. At this point, the other person now returns to the room.*

After reading the vignette, participants were required to indicate which of the following statements best described what happened in the scenario: 1. "I stayed in the room and the other person left the room"; 2. "Both of us stayed in the room"; or 3. "The other person stayed in the room and I left the room"[1]. Participants were then instructed as follows: "Please remember the details of the scenario you read above as you will require information from this scenario to successfully complete the next part of the experiment."

Control Vignette. The control vignette described a similar scenario to the false belief vignette, but critically there was no change in the location of items. The English translation is as follows:

> *The following sentences describe a scenario that involves you and the person in the second picture that you brought with you today. I want you to imagine that you and this person are in a room. You have a box in front of you and the other person has a different box in front of them. When you both open the lid of the boxes, you both see*

[1] The researchers checked responses to these questions after the experiment to ensure that they were consistent with the vignette; the data were lost for five participants due to a software overwriting error, but consistency was obtained for all other participants.

together that you have a glove in your box and the other person has a scarf in their box. Then the lids are placed back on the boxes.

The same three statements, and the instruction presented above, were used to ensure that participants understood the vignette.

Strategy Questionnaire. This questionnaire was designed to identify any strategies that participants may have engaged in to successfully complete the belief IRAP (see Appendix A for questionnaire translated to English). Specifically, participants were asked three questions concerning the potential influence of the preceding vignette on their responding during the belief IRAP, rated from 1 ("*not much"*) to 5 ("*a lot"*). One question asked "How much of your responding on the computer task was influenced by the scenario that you read before and throughout the task?" The questionnaire also contained a single open-ended question regarding any strategy they may have used to complete the task. Participants, in the False Belief Condition only, were also asked about the degree of success they believed they had in taking the perspective of the other person, rated from 1 ("*not successful*") to 5 (*"very successful").*

Similarity Questionnaire. This questionnaire was designed to identify perceived similarities between the self- and other-picture (see Appendix A). Specifically, participants were asked a single question concerning overall perceived similarity between self and other, rated from 1 ("*not similar at all*") to 5 ("*very similar"*). Another five questions pertained to perceived similarity in terms of 1. hair colour, 2. age, 3. eye colour, 4. skin colour and 5. facial expression, rated from 1 ("*not similar at all*") to 5 ("*very similar"*). Two final questions focused on the attractiveness of the self-picture and the other-picture, both rated from 1 ("*not attractive*") to 5 (*"very attractive").*

The Community Assessment of Psychic Experiences (CAPE; Stefanis et al., 2002). The CAPE measures psychotic-like experiences in the general population and was employed because perspective-taking has been implicated in psychotic-like experiences (e.g. Savla, et al., 2013). The scale consists of 42 symptom items rated along three sub-scales: positive symptoms (20 items, e.g., "Do you ever feel as if there is a conspiracy against you?"), negative symptoms (14 items, e.g., "Do you ever feel that you experience few or no emotions at important events?") or depressive symptoms (eight items, e.g., "Do you ever feel sad?"). Each item is rated on two 4-point Likert scales from 0 (*never*) to 3 (*nearly always*) to indicate (1) the frequency of symptoms and (2) the level of distress associated with each symptom. The CAPE provides overall frequency

and distress scores of psychic experiences, and total frequency and distress scores for each of the three sub-scales. In order to account for partial non-responses, all scores are weighted for the number of valid answers per subscale (i.e. sum score divided by number of items completed). Overall frequency and distress scores are also weighted. In all cases, higher scores indicate greater frequency or distress regarding symptoms, although there are no clinical cut-offs for this measure. The Dutch version was completed by participants. The scale has demonstrated adequate reliability: positive dimension alpha = 0.63, negative dimension alpha = 0.64, and depressive dimension alpha = 0.62 (Konings, et al., 2006).

Psychological Flexibility Index (PFI). The PFI is a measure of psychological flexibility that was being developed, when the current study was conducted, by Bond and colleagues. At that time, the measure included 82 items. Each item is rated on a 6-point Likert scale from 1 (*disagree strongly*) to 6 (*agree strongly*), with a minimum of 82 and a maximum of 492, generated by reversing relevant items and then summing the scores. Higher scores indicate higher levels of psychological flexibility, with lower scores indicating lower flexibility. At present, there are no reliability data on this measure.

Experiences in Close Relationships-Relationship Structures questionnaire (ECR-RS; Fraley, et al., 2011). The ECR-RS assesses attachment patterns in four close relationships (mother, father, romantic partner, and best friend). Each of the four relationships is rated as a separate domain along two subscales: a) anxious attachment and b) avoidant attachment. The *anxious attachment* subscale comprises 3 items (e.g. “I'm afraid that this person may abandon me”) with a maximum possible score of 21 and a minimum of 3. The *avoidant attachment* subscale comprises 6 items (e.g. “I don't feel comfortable opening up to this person”), with a maximum possible score of 42 and a minimum of 6. Each item is rated on a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher scores indicate higher levels of avoidant attachment and anxious attachment. According to Fraley et al., the *alpha* reliabilities for the four relationship domains in the avoidant subscale are between .81 and .92, with the anxiety subscale between .83 and .87. Test-retest reliability is available for only two domains on each subscale, but is adequate (*alpha* =.65 for romantic relationships and .80 for parental relationships).

Inclusion of Other in the Self (IOS; Aron, et al., 1992). The IOS is a measure of closeness in relationships, comprising two sets of seven Venn diagrams. All Venn diagrams contain one circle that represents the self, while the other circle represents a “best friend” or “other people generally”. As such, each set of Venn diagrams represents the relationship between self and a significant other (best friend) or between self and a non-

significant other (other people generally). Seven Venn diagrams were presented in each set, with each Venn diagram differing systematically in terms of the extent of the overlap between the two circles. Specifically, in the first Venn diagram, the two circles are completely separate, whereas in the seventh Venn diagram, the two circles are almost fully overlapping, with each Venn diagram in-between showing some variation from one extreme to the other. In order to yield one overall score for the relationship between self and best friend, and one overall score for the relationship between self and other people generally, each Venn diagram is allocated a number between 1 and 7, where 1 represented the least overlap, and 7 represented the most. Hence, the maximum score for best friend/other people generally was 7, with the minimum score 1. The IOS has demonstrated adequate reliability (*alpha* = .93, see Aron et al.).

Experiencing of Self Scale (EOSS; Kanter, et al., 2001). The EOSS measures the control of others over the experience of the self. It consists of 20 items rated along four subscales (each with 5 items): casual acquaintances-absent (e.g. "My feelings are influenced by casual acquaintances when I am alone"); casual acquaintances-present (e.g. "My wants are influenced by casual acquaintances when I am with them"); close relationships-absent (e.g. "My attitudes are influenced by close relationships when I am alone"); and close relationships-present (e.g. "My actions are influenced by close relationships when I am with them"). Each item is rated on a 7-point Likert scale from 1 (*never true*) to 7 (*always true*). The maximum overall score is 140 and the minimum is 20, with high scores indicating greater control of others over the experience of self. According to Kanter et al., the scale overall has high internal consistency (*alpha* = 0.91), with internal consistency in the subscales ranging from *alpha* .83-.93.

Self-warmth Thermometer. A feeling-thermometer adapted from Vahey et al. (2009) was used as a measure of subjective self-warmth. The current measure composed an illustrated thermometer with a continuous horizontal scale from 0 (*cold*), rising in intervals of 10, to 100 (*warm*). Participants were required to indicate their self-warmth from 0-100. Given that this is not a standardised measure, there are no reliability data.

Procedure. Experiment 1 comprised three stages, with all participants completing those stages as follows: 1. Familiarisation IRAP; 2. Condition vignette and belief IRAP; and 3. Questionnaires (Strategy Questionnaire, Similarity Questionnaire, CAPE, PFI, IOS, ECR-RS, EOSS and the Self-warmth Thermometer, always presented in this order).

Stage 1: Familiarisation IRAP. The familiarisation IRAP was employed to establish competent performances on a simple word-based IRAP (*Fruits* vs. *Vegetables*) prior to completion of the belief IRAP. Participants were simply instructed to determine, based on individual trial feedback, what the task involved. Consider a trial with the label "Fruits" and the target "Pear*"*. Participants responded on each trial using either the "d" key for the response option on the left or the "k" key for the response option on the right. The locations of the response options (the words, "Yes" and "No") alternated from trial to trial in a quasi-random order, such that they did not remain in the same left-right locations for more than three successive trials.

Consistent trial blocks required responding that was in accordance with the pre-experimental verbal history of the participants: *Fruit-Fruit/Yes, Vegetable-Vegetable/Yes, Fruit-Vegetable/No* and *Vegetable-Fruit/No*. *Inconsistent* trial blocks required responding that was *not* in accordance with pre-experimental verbal relations: *Fruit-Fruit/No, Vegetable-Vegetable/No, Fruit-Vegetable/Yes* and *Vegetable-Fruit/Yes*. The familiarisation IRAP always commenced with a consistent block of trials. When participants selected the response option that was deemed correct within that block, the label, target and response option stimuli were immediately removed from the screen, and the next trial was presented after an inter-trial interval of 400 ms (the label, target and response option stimuli then appeared simultaneously at the beginning of the next trial). When participants selected the response option that was deemed incorrect for that block, the stimuli remained on the screen and a red "X" appeared beneath the target stimulus. The participants were required to select the correct response option, and only then did the program proceed directly to the 400 ms inter-trial interval (followed immediately by the next trial). Participants were required to achieve both accuracy (≥ 80% correct responding) and latency criteria (≤ 2,000 milliseconds) in every block. As is typical in IRAPs, performance feedback was presented at the end of each block. The program automatically recorded response accuracy (based on the first response emitted on each trial) and response latency (time in ms between trial onset and the emission of a correct response) on each trial.

The familiarisation IRAP differed from a typical IRAP in that it contained only practice blocks (i.e. these were not followed by test blocks). Participants were exposed to a maximum of five pairs of blocks, with 24 trials per block (12 for each type of target stimulus, fruit or vegetable). If participants achieved both accuracy and latency criteria on the first, second, third, fourth or fifth pair of blocks, they proceeded to the condition vignette and belief IRAP.

Stage 2: Condition Vignette and Belief IRAP. Following the familiarisation IRAP, participants completed the condition vignettes and belief IRAP. These were presented in a sequence, such that a condition vignette was presented before each of the three test block pairs of the belief IRAP (i.e. condition vignette-first test block pair- condition vignette-second test block pair-condition vignette-third test block pair). After reading the condition vignette, participants were required to indicate which statement out of three possible statements best described what happened in the scenario that they just read. Participants were then instructed as follows: "Please remember the details of the scenario you read above as you will require information from this scenario to successfully complete the next part of the experiment." Participants were then presented with a test block pair from the Belief IRAP. The first block always required a response pattern that was deemed consistent with the vignette (e.g. participant's picture/ "Thinks there is a glove"/ "Yes").

On each trial of the belief IRAP, the label (participant's picture or other's picture) appeared at the top of the screen, with a target statement (belief about a scarf or belief about a glove) in the centre, and the two response options ("Yes" and "No") at the bottom. The instruction "The previously correct and incorrect answers have been reversed" was presented between the first block and second block of each test block pair.

When participants completed a block of trials, the IRAP program delivered feedback on their performance during that block. A fixed set of three test block pairs was presented with no accuracy or latency criteria required for participants to progress from one block to the next. However, percentage correct and median latency were presented at the end of each block to encourage participants to maintain the accuracy and latency levels.

Stage 3: Questionnaires. The Strategy and Similarity Questionnaires were presented in paper format and all other questionnaires were presented to participants via computer using the program Psychopy (Peirce, 2007).

Results and Discussion

Questionnaire data. A summary of the means and standard deviations for questionnaires, divided according to the two conditions (False Belief and Control), are presented in Appendix B. The scores divided across the conditions did not appear to differ substantively. Independent *t*-tests indicated that there were no significant differences between conditions, except for one comparison; casual acquaintances-present sub-scale of the EOSS, $t(56) = 2.33$, $p = .02$ (all other $ps > .3$). Given the large number of comparisons (25), this single significant effect was considered to be a false positive.

The open-ended strategy question was also assessed ($N = 49$). The open-ended answers were read by one researcher who developed an initial coding frame to organise the data. These codes were then grouped into categories, according to how they were related. Following this, a second researcher independently reviewed the coding and categories developed by the first researcher. Any inconsistencies or issues raised by the second researcher were discussed and the categories were adjusted accordingly. A list of the final 11 categories is presented in Appendix C. The most common strategy recorded by participants was linking the object words and picture words together ($N = 25$). Ten participants reported that they focused only on the object word and did not read the full sentences. Five participants reported that they rehearsed the link between the object and person before starting an IRAP block.

IRAP data. Consistent with standard practice in IRAP research, mean response latencies for consistent and inconsistent blocks were initially divided according to trial-type and calculated for each participant. Based on the latency and accuracy criteria, eight participants failed to complete the familiarisation IRAP (and did not proceed to the belief IRAP). Exclusion criteria were also applied to the belief IRAP, such that participants were required to maintain an accuracy level of ≥ 79% correct and a median latency ≤ 2,000ms on the *third block pair*.[2] Eleven participants failed to maintain these criteria, five from the False Belief Condition and six from the Control Condition. Their data were excluded from further analysis (False Belief, final $N = 30$; Control, final $N = 30$).

D_{IRAP}-scores. D_{IRAP}-scores for the belief IRAP were calculated for each of the four trial types, such that positive D_{IRAP}-scores during vignette-consistent blocks indicated responding "Yes" more quickly than "No" on *Self-Glove* and *Other-Scarf* trial types and responding "No*"* more quickly than "Yes*"* on *Self-Scarf* and *Other-Glove* trial types. Negative D_{IRAP}-scores indicated the opposite pattern: responding "No" more quickly than "Yes" on *Self-Glove* and *Other-Scarf* trial types and responding "Yes" more quickly than "No" on *Self-Scarf* and *Other-Glove* trial types.

[2] Initially, we planned to include analysed data from all three block pairs, or at least two block pairs, but 47 of the 75 participants failed to meet the performance criteria on the first and/or second test block pair. Thus, only data for the third block pair were analysed and, even then, 17 participants failed to meet the performance criteria. Nevertheless, focusing on the third block pair yielded data that could be analysed for 60 participants. Note, the high attrition rate was likely due to the use of a familiarisation IRAP in the place of the usual practice blocks; this issue is addressed, however, in the next study.

The mean D_{IRAP}-scores and standard errors for each trial type, from the third test block pair, are presented in Figure 2. Positive scores were recorded for three of the four trial types, with negative scores recorded on *Self-Scarf* in both of the conditions. For each of the four trial types, the difference between the two IRAPs appeared relatively modest. A 2x4 mixed repeated measures ANOVA produced a main effect for trial type [$F(1, 58) = 8.65$, $p < .0001$, $\eta_p^2 = .29$], but not for condition ($p > .8$), or for the interaction ($p > 0.3$). Post-hoc comparisons, with the trial type effects collapsed across the two conditions (see Table 2), indicated that *Self-Scarf* differed from every other trial type. Eight one-sample *t*-tests indicated that the effects were significantly different from zero ($ps < .05$) for both the *Self-Glove* and *Other-Scarf* trial type in both conditions and for the *Self-Scarf* trial type in the Control Condition.

INSERT FIGURE 2 AND TABLE 2 HERE

Correlations. Given that no main effect emerged in comparing the False Belief and Control Conditions, the D_{IRAP}-scores for both conditions were collapsed before being subjected to correlational analyses with the questionnaires (i.e. a total of 100 correlations; 25 for each trial type). Only five significant correlations (at $p < 0.05$) emerged, with no obvious pattern or clustering around a particular trial type or self-report measure. Specifically, *Self-Glove* correlated positively with the frequency of negative psychotic-like symptoms [$r(58) = .32$, $p < 0.02$] and with distress associated with these symptoms [$r(58) = .27$, $p < 0.04$]. *Self-Scarf* correlated with the casual acquaintances-present sub-scale of the EOSS [$r(58) = .29$, $p < 0.03$]. Finally, *Other-Scarf* correlated with closeness to best friend [$r(58) = .27$, $p < 0.05$] and with greater control of others over the experience of self [$r(58) = .26$, $p < 0.05$]. None of these remained significant after Bonferroni corrections.

Summary and Conclusion. The primary objective of Experiment 1 was to explore the potential impact of a false belief and control vignette on performances in a related belief IRAP. The results indicated vignette-consistent scores for three of the four trial types, with vignette-inconsistent effects recorded on *Self-Scarf* in both conditions. There was little evidence that the two vignettes impacted differentially upon the IRAP performances. The correlational analyses failed to indicate any clear relationships between the self-report measures and the IRAP. Despite there being no significant difference between IRAP performances across the two conditions, the pattern of results suggest that both vignettes, to some degree, impacted the IRAP effects, given that 6 of the 8 trial types were in a vignette-consistent direction. Of course, the vignette-inconsistent effect for the *Self-Scarf* trial type seems somewhat anomalous. On balance, this result could be interpreted as a type of self-positive bias

effect, in which participants tended to choose a positively valanced response option (i.e. Yes) on a trial type that presented a self-related label (i.e. a picture of the self). The bias towards responding "Yes" on the *Other-Scarf* trial type would thus be seen as driven largely by the vignette rather than a self-positivity bias effect (see Finn, et al., 2018, for empirical evidence for, and a detailed discussion of, the complex manner in which individual trial types may differentially influence responding on the IRAP). As an aside, it is interesting to note that the anomalous effect for the *Self-Scarf* trial type was substantively larger for the Control Condition, where "Yes" could be interpreted as a vignette-consistent response because the self knows that there are two items (one in each box). On balance, the inferential statistics did not yield a significant difference for condition and thus further speculation seems unwarranted.

In reflecting upon the results obtained in Experiment 1, a number of issues emerged that seemed important to address in a follow-up experiment. First, it became apparent that participants in the Control Condition may have found the relationship between the vignette and the IRAP trial types somewhat ambiguous. Specifically, the control vignette specified that there were two boxes present in the room (one in front of the participant and another in front of the other person), whereas the belief IRAP presented statements that specified only one box. As such, it is difficult to interpret the IRAP effects that were observed for the Control group. A related issue pertains to the fact that the order in which the IRAP blocks were presented was not counterbalanced (i.e. the first block of the belief IRAP was always vignette-consistent). It is possible, therefore, that the IRAP performances for the False Belief group were determined largely by the vignette, whereas the performance of the Control group was simply determined by the pattern that they were required to produce on the first block. Indeed, it could be the case that the IRAP effects for the False Belief group may also have been determined, at least to some degree, by the pattern required by the first block. If this was the case, it could explain why there was limited evidence for a significant difference between the two conditions. With these issues in mind, we designed a subsequent experiment that once again sought to develop an IRAP that would show some sensitivity to a false belief vignette. In Experiment 2, therefore, the content of the vignettes and IRAP block-order were manipulated. These manipulations were designed to counterbalance the correspondence between vignette content and the contingencies in effect during initial contact with the IRAP.

Experiment 2

The main aim of Experiment 2 was to determine the extent to which false belief vignettes presented before each block of trials in a belief IRAP would impact the performances observed on that IRAP. In doing so,

two specific variables were manipulated across four conditions; 1. The sequence in which the critical stimuli were specified in the false belief vignette; and 2. the order in which the IRAP blocks were presented (i.e. vignette-consistent followed by vignette-inconsistent versus the opposite block sequence). Specifically, participants were presented with one of two false belief vignettes, both similar to that presented in Experiment 1. Half of the participants were presented a vignette in which a scarf was initially found in the box and this was later replaced with a glove; the other half were presented with a vignette in which a glove was initially found in the box and this was later replaced with a scarf. The only difference between both vignettes, therefore, was the sequence in which the stimuli were specified (i.e. scarf then glove versus glove then scarf). The main rationale for employing these versions of the same vignette was to determine if clear differential patterns of responding, consistent with the specified sequence in the vignette, would be observed in the belief IRAP performances. As noted above, the sequence in which the IRAP blocks were presented in Experiment 1 was not manipulated and thus the extent to which the vignette *per se* determined performance remained unclear. In Experiment 2, therefore, IRAP block sequence was also manipulated, thus creating a mixed 2x2 factorial design: (i) scarf-glove sequence/vignette-consistent first, (ii) glove-scarf sequence/vignette-consistent first, (iii) scarf-glove sequence/vignette-inconsistent first and (iv) glove-scarf sequence/vignette-inconsistent first.

Before continuing, it is important to note that pilot research for Experiment 2 suggested that specific parameters of the belief IRAP could be altered to both reduce attrition rates and increase the impact of the vignettes. First, the familiarisation IRAP employed in Experiment 1 was now replaced by exposure to practice blocks in the belief IRAP. Second, the vignettes were presented before exposure to each block of the IRAP (rather than before each pair of blocks). To avoid any perceived conflict between the vignette and the between-block instructions that were presented in Experiment 1, the latter were removed and replaced with general task instructions presented once at the beginning of the experiment. Given the large number of changes in Experiment 2, relative to 1, the former was deemed to be largely exploratory and thus we did not make any formal predictions.

Method

Design. Experiment 2 comprised four conditions: (i) scarf-glove sequence/vignette-consistent first, (ii) glove-scarf sequence/vignette-consistent first, (iii) scarf-glove sequence/vignette-inconsistent first, and (iv) glove-scarf sequence/vignette-inconsistent first. Participants were given general task instructions before exposure to the vignette and the IRAP. The Similarity and Strategy Questionnaires that were employed in the

previous experiment were presented after completing the IRAP; given the focus of the current experiment, the six questionnaires employed in the previous experiment were not presented here.

Participants. Seventy-four participants were recruited for the current experiment, 58 females and 16 males. Participants ranged from 18-25 years (M = 20.97). The general strategy for recruiting numbers of participants was similar to that previously described in Experiment 1.

Materials and Apparatus

General Task Instructions. Participants were presented with a sheet that provided general instructions for completing the IRAP. The English translation of the instructions presented in Dutch is as follows:

> *You will soon be performing different tasks on the computer. Before each part of the task, you will be presented with a story about you and the person whose picture you brought here today. This story will sometimes be consistent with that computer task you need to perform and other times it will be inconsistent with the computer task you need to perform.*

Condition vignettes. Each of the two vignettes was presented on-screen in a word document. Each comprised a written paragraph that described one of two scenarios involving the participant and the other person depicted in the other-face picture. Both vignettes were similar to the false belief vignette employed in the previous experiment. Based on pilot research the vignettes were modified to reduce potential ambiguity in the event described (see below).

Scarf-Glove Sequence Vignette. The scarf-glove sequence vignette specified that there was a scarf in the box first and this was replaced with a glove. The English translation of the vignette presented in Dutch is as follows:

> *The following sentences describe a scenario that involves you and the person in the second picture that you brought with you today. I want you to imagine that you and this person are in a room. In front of you both is a box. When you both open the lid of the box, you both see together that there is a scarf in the box and then you place the lid back on the box again. At this point, the other person leaves the room. Therefore, they cannot see what happens in the room but you still are in the room and you can still see*

what happens. When they are no longer in the room (but you still are) the scarf is removed from the box and is replaced with a glove. The lid is then put back on the box. At this point, the other person now returns to the room and they are not allowed to take the lid off the box.

After reading through the vignette, participants were required to indicate which of the following statements best described what happened in the scenario: 1. "I stayed in the room and the other person left the room when the items in the box were changed"; 2. "We both stayed in the room when the items in the box were changed"; or 3. "The other person stayed in the room and I left the room when the items in the box were changed"[3]. Participants were then instructed as follows: "Please remember the details of the scenario you read above. as you will require information from this scenario to successfully complete the next part of the experiment."

Glove-Scarf Sequence Vignette. The glove-scarf sequence vignette described a similar scenario to the scarf-glove sequence vignette. The only difference between both vignettes, therefore, was the sequence in which the stimuli were specified (i.e. scarf then glove versus glove then scarf). The same three statements, and the instruction presented above, were used to ensure that participants understood the vignette.

Belief IRAP. The format for the belief IRAP was similar to that presented in the previous experiment, except for the three following modifications; 1. A maximum of five practice blocks pairs were now presented before a fixed number of six test block pairs; 2. The order in which the IRAP blocks (i.e. consistent followed by inconsistent versus inconsistent followed by consistent) was manipulated across the four conditions; 3. Only the vignettes and the three statements, which ascertained participant understanding, were presented before each block (i.e. no additional rules or instructions were used).

The Strategy and Similarity Questionnaires described in the previous experiment were again used to assess performance strategies and perceived physical similarities between the self and other.

Procedure. Experiment 2 took place on an individual basis in sound-proof cubicles at the **XX** University.

[3] The researchers checked responses to these questions after the experiment to ensure that they were consistent with the vignette; only one participant in the scarf-glove sequence/vignette-inconsistent-first condition identified the incorrect scenario for one block out of the 16 presented.

General task instructions. The researcher gave participants a copy of the general task instructions. If participants asked for any clarification, the researcher provided this verbally in a brief and concise manner.

Condition vignette and belief IRAP. Participants were exposed to the belief IRAP, with the same vignette presented before each practice and test block throughout the IRAP.

Consistent blocks required responding that was in accordance with the vignette, which was labelled as follows: *Self-Correct/Yes, Self-Incorrect/No, Other-Incorrect/No* and *Other-Correct/Yes.* Inconsistent blocks required the opposite, labelled as follows: *Self-Correct/No, Self-Incorrect/Yes, Other-Incorrect/Yes* and *Other-Correct/No.*

Questionnaires. Participants completed the two questionnaires immediately after completing the belief IRAP.

Results and Discussion

Questionnaire data. A summary of the means and standard deviations for questionnaires, divided according to the four IRAP conditions, are presented in Appendix D. The similarity scores were relatively high, indicating that participants confirmed that they looked similar to the person depicted in the other picture. In general, the scores divided across the four IRAP conditions did not appear to differ substantively. For the purposes of statistical analysis only the scores from the general similarity question were entered into a one-way between-participants ANOVA and this proved to be non-significant ($p = .23$).

The means and standard deviations for the strategy scores indicate that participants perceived that they were relatively successful at taking the other person's perspective and that they felt that the vignette controlled their responding on the IRAP. The scores divided across the four IRAP conditions did not appear to differ substantively and two one-way between-participant ANOVAs, one for each question, both proved to be non-significant; given the separate analyses for each question a Bonferroni correction was applied ($p < .025$).

In addition, the open-ended strategy question was assessed ($N = 51$). The open-ended answers were read by one researcher who developed an initial coding frame to organise the data. These codes were then grouped into categories, according to how they were related. Following this, a second researcher independently reviewed the coding and categories developed by the first researcher. Any inconsistencies or issues raised by the second researcher were discussed and the categories were adjusted accordingly. A list of the final 11 categories

is presented in Appendix E. The most common strategies recorded by participants was linking the object words and picture words together ($N = 16$) and focusing on the scenario to help complete the IRAP ($N = 16$). Fourteen participants reported that they relied on the IRAP feedback and 11 reported that they focused only on the object word and did not read the full sentences. Nine participants reported that they rehearsed the link between the object and person before starting an IRAP block.

IRAP data. Due to a technical issue, the IRAP data for two participants were lost and thus were not included in the analyses. The data for a third participant were also excluded after the participant reported prior familiarity with similar IRAPs, which recent research indicates may influence IRAP performance (see Finn et al., 2018).

As noted previously, practice blocks required an accuracy level of ≥ 80% and a median latency of ≤ 2,000ms; three participants failed to achieve these criteria across five exposures and thus they did not proceed to the test blocks. Test blocks required an accuracy level of ≥ 79% and a median latency of ≤ 2,000ms (on two of the three successive pairs), which four participants failed to achieve, and thus the data for these participants were not included in subsequent analyses. Seven participants failed to maintain the accuracy and latency criteria for one of the pairs of test blocks, and thus their scores were calculated from the remaining two pairs (see Nicholson & Barnes-Holmes, 2012). The final analyses contained $N = 64$ (scarf-glove sequence/vignette-consistent-first, $N = 17$; scarf-glove sequence/vignette-inconsistent-first, $N = 15$; glove-scarf sequence/vignette-consistent-first, $N = 16$; and glove-scarf sequence/vignette-inconsistent-first, $N = 16$).

D_{IRAP}-scores. As noted previously, participants were divided into four conditions based on the vignette sequence (i.e. scarf-glove versus glove-scarf) and the IRAP block sequence (i.e. vignette-consistent-first versus vignette-inconsistent-first). D_{IRAP}-scores were calculated for each trial type, such that positive scores indicated a response bias that was consistent with the scarf-glove vignette and negative scores indicated a response bias that was consistent with the glove-scarf sequence. The data were entered into a preliminary 2x2x4 mixed repeated-measures ANOVA and this yielded a significant three-way interaction, $F(1, 60) = 18.85$, $p < .0001$, $\eta_p^2 = .24$. A main effect for vignette, $F(1, 60) = 4.41$, $p = .04$, $\eta_p^2 = .06$, and a two-way interaction for vignette and IRAP block sequence, $F(1, 60) = 23.54$, $p < .0001$, $\eta_p^2 = .28$ were also recorded. Given the highly significant three-way interaction with IRAP trial type it was decided at this point to conduct four separate 2x2 independent

ANOVAs, one for each trial type; a Bonferroni correction ($p < .0125$) was applied to the multiple follow-up ANOVAs.

A graphical representation of the four ANOVAs is presented in Figure 3. The following explanation may assist in the interpretation of the figure. If the vignette controlled the IRAP performances, then the two bars for the scarf-glove sequence should be in a positive direction, whereas the two bars for the glove-scarf sequence should be in a negative direction. The top left panel shows that the D_{IRAP}*-scores* for the *Self-Correct* trial type were vignette-consistent for both vignette sequences (scarf-glove and glove-scarf), but only when the IRAP commenced with a vignette-consistent block. The opposite appeared to be the case when the IRAP commenced with a vignette-inconsistent block. The descriptive analysis was supported by the 2x2 ANOVA, which yielded a significant interaction, $F(1, 60) = 83.83$, $p < .0001$, $\eta_p^2 = .58$, but no significant main effects (*ps* > .4). The top right panel shows that the D_{IRAP}*-scores* for the Self-Incorrect trial type were vignette-consistent, and only marginally so, for the scarf-glove sequence when the IRAP commenced with a vignette-consistent block. The 2x2 ANOVA yielded no significant main or interaction effects (all *p*s > .04). The bottom left panel shows that the D_{IRAP}*-scores* for the Other-Incorrect trial type were relatively small and all vignette-inconsistent**.** The 2x2 ANOVA yielded no significant main or interaction effects (all *ps* > .09). The bottom right panel shows that the pattern of D_{IRAP}*-scores* for the *Other-Correct* trial type were similar (albeit weaker) to the pattern observed for the *Self-Correct* trial type. The 2x2 ANOVA, yielded a significant interaction, $F(1, 60) = 15.23$, $p = .0002$, $\eta_p^2 = .20$, but no significant main effects (*ps* > .4). Overall, therefore, there was little evidence that the vignette controlled the IRAP performances for any of the four trial types. Indeed, for the *Self-Correct* and *Other-Correct* trial types the pattern of IRAP effects suggests that the primary controlling variable was the order in which the IRAP blocks were presented (vignette-consistent-first versus vignette-inconsistent-first).

INSERT FIGURE 3 HERE

At this point in the analyses of the data from the IRAP it appeared that the false belief vignette had virtually no impact on the response biases recorded during the test blocks. Instead, the response pattern required during exposure to the first block of trials seemed to drive the IRAP response biases. In drawing this conclusion, however, it may be premature to assume that the vignette had *no impact whatsoever* on the IRAP performances. For example, perhaps the vignette was a controlling variable, but only when it cohered with the initial exposure to the IRAP. Or to put it another way, if participants perceived the vignette to be an accurate guide on how to

respond on the IRAP they simply continued to be guided by both sources. If, however, participants perceived the vignette to be an inaccurate guide, then they simply ignored the vignette and treated the first block of IRAP trials as the 'correct' pattern. If this post-hoc interpretation of the findings is correct, then perhaps performances on the IRAP may differ during the practice blocks (i.e. when participants first encounter either coherence or incoherence between the vignette and the feedback contingencies of the IRAP). To test this suggestion, we simply compared the number of practice blocks that participants required in the vignette-consistent-first versus the vignette-inconsistent-first conditions. The difference proved to be significant with the consistent-first group requiring a mean of 2.03 (SD = 1.12) practice blocks to reach the accuracy and latency criteria versus a mean of 2.93 (0.96) for the inconsistent-first group.

To further explore the potential impact of coherence between the vignette and initial exposure to the IRAP, we analysed the individual D_{IRAP}-scores from the first pair of practice blocks. Although this analytic strategy is rarely if ever adopted in IRAP research (because the IRAP performances could not be considered relatively stable in terms of the desired stimulus control) it seemed reasonable to adopt it here to address the post-hoc question we were asking. We restricted our analysis to the first pair of practice blocks because a large number of the participants (N = 18), particularly in the consistent-first condition (N = 16), only required one pair of practice blocks before proceeding to the test blocks.

The data from the first pair of practice blocks were entered into a preliminary 2x2x4 mixed repeated measures ANOVA, and this yielded a two-way interaction for vignette and IRAP block sequence, $F(1, 60) = 62.71$, $p < .0001$, $\eta^2_\rho = .51$ and a main effect for vignette, $F(1, 60) = 52.36$, $p < .0001$, $\eta^2_\rho = .62$. A graphical representation of the interaction for vignette and IRAP block sequence is presented in Figure 4. The graph shows that the D_{IRAP}-*scores* for the vignette-consistent-first conditions was marginally vignette-inconsistent for both vignette sequences (scarf-glove sequence and glove-scarf sequence). Similar, but far stronger effects, were observed for the vignette-inconsistent-first conditions. Given the highly significant two-way interaction, it was decided to conduct four follow-up unpaired *t*-tests; a Bonferroni correction ($p < .0125$) was applied. Only one of the four *t*-tests proved to be non-significant; the comparison between the two vignette-consistent-first conditions ($p > .23$; remaining $ps < .0001$). The pattern of effects for the first pair of practice-blocks suggests that when the vignette and the initial IRAP contingencies cohered, block sequence had a limited impact on IRAP performance. However, when the vignette and the IRAP contingencies did not cohere (during initial contact with the IRAP), block sequence was a dominant controlling variable.

INSERT FIGURE 4 HERE

Overall, the results of this second experiment appear to confirm that the false belief vignettes had a limited impact on the IRAP performances during the test blocks. Indeed, the primary controlling variable, at least with respect to the *Self-Correct* and *Other-Correct* trial types, was the order in which the two types of IRAP blocks were presented. In other words, the observed IRAP effects were in a direction that was consistent with the contingencies that were contacted during the first block of trials on the IRAP, rather than the content of the vignettes. On balance, the vignettes did not appear to be completely inert, as controlling variables, because the participants in the vignette-consistent-first conditions required fewer practice blocks to reach criteria than participants in the vignette-inconsistent-first conditions. Furthermore, there were large and significant differences in the actual overall IRAP effects during the initial pair of practice blocks across the two sequences (i.e. vignette-consistent-first versus vignette-inconsistent-first).

General Discussion

The primary objective of Experiment 1 was to determine if the presentation of a false belief vignette before exposure to a single IRAP would influence performance in a vignette-consistent direction, thus showing that the IRAP could capture complex perspective-taking. Specifically, the false belief vignette comprised a written paragraph that described a scenario involving the participant and the other person depicted in the other-face picture. In this scenario, the participant observed that items in a box switched locations when the other person left the room. A belief IRAP was then presented that required the participant to respond to what they thought was in the box and what they thought that the other person thought was in the box. Given that the items had been switched when the other person had the left the room, the self and other should differ in terms of what they believed to be in the box. The key question was, would the observed IRAP effects be consistent with the false belief vignette? A control vignette was presented to half of the participants, in which there was no exchange of the items in the box and therefore no false belief attribution was required. However, the results were inconclusive and there is a number of possible reasons for this outcome. (1) The vignette presented in the Control Condition could have been interpreted as ambiguous for participants. (2) High rates of attrition restricted data analyses to the final pair of test blocks. (3) The order in which the IRAP blocks were presented (i.e. vignette-consistent-first versus vignette-inconsistent-first) was not counterbalanced.

Naturally, the primary aim of Experiment 2 was to rectify these three issues by: (i) including an active control design in which the two vignettes specified the opposite states of affairs; (ii) included practice blocks in the IRAP; and (iii) counterbalancing the block sequence of the IRAP. The results Experiment 2 were more conclusive, but still suggested that the primary controlling variable was the sequence in which IRAP blocks were presented, rather than the actual content of the vignettes. Indeed, the primary controlling variable, at least with respect to the *Self-Correct* and *Other-Correct* trial types, was the order in which the two types of IRAP blocks were presented. Thus, the extent to which participants were responding to self versus other appears to be limited. Nevertheless, post-hoc analyses did indicate that the vignettes did impact on performance, but only when the vignette and the initial contingencies of the IRAP cohered with each other.

It remains unclear exactly what relational repertoires seem to be required to complete the types of task that aim to assess what is described as false belief, such as the Deceptive Container Task or the Sally-Anne Test, which was employed in the current research. Although largely speculative, we present below a model of the relational repertoires that may be required as stipulated in a recent article by Kavanagh, et al. (2020). The details of that model are quite extensive and so they will not be repeated here. Rather a brief summary of the model will be presented so that the reader may appreciate why the IRAP may have failed to capture the perspective-taking effects we attempting to analyze in the current research.

The reader should first examine Figure 5, which provides a graphical representation of the suggested functional-analytic processes involved in responding correctly to the classic Unexpected Location (false belief) Task.

INSERT FIGURE 5 HERE

The left-hand side of the figure indicates that initially (at Time 1) both the self and other observe a glove being placed into a box; based on this the self can conclude that both self and other know that there is a glove in the box. The right-hand side of the figure indicates that subsequently (at Time 2) the self observes the glove in the box being replaced with a scarf when the other is not in the room; based on this, the self can conclude that only the self will know that there is a scarf (rather than a glove) in the box. The double-headed arrow linking the left and right sides of the figure indicates that responding correctly to the false belief task requires that the self relates the two relational networks as distinct in terms of what the self and other know after Time 2. The critical point here is that if the self simply reported that the other *does not know* what is in the box after Time 2, that would indicate relating relations If, however, the self-reports that the other thinks that the box contains a glove,

that requires the relating of relations at Time 2 to the relating of relations at Time 1. More informally, the self has to understand that what the other knew at Time 1 is what they still think at Time 2.

Deconstructing a classic false belief task in terms of relating relations, as we have done here, clearly reveals the layers of complexity involved in this widely used task and may explain to some extent why many young children struggle to solve it correctly. At this point, we should be clear that the current relational interpretation of the false belief task remains highly speculative. Nevertheless, if the foregoing relational analysis of perspective-taking is at least partially correct, it may help to explain why it has proven so difficult to capture perspective-taking using the IRAP. Consider two key issues in this regard. 1. Responding to the IRAP requires participants to respond under time pressure (typically < 2000ms). When considering the above conceptual analysis it would appear unlikely that participants could engage in such complex relational responding within such a short time period. 2. The repeated presentation of similar trial types may also impact upon the likelihood of participants responding with 'genuine perspective-taking'. For example, it may be the case that participants initially engage in 'genuine perspective-taking' during the first trials of the IRAP, but thereafter simply maintain 'correct' responding. In other words, participants can respond correctly across latter trials that require only simple relational responses, such as mutual entailment, but they are not relating relational networks. Overall therefore, it may be the case that in the format presented in the current study, the IRAP as a methodology is,limited in its ability to capture perspective-taking 'in flight' at least after participants have completed the initial trials in the first two blocks of the procedure. However, if the IRAP was combined with a "think aloud" procedure and data were collected from the practice blocks this might allow us to track the shift from relating relational networks (i.e., when the participants are actually perspective-taking) to something closer to mutual entailing (i.e., when participants are simply confirming self-glove/other-scarf, etc. without working through all of the networked relations to derive the correct response). In this way, the IRAP may better investigate the dynamics of "genuine" perspective-taking through the recording of think-aloud protocols rather than simply differences in reaction times.

Availability of Data and Materials

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Compliance with Ethical Standards

Conflict of Interest Declaration: On behalf of all authors, the corresponding author states that there is no conflict of interest

Ethical approval: All procedures in the current study were in accordance with the ethical standards of the institutional research committee, and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical approval was approved by the Ethics Committee of the Psychology Department, Ghent University, Belgium. Informed consent was obtained for all participants.

References

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*, 596-612. DOI: 10.1037/0022-3514.63.4.596

Barbero-Rubio, A., López-López, J. C., Luciano, C., & Eisenbeck, N. (2016). Perspective-taking measured by implicit relational assessment procedure (IRAP). *The Psychological Record, 66(2),* 243-252. DOI: 10.1007/s40732-016-0166-3

Barnes-Holmes, Y. (2001). *Analysing relational frames: Studying language and cognition in young children* (Unpublished doctoral thesis). National University of Ireland Maynooth.

Bernstein, R. F., Laurent, S. N., Nelson, B. W., & Laurent, H. K. (2015). Perspective-taking induction mitigates the effects of partner attachment avoidance on self-partner overlap. *Personal Relationships, 22*, 356-367. DOI: 10.1111/pere.12085

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect in the IRAP: Developing a differential arbitrarily applicable relational responding effects (DAARRE) Model. *The Psychological Record, 68(1),* 11-25. DOI: 10.1007/s40732-017-0262-z

Fraley, R. C., Heffernan, M. E., Vicary, A. M., & Brumbaugh, C. C. (2011). The Experiences in Close Relationships-Relationship Structures questionnaire: A method for assessing attachment orientations across relationships. *Psychological Assessment, 23*, 615-625. DOI:10.1037/a0022898

Galinsky, A. D., & Ku, G. (2004). The effects of perspective-taking on prejudice: The moderating role of self-evaluation. *Personality and Social Psychology Bulletin, 30(5)*, 594-604. DOI: 10.1177/0146167203262802

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective-taking and empathy in negotiations. *Psychological Science, 19(4),* 378-384. DOI: 10.1111/j.1467-9280.2008.02096.x

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The implicit relational assessment procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science*, *2(3-4)*, 105-119. DOI: 10.1016/j.jcbs.2013.05.002

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition.* New York: Kluwer Academic.

Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development, 13(4),* 590-619. DOI: 10.1111/j.1467-9507.2004.00285.x

Kanter, J. W., Parker, C. R., & Kohlenberg, R. J. (2001). Finding the self: A behavioral measure and its clinical implications. *Psychotherapy, 38*, 198-211. DOI:10.1037/0033-3204.38.2.198

Kavanagh, D., Barnes-Holmes, Y., & Barnes-Holmes, D. (2020). The study of perspective-taking: Contributions from mainstream psychology and behavior analysis. The Psychological Record, 70, 581-604. DOI:10.1007/s40732-019-00356-3s

Kavanagh, D., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Finn, M. (2018). Exploring differential trial type effects and the impact of a read-aloud procedure on deictic relational responding on the IRAP. *The Psychological Record, 68*, 163-176. DOI:10.1007/s40732-018-0276-1

Kavanagh, D., Matthyssen, N., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Vastano, R. (2019). Exploring the use of pictures of self and other in the IRAP: Reflecting upon the emergence of differential trial type effects. *International Journal of Psychology and Psychological Therapy, 19(3),*323-336.

Kavanagh, D., Roelandt, A., Van Raemdonck, L., Barnes-Holmes, Y., Barnes-Holmes, D., & McEnteggart, C. (2019). The On-going Search for Perspective-taking IRAPs: Exploring the Potential of the Natural Language IRAP. *The Psychological Record, 69(2),* 291-314.

Konings, M., Bak, M., Hanssen, M., Van Os, J., & Krabbendam, L. (2006). Validity and reliability of the CAPE: A self-report instrument for the measurement of psychotic experiences in the general population. *Acta Psychiatrica Scandinavica, 114*, 55–61. DOI:10.1111/j.1600-0447.2005.00741.x

Mead, G. H. (1934). *Mind, self, and society*. Chicago: Chicago University Press.

Montoya-Rodríguez, M. M., Molina, F. J., & McHugh, L. (2017). A review of relational frame theory research into deictic relational responding. *The Psychological Record, 67(4),* 569-579. DOI: 10.1007/s40732-016-0216-x

Nicholson, E., & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record*, *62*, 263–278. DOI: 10.1007/BF03395801

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8-13. DOI: 10.1016/j.jneumeth.2006.11.017

Perner, J., Frith, U., Leslie, A. M., & Leekman, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development, 60*, 689-700. DOI: 10.2307/1130734

Piaget, J. (1948). *The moral judgment of the child*. New York: Free Press.

Savla, G. N., Vella, L., Armstrong, C. C., Penn, D. L., & Twamley, E. W. (2013). Deficits in domains of social cognition in schizophrenia: A meta-analysis of the empirical evidence. *Schizophrenia Bulletin, 39(5),* 979-992. DOI: 10.1093/schbul/sbs080

Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental Psychology, 51(9),* 1190-1200. DOI: 10.1037/dev0000039

Stefanis N. C., Hanssen M., Smirnis N. K., Avramopoulos, D. A., Evdokimidis I. K., Stefanis, C. N.,…& Van Os, J. (2002). Evidence that three dimensions of psychosis have a distribution in the general population. *Psychological Medicine*, *32*, 347–358. https://doi.org/10.1017/S0033291701005141

Vahey, N., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). A first test of the Implicit Relational Assessment Procedure (IRAP) as a measure of self-esteem: Irish prisoner groups and university students. *The Psychological Record, 59*, 371-388. DOI:10.1007/BF03395670

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the implicit relational assessment procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry, 48*, 59-65. DOI: 10.1016/j.jbtep.2015.01.004

Vescio, T. K., Sechrist, G. B., & Paolucci, M. P. (2003). Perspective taking and prejudice reduction: The mediational role of empathy arousal and situational attributions. *European Journal of Social Psychology, 33(4)*, 455–472. DOI: 10.1002/ejsp.163

World Health Organization (2017). Process of translation and adaptation of instruments. Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en