






Differential Trial-Type Effects in an Implicit Relational Assessment Procedure: Extending the DAARRE Model

Rodrigo Vianna de Almeida¹  · Dermot Barnes-Holmes¹  · Julian C. Leslie¹ 

Accepted: 25 June 2024
© The Author(s) 2024

Abstract

The differential arbitrarily applicable relational responding effects (DAARRE) model explains two effects commonly observed with the implicit relational assessment procedure (IRAP): the single trial-type dominance effect (STTDE) and the dissonant-target trial-type effect (DTTTE). We sought to explore variables that modulate these effects. Forty-two participants who completed an IRAP met the practice criteria (median latency $\leq 2,000$ ms and accuracy $\geq 80\%$ correct) and thus proceeded to six test blocks of trials. The IRAP had four trial-types: trial-type 1: happy-face—happy-word; trial-type 2: happy-face—fear-word; trial-type 3: fear-face—happy-word; trial-type 4: fear-face—fear-word. Participants were randomly assigned to start with a consistent (respond True, False, False, True, respectively, to the four trial-types) or inconsistent (opposite responding) block. Difference (D_{IRAP}) scores were calculated (inconsistent minus consistent latencies). A significantly larger D_{IRAP} score was observed in trial-type 2 relative to trial-type 3 (a DTTTE), but only for participants who started with the inconsistent block. A “happiness superiority” STTDE (i.e., larger D_{IRAP} score on trial-type 1 relative to trial-type 4) was observed, but only for participants who failed to maintain the criteria at the trial-type level (in any of the trial-types). Overall, the findings highlight the importance of distinguishing between the functional versus relational properties of the stimuli presented within an IRAP. The modulation of the DTTTE is interpreted in terms of functional coherence between the target and response option in inconsistent blocks. A novel data-analytic algorithm is presented to identify performance criteria violations at the trial-type level. The findings suggest that a greater focus is required in relational frame theory on the impact of the functional properties of stimuli on relational framing.

Keywords DAARRE model · IRAP · STTDE · DTTTE · Trial-type

The study of derived relational responding may be traced back to the seminal work of Murray Sidman (1971), and in particular his research on stimulus equivalence relations (Sidman, 1994). The basic phenomenon involves training a series of overlapping matching to sample responses using arbitrary stimuli, and then observing novel or emergent matching responses in the absence of differential reinforcement. For example, a participant might be trained first in two overlapping stimulus relations (e.g., A1–B1, B1–C1). Subsequently, the participant may demonstrate symmetrical

relational responding (i.e., B1–A1, C1–B1), transitive relational responding (A1–C1), and combined symmetrical and transitive relational responding (i.e., C1–A1). When each of these derived relational responses emerge, the three stimuli are said to participate in an equivalence class. Research on stimulus equivalence, and derived relational responding in general, has attracted increasing attention within the behavior analytic literature (e.g., Barnes-Holmes et al., 2017; Dougher, 2020; Tonneau, 2001). One reason for this interest is that equivalence responding emerges in the absence of differential reinforcement, and thus presents a challenge to a three-term contingency explanation for the behavior. A second reason for the interest is that there appears to be a relationship between equivalence responding and human language (Barnes-Holmes et al., 2005; Bortoloti et al., 2014; Devany et al., 1986; Dias et al., 2020; Sidman, 2018).

One line of research in the study of equivalence and derived relational responding has involved using some of

This work is part of Rodrigo de Almeida’s doctoral research at Ulster University as a requirement for the degree of PhD.

✉ Rodrigo Vianna de Almeida
de_almeida-r@ulster.ac.uk

¹ School of Psychology, Ulster University, Cromore Road, Coleraine BT52 1SA, Northern Ireland

the concepts and procedures employed in the area to develop methods that may be sensitive to specific verbal histories that occurred outside of the experimental laboratory. The basic approach involves training participants in a series of matching tasks that should generate specific equivalence relations, but these relations likely conflict with previously established verbal relations. In one of the earliest studies in this area, participants resident in Northern Ireland were trained to match stimuli that, in principle, would lead to the formation of equivalence relations between Catholic names and Protestant symbols (Watt et al., 1991). In the social context of Northern Ireland, however, these two sets of verbal stimuli would typically be seen as different or opposite, rather than equivalent. Consistent with their wider social context, the majority of Northern Ireland participants failed the equivalence test, but all of the participants from an English background (the control group) successfully matched the Catholic names with the Protestant symbols. This basic effect, in which preexperimental history appears to affect derived relational responding, has been replicated across a number of studies (e.g., Barnes et al., 1996; Dixon et al., 2006; Haydu et al., 2015, 2019; Leslie et al., 1993).

The conceptual basis of the foregoing studies, in which the preexperimental and experimental histories are conflicted, led to the development of a procedure that was designed to assess relational responding “in flight.” The method, known as the implicit relational assessment procedure (IRAP; Barnes-Holmes & Harte, 2022; Barnes-Holmes et al., 2008; McKenna et al., 2007), emerged out of an account of equivalence, and derived relational responding in general, known as relational frame theory (RFT; Hayes et al., 2001). According to the theory, equivalence is but one class of relational responding, and the IRAP was designed to measure responding in accordance with networks of multiple relations, rather than equivalence alone.

The IRAP employs three sets of stimuli: label stimuli, target stimuli, and response option stimuli. For example, labels can be pictures of faces and targets can be adjectives (e.g., “happy,” “fearful”). There are two classes of labels and two classes of targets, usually in contrast to each other. On each trial, participants are presented with a label stimulus at the top of the screen and a target stimulus at the bottom of the screen. The label (L) and the target (T) stimuli are selected from classes 1 and 2, thus yielding four possible combinations: L1—T1, L1—T2, L2—T1, and L2—T2. The two response options (e.g., “true” and “false”) appear on every trial, and are used to indicate the relational coherence or incoherence between label and target stimuli. For example, assume L1 are pictures of happy faces and L2 are pictures of fearful faces, and T1 are happiness words (e.g., “cheerful”) and T2 are fear words (e.g., “fearful”). If the response options are “true” and “false,” then coherent relational responses (based on participants’ verbal histories) would

be: L1—T1 is true, L1—T2 is false, L2—T1 is false, and L2—T2 is true; and incoherent relational responses would be the opposite (e.g., L1—T1 is false). One out of these four possible combinations between a label and a target stimulus is presented on each trial; we will therefore refer to them as trial-type 1: happy-face—happy-word; trial-type 2: happy-face—fear-word; trial-type 3: fear-face—happy-word; and trial-type 4: fear-face—fear-word.

Within blocks of trials, participants are asked to respond to these relations under time and accuracy performance criteria. For example, participants may be asked to pick a response option within 2 s and perform with at least 80% correct responses across the trials of a block. For each block of trials, participants are required to respond in one of two opposing patterns: one is deemed coherent with their history, and the other is deemed incoherent. These two types of blocks are alternated such that if an IRAP starts with the coherent block, then the following blocks will be incoherent, coherent, incoherent, and so on; or, if an IRAP starts with the incoherent block, then the following sequence will be coherent, incoherent, coherent, and so on. Response latencies are recorded across coherent blocks and across incoherent blocks. D_{IRAP} scores are calculated by subtracting the latencies of the coherent blocks from that of the incoherent blocks, and dividing them by the standard deviation across blocks. Thus, if participants’ response latencies are on average smaller in coherent than in incoherent blocks, their D_{IRAP} scores will be positive; otherwise, they will be negative. Four D_{IRAP} scores, one for each trial-type, are typically calculated for each participant (see Barnes-Holmes et al., 2010a, p. 533).

The basic assumption behind the IRAP is that history-consistent relational responding is expected to be more probable and quicker than history-inconsistent responding, and this is reflected in the differential response latencies across coherent versus incoherent blocks. This assumption has been supported across numerous empirical studies in which participants respond more quickly in blocks of trials that are coherent than incoherent with their verbal histories (e.g., Barnes-Holmes et al., 2009, 2010c; Kelly & Barnes-Holmes, 2013; Rabelo et al., 2014; Roddy et al., 2010; Sereno et al., 2021; Timmins et al., 2016). An early explanation for such IRAP performances was formalized in the relational elaboration and coherence (REC) model (Barnes-Holmes et al., 2010a). The REC model focused largely on the coherence of the relationship between the label and the target; thus, responses on the coherent blocks were assumed to be faster than those on the incoherent blocks (e.g., given a happy face and a happy word, participants would pick the response option “true” more quickly than “false”). This explanation focused on performance differences between coherent and incoherent blocks, but not between trial-types within the blocks. However, differential trial-type effects have been

observed (e.g., Finn et al., 2018, 2019), which the REC model could not readily explain.

There are two specific differential trial-type effects that have been observed with the IRAP. For example, Kavanagh et al. (2019) exposed participants to IRAPs that involved presenting, as label stimuli, face words (i.e., “Face,” “Head,” and “Person”) and pen words (i.e., “Pen,” “Stylo,” and “Bic”), and, as target stimuli, pictures of a face or a pen. Participants were required, during coherent blocks, to respond “Yes” when a face word was presented with a picture of a face (trial-type 1), and when the pen word was presented with a picture of a pen (trial-type 4); in addition, during incoherent blocks, they were required to respond “No” to these combinations of stimuli. Furthermore, in coherent blocks, participants were required to respond “No” when presented with a face word and a picture of a pen (trial-type 2), and when presented with a pen word and a picture of a face (trial-type 3); in addition, during incoherent blocks, they were required to respond “Yes.”

Results from Kavanagh et al. (2019) indicated that all group effects were in the predicted direction (i.e., shorter latencies in the coherent relative to the incoherent blocks). However, two key differences emerged between the trial-types that could not be explained based solely on the relations between label and target stimuli. In particular, the difference in latencies for the face-face trial-type (i.e., trial-type 1) was significantly larger than for the pen-pen trial-type (i.e., trial-type 4). Furthermore, the difference in latencies for the face-pen trial-type (i.e., trial-type 2) was significantly larger than for the pen-face trial-type (i.e., trial-type 3). In both cases, the two trial-types shared the same response option within blocks. That is, the face-face and the pen-pen trial-types (i.e., 1 and 4) both required responding “Yes” during coherent blocks, and “No” during incoherent blocks; moreover, the face-pen and the pen-face trial-types (i.e., 2 and 3) both required responding “No” during coherent blocks, and “Yes” during incoherent blocks. Perhaps any difference between the face-face and pen-pen trial-types, relative to the face-pen and pen-face trial-types, could be explained by the fact that they required choosing different response options within blocks of trials (Barnes-Holmes et al., 2010b). However, differential response options could not explain the difference between the face-face and the pen-pen trial-types, or the difference between the face-pen and the pen-face trial-types. In order to explain these differences, Kavanagh et al. (2019) drew on a model of IRAP performances that had been recently proposed in the literature (Finn et al., 2018; Kavanagh et al., 2018).

The model is referred to as the differential arbitrarily applicable relational responding effects (DAARRE) model. This model is shown in Fig. 1, which presents the stimuli used in the current study. In particular, pictures of happy and fearful faces are presented as label stimuli, and words

denoting happiness and fearfulness are presented as target stimuli, with two response options: the words “true” and “false.” Similar to the REC model, the DAARRE model incorporates the relationship between the label and the target (e.g., whether the relationship between the face and the word is coordinate or distinct, defined in RFT as the C_{rel} property). However, the DAARRE model also incorporates the functional properties (e.g., orienting and evoking, defined in RFT as the C_{func} property) of all of the events, including the response options.¹ In Fig. 1, the C_{func} properties of the happy faces, the happy words, and the “True” response option are all labeled with a plus sign to indicate a generally positive valence relative to the fearful faces, fear words, and the “False” response option (for this reason, the latter are marked with a minus sign).

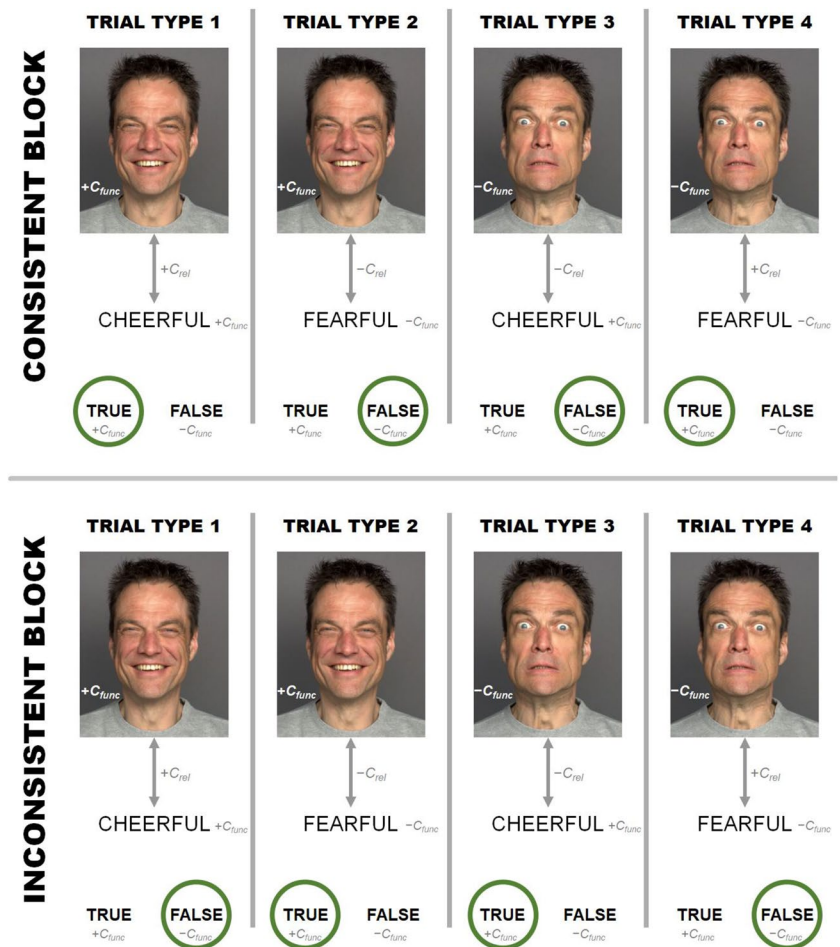
The DAARRE model explains the two differential trial-type effects mentioned above by appealing to the level of coherence among the C_{rel} and C_{func} properties contained within each of the trial-types. Consider first, the predicted² differential effect between trial-types 1 and 4; that is, the D_{IRAP} scores for the happy-happy trial-type will be larger than for the fear-fear trial-type. According to the DAARRE model, for the happy-happy trial-type, there is maximal coherence among the two C_{rel} and the two C_{func} properties during coherent blocks (i.e., four plus signs). In contrast, for the fear-fear trial-type, there is reduced coherence in that the C_{func} properties for the label and target are both negative, but the C_{rel} and the C_{func} properties for the response option are both positive. This difference in coherence between these two trial-types explains the dominance of trial-type 1 (maximal coherence) over trial-type 4 (reduced coherence). The prediction is that participants will find it easier, all things being equal, to respond on a trial-type in which all of the controlling elements cohere with each other than when they do not. We will refer to this effect as a single trial-type dominance effect (STTDE).

Now, consider the predicted difference between trial-types 2 and 3; that is, the D_{IRAP} scores for the happy-fear trial-type will be larger than for the fear-happy trial-type.

¹ The concepts of C_{rel} and C_{func} properties refer the functions of stimuli that are selected by C_{rel} and C_{func} contextual cues in any instance of arbitrarily applicable relational responding. The term “property” does not therefore refer to some intrinsic feature or essence of the stimuli.

² The DAARRE model does not make generic predictions in a hypothetico-deductive fashion, rather it provides a conceptual instrument for inductive research. The differences between trial-types 1 and 4, as well as the differences between trial-types 2 and 3, are effects commonly observed in IRAP research, which may be explained by the DAARRE model based on the dynamic interplay between C_{rel} and C_{func} properties of stimuli that arise over the course of individual learning histories (for detailed discussions, see Harte et al., 2022; Harte & Barnes-Holmes, 2024).

Fig. 1 The Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) Model with Stimuli Used in This Study's IRAP Preparation. *Note.* Participants only saw the face pictures (label stimuli), emotionally related words (target stimuli), and the response options "TRUE" and "FALSE," which were randomly replaced at either side of the screen from trial to trial. The responses that were reinforced in each block are circled. Coherent (i.e., history-consistent) blocks were named "consistent". Incoherent (i.e., history-inconsistent) blocks were named "inconsistent." C_{func} stands for "contextual stimuli controlling psychological functions." C_{rel} stands for the "relational context controlling arbitrarily applicable relational" responding



It is difficult to explain this difference based solely on the coherence/incoherence among *all* of the elements in the trial-types because, in both trial-types there is an equal number of $+C_{func}$ and $-C_{func}$, as well as the same $-C_{rel}$. However, the two trial-types may be distinguished based on the C_{func} properties of the label and the target. In trial-type 2, the label is positive and the target is negative whereas in trial-type 3, the label is negative and the target is positive. It is critical to note that during history-consistent blocks, in trial-type 2, the $-C_{func}$ property of the target (i.e., fearful words) coheres with the $-C_{func}$ property of the correct response option (i.e., "false"); however, in trial-type 3, the C_{func} properties of the target and the correct response option are incoherent (still, during the history-consistent blocks). According to the DAARRE model, the coherence between the spatially contiguous target and response option in trial-type 2 may facilitate more rapid responding during history-consistent blocks relative to trial-type 3, where the spatially contiguous target and response option possess incoherent C_{func} properties. In other words, participants may find easier to respond negatively when the target is negative than when the target is positive. Consistent with Kavanagh et al. (2019), we will call this a dissonant target trial-type effect (DTTTE).

Previously published studies (Finn et al., 2018; Gomes et al., 2019; Kavanagh et al., 2018; Pidgeon et al., 2021; Pinto et al., 2020) that have considered IRAP performances in light of the DAARRE model have tended to employ it in a post-hoc manner (i.e., the DAARRE model has been used to interpret effects that were not explicitly predicted). One recent exception was a study that successfully generated the STTDE by establishing a "true" function for a stimulus that was subsequently presented in an IRAP (Finn et al., 2019). However, this study did not focus on the other DAARRE model effect previously referred to as the DTTTE (see Schmidt et al., 2021). The primary purpose of the current study was to present participants with an IRAP for which the DAARRE model would predict both STTDE and DTTTE effects. In particular, happy faces and fearful faces were presented with semantically related words (i.e., happiness and fear words). We chose faces because previous research had reported differential trial-type effects using such stimuli (Bortoloti et al., 2019; Kavanagh et al., 2019; Perez et al., 2019; Pinto et al., 2020; Schmidt et al., 2021; see also Bortoloti et al., 2020).

In addition to testing the predictions of the DAARRE model (the STTDE and the DTTTE), we also sought to

explore the potential impact of two other variables in the IRAP. The first of these was the order in which the IRAP blocks are presented (coherent-first vs. incoherent-first). In particular, we sought to determine if block order affected the STTDE and/or DTTTE effects. The second variable we sought to explore was whether or not participants maintained criteria at the trial-type level. Research employing the IRAP typically requires participants to maintain performance criteria at the level of the block (e.g., median latency $\leq 2,000$ ms, accuracy $\geq 80\%$). Given the focus on the explicitly predicted differences between individual trial-type scores in the current study (i.e., the STTDE and the DTTTE), we analyzed the data in terms of whether or not participants maintained the performance criteria at the level of the individual trial-type, rather than at the level of the overall block. Given that this was a performance-based variable, it is best considered as an attribute variable rather than a directly manipulated variable (i.e., it is an explanatory variable based on an attribute of the participants). Again, we sought to determine if this particular performance variable affected the STTDE and/or the DTTTE. Notwithstanding, we should emphasize that analyzing the impact of these two explanatory variables (block order and performance criteria) was largely exploratory and therefore we made no specific predictions about the nature of the impact that these variables could exert on the two DAARRE model effects. These variables are, nevertheless, crucial in research employing the IRAP. Block order influences the history participants are exposed to during practice blocks before starting test blocks. Trial-type level performance is important because the DAARRE model focuses on performances at trial-types, but IRAP research has not yet looked into performance criteria at the trial-type level. The present article employs for the first time a data-analytic algorithm that was developed to identify the performance at the trial-type level.

Method

Participants

Participants were University students recruited via convenience sampling using Ulster University resources, such as Sona Systems (<https://www.sona-systems.com/>). Participants under the age of 18 or above 65, or with atypical (uncorrected) visual functioning, were not recruited. From the 50 recruited participants, 1 chose not to continue. The final sample consisted of 49 participants (34 females, 15 males). The final sample ranged from 18 to 47 years old ($M = 25.5$, $SD = 7.8$). All participants had the opportunity to read the information sheet and pose questions before filling out the consent form. Only participants who agreed to

participate were recruited. The study received ethical clearance from the local psychology research ethics committee.

Materials

A Testing-IRAP (GO-IRAP from <https://balc-i.net/go-irap/>) was configured using happy or fearful faces as label stimuli (extracted from the FACES database; Ebner et al., 2010³) and happiness-related or fear-related words⁴ as target stimuli. This generated four trial-types, as exemplified in Fig. 1. Trial-type 1 was a happy face with a happy word, trial-type 2 was a happy face with a fearful word, trial-type 3 was a fearful face with a happy word, and trial-type 4 was a fearful face with a fearful word. In consistent blocks, participants were required to respond True, False, False, and True when presented with trial-types 1, 2, 3, and 4, respectively, in order to proceed to the next trial. In inconsistent blocks, participants were required to make the opposite responses (i.e., False, True, True, and False) in order to proceed to the next trial. Each block consisted of 32 trials, with each of the four trial-types presented eight times in a quasi-random fashion.

At the start of the IRAP session, participants were presented with the following message:

This task will present sets of words and images. You will be asked to relate the words to the images. If you make a mistake you'll see a red X. Provide the correct response to continue. You will not be able to proceed unless you respond correctly. Respond as accurately as you can. When you've learned to be accurate you'll naturally speed up too.

At the end of a block, participants received feedback information on their accuracy (in correct percent) and latency (in milliseconds). Participants were also presented with one of the following messages, dependent upon their performance in the immediately preceding block. That is, the feedback messages were designed to encourage participants to maintain their performance if it met both criteria, or to adjust their performance appropriately if it failed to meet one or both criteria. In particular, if a participant met both the accuracy and latency criteria, the following message was displayed: "Continue responding both as accurately and quickly as you can." If a participant met the accuracy but not the latency criteria, the following message was displayed:

³ We used the following young and middle-aged male and female facial expressions of happiness and fear from the FACES database (<https://faces.mpg.de/imeji/>): 066_y_m_h_b.jpg, 140_y_f_h_b.jpg, 116_m_m_h_a.jpg, 168_m_f_h_a.jpg; 066_y_m_f_b.jpg, 140_y_f_f_b.jpg, 116_m_m_f_a.jpg, 168_m_f_f_b.jpg.

⁴ The words were: HAPPY, GLAD, SMILING, CHEERFUL; SCARED, ALARMED, FRIGHTENED, FEARFUL.

“Continue responding as accurately as you can. You'll naturally go quickly when your responses are accurate.” If a participant met the latency but not the accuracy criteria, the following message was displayed: “Learn to accurately follow the rule before attempting to respond quickly.” If a participant did not meet either criteria, the following message was displayed: “Learn to accurately follow the rule before attempting to respond.” At the end of the first block, and every block thereafter, participants were presented with the following message prior to the first trial of the next block: “The previously correct and incorrect answers have now been reversed.”

The response options “True” and “False” switched positions quasi-randomly from trial to trial. Participants had to press “D” on the keyboard for the response-option at the left side of the screen, or “K” for the response-option at the right side of the screen. A blank screen was presented for 400 ms in the intertrial intervals. When participants responded correctly, they proceeded to the next trial without any other differential consequence. When participants responded incorrectly, a red “X” was presented, and the trial did not proceed until the correct response was emitted. If a participant did not respond within 2,000 ms from the beginning of a trial, the warning message “2 s” appeared in red; to avoid “overloading” the participant at the beginning of the experiment, this warning message did not appear during the first pair of practice blocks.

A 7-point self-report scale was developed to measure participants' explicit ratings of all the stimuli used in the IRAP, namely: the four happy faces, the four fearful faces, the four happiness-related words, and the four fear-related words. The seven response options were: *Extremely Happy*, *Happy*, *Mildly Happy*, *Neither Happy Nor Fearful*, *Mildly Fearful*, *Fearful*, and *Extremely Fearful*. The scale was presented once for each of these stimuli (16 times in total). The order of the stimuli was randomized, but all pictures preceded all words. The scale had high face validity, but we did not pursue a formal study of its psychometric properties. Rating results for each stimulus are shown in Table 1, which indicated no overlap between the rating of happy or fearful stimuli.⁵

Procedures

Before taking part, participants were presented with an information sheet, followed by the consent form. Then, all participants undertook, in the same order, the Testing-IRAP, the self-report rating scales, and the DASS, respectively. Participants were finally debriefed by receiving a debriefing letter.

The Testing-IRAP presented practice blocks before test blocks. Participants were required to achieve specific performance criteria at the block level (median accuracy $\geq 80\%$

and latency $\leq 2,000$ ms), in both blocks of a pair of practice blocks, in order to proceed to the test blocks. A pair encompasses one history-consistent and one history-inconsistent block (the order of which depended on the starting block). Practice blocks performances were not used to calculate the D_{IRAP} scores (explained subsequently); these scores were calculated exclusively from the test blocks.

Participants received a maximum of six pairs of practice blocks as an opportunity to learn to perform the task. The blocks alternated from history-consistent to history-inconsistent or vice versa. Participants who were randomly assigned the first block as history-consistent went on to experience the IRAP in a sequence of consistent–inconsistent pairs of blocks. Participants who were randomly assigned the first block as history-inconsistent went on to experience the IRAP in a sequence of inconsistent–consistent pairs of blocks. The same sequence continued over test blocks. Participants who achieved performance criteria in a pair of practice blocks went on to undertake exactly three pairs of test blocks.

A participant who exhausted the six pairs of practice blocks without achieving the performance criteria was given the opportunity to try again and complete the IRAP from the start (i.e., including practice blocks). Three participants agreed to complete a second IRAP, and went on to successfully complete the test blocks in their second attempt. Seven participants chose not to try again, and therefore did not proceed to the test blocks. As a result, no D_{IRAP} scores were registered for these participants. Nevertheless, their self-reports were incorporated into the data presented in Table 1 because they rated the visual stimuli, and no formal analyses associated these data with IRAP results.

Data Processing

The D_{IRAP} scores were calculated from test blocks for each trial-type, as described in Barnes-Holmes et al., (2010a, p. 533). Note, however, that in the case of the present research, we did not exclude data from participants who failed to maintain the criteria at the block level (see Hussey et al., 2015, p. 159), because one of the purposes of the study was to examine the impact of maintaining or not maintaining the performance criteria. From the table in Appendix A, it can be seen that no participant who failed

⁵ After completing the happy/fear self-report rating scales, the final material that the participants completed before being debriefed was the depression, anxiety, and stress scales (DASS; Lovibond & Lovibond, 1995). No obvious trends associating IRAP results with higher levels of symptoms were observed. Given that such symptoms were infrequent, because the sample was overall normative, these analyses are presented as Supplementary Materials (Table S1, Table S2, Figure S1) and are not further discussed here.

Table 1 Results from the Self-Report Rating Scales

	Extremely Fearful	Fearful	Mildly Fearful	Neither Happy Nor Fearful	Mildly Happy	Happy	Extremely Happy
Pictures							
Young male happy	—	—	—	—	14	31	3
Young female happy	—	—	—	—	2	30	16
Middle-aged male happy	—	—	—	—	1	9	38
Middle-aged female happy	—	—	—	—	3	32	13
Young male fearful	17	19	12	—	—	—	—
Young female fearful	27	18	3	—	—	—	—
Middle-aged male fear	32	14	2	—	—	—	—
Middle-aged female fearful	25	22	1	—	—	—	—
Words							
GLAD	—	—	—	2	27	15	4
HAPPY	—	—	—	—	2	43	3
SMILING	—	—	—	3	21	24	—
CHEERFUL	—	—	—	1	7	25	15
FRIGHTENED	24	18	5	1	—	—	—
FEARFUL	5	40	3	—	—	—	—
SCARED	14	23	9	2	—	—	—
ALARMED	5	12	26	5	—	—	—

The number of participants refer to count of people who rated in the category. Data from one participant was lost due to a technical error ($n = 48$). Zeros are indicated by dash

to maintain the criteria at the trial-type level produced latencies over 2,500 ms or accuracies below 60%. In fact, only four of the 76 scores (from the 19 participants who failed at least one criterion) fell below 70%. In effect, the participants who failed to maintain the criteria at the trial-type level generally responded close to the performance criteria.

In order to categorize participants according to whether or not they maintained the performance criteria at the trial-type level, we developed a so-called policing algorithm. The source code of the “policing algorithm” can be found online at https://github.com/rodva/TrialType_Criterion_Police. A brief presentation of the source code is presented in Appendix B. To avoid requiring specific configurations on user’s local computer, an interactive version of the algorithm that was adapted to run on web browsers, which can be accessed at <https://policing-irap.rd.psc.br/>. Although the present article presents these codes for the first time, a thorough explanation of the algorithms is beyond the scope of the current work.

To categorize participants as those who did or did not maintain the performance criteria, we checked the average latency to correct response and accuracy at each trial-type across all test blocks against the criteria (accuracy $\geq 80\%$, median latency $\leq 2,000$ ms). Participants who failed those criteria at any of the trial-types were categorized as not

having maintained the performance criteria; otherwise, they were grouped as participants who maintained the performance (see Appendix A). We did not use block-level performance for this categorization.

In the current study, we focused initially on individual-participant data for the STTDE by subtracting the trial-type 4 D_{IRAP} score from the trial-type 1 D_{IRAP} score, and for the DTTTE by subtracting the trial-type 3 D_{IRAP} score from the trial-type 2 D_{IRAP} score. In the current report, the acronym STTDE refers to the relative dominance of trial-type 1 over 4 or 4 over 1. A positive STTDE difference indicates that the D_{IRAP} score for trial-type 1 was larger than for trial-type 4 (i.e., a “happiness superiority effect”); conversely, a negative STTDE difference indicates that the D_{IRAP} score for trial-type 4 was larger than for trial-type 1 (i.e., a “fear superiority effect”). Also in the current report, we use the acronym DTTTE to refer to the dominance of trial-type 2 over 3, or 3 over 2. Therefore, a positive DTTTE difference indicates a larger D_{IRAP} score for trial-type 2 relative to trial-type 3; conversely, a negative difference indicates a larger D_{IRAP} score for trial-type 3 relative to trial-type 2. Following the analyses of the individual-participants data, group-based statistical analyses were used to compare the D_{IRAP} scores of trial-type 1 with those of trial-type 4, and the D_{IRAP} scores of trial-type 2 with those of trial-type 3.

Design

The dependent variable of interest corresponded to the D_{IRAP} scores for the four trial-types. Because the four trial-types were presented concurrently for eight times each, in random order, within each 32-trial block, each trial-type constitutes a repeated measure collected within-subject. Therefore, the dependent variable (D_{IRAP}) encompasses a within-subject factor of concurrent (simultaneous) measures. However, it is not of theoretical relevance to explore all six possible pairwise comparisons among trial-types. This is because trial-types with a different response option within a block cannot be compared to each other in a theoretically meaningful way. Therefore, the comparisons of interest are only between trial-types 1 and 4, and between 2 and 3.

There are two explanatory variables of interest: block order and performance at the trial-type level. Each of them has two between-subjects levels. The design does not include the interaction of these two factors. Therefore, we explored the effects of these two explanatory variables on D_{IRAP} scores separately. Taking advantage of the economy provided by the within-subjects factor, we employed four 2×2 factorial designs: [1] block order (starting with the consistent or the inconsistent block) with STTDE (trial-types 1 and 4), [2] block order with DTTTE (trial-types 2 and 3), [3] maintaining performance in all four trial-types over all three pairs of test blocks (yes or no) with STTDE, and [4] maintaining performance with DTTTE.

From the between-subjects factors, the only independent variable we could directly manipulate was block order. Test-blocks performance at the trial-type level could not be directly manipulated because this was a participant attribute, but it is still a between-subjects factor. For each of the four designs, mixed-effects ANOVAs were but one of the statistical techniques we employed: we also used binomial tests for proportions and chi-square tests of association, and Bayesian nonparametric analyses.⁶

Results

Figure 2 presents the participant flowchart. Of the 50 participants recruited, 1 dropped out and 7 did not attain practice-blocks performance criteria (i.e., accuracy $\geq 80\%$, median latency $\leq 2,000$ ms), leaving 42 participants in with their D_{IRAP} scores (29 females and 13 males; mean age = 25.4,

$SD = 7.7$). The policing algorithm identified participants who did not maintain the accuracy and latency criteria for each trial-type during test blocks. Based on this, we categorized participants into those who maintained both criteria across all four trial-types ($n = 23$) and those who did not ($n = 19$) (exact binomial test: $p = 0.644$, one-sample proportions test: $\chi^2_{(1)} = 0.21$, $p = 0.643$). That is, any participant who did not maintain the accuracy and/or the latency criteria (averaged across test blocks) on at least one trial-type was categorized as not maintaining the criteria.⁷

All participants were randomly allocated to start with either the consistent or the inconsistent block. Nevertheless, 3 out of the 42 participants required a second exposure to the practice blocks. The second exposure to the practice blocks commenced with a block opposite to that of the first exposure (e.g., a participant who started with the consistent block in the first exposure started with the inconsistent block in the second exposure). As a result, it was not possible to categorize them unambiguously as consistent- or inconsistent-first. Hence, we excluded them from any analysis based on block order; therefore, we only considered 20 participants as starting with the consistent block and 19 participants starting with the inconsistent block (exact binomial test: $p = 1$, one-sample proportions test: $\chi^2_{(1)} = 0.00$, $p = 1$). Nevertheless, when analyzing data based on maintaining criteria across trial-types, all 42 participants were included in the analyses. Given that the current study aimed to replicate the STTDE and the DTTTE observed in prior research, we conducted a preliminary analysis across the entire sample, and both were significant: trial-type 1 was larger than trial-type 4 ($t_{(41)} = 3.58$, $p < 0.001$ one-tailed) and trial-type 2 was larger than trial-type 3 ($t_{(41)} = 2.01$, $p = 0.026$ one-tailed).

Frequency-Based Categorical Comparisons of DAARRE Effects

We obtained a measure of each participant's STTDE by performing the following subtraction between two of their D_{IRAP} scores: trial-type 1 minus trial-type 4. The more positive the result of this subtraction, the larger the score for trial-type 1 relative to trial-type 4 (i.e., a "happiness superiority effect"). In contrast, the more negative the result of the subtraction, the larger the score for trial-type 4 relative to 1 (i.e., "fear superiority effect"). For the purposes of communication, we will refer to both effects under the label of STTDE. Likewise, we estimated participants' DTTTE by performing the following D_{IRAP} scores subtraction: trial-type

⁶ The binomial and chi-square tests were conducted in SPSS and double-checked in R. Finn graphs were produced using the ggplot2 package in R. The complete ANOVAs were conducted in SPSS and some effects were double-checked in R. All Bayesian analyses were run in JASP. Analyses reports can be found on OSF.

⁷ The table in Appendix A shows the D_{IRAP} scores for each participant, as well as the proportions of accurate first responses and correct response latencies in milliseconds per trial-type (as averaged across all test blocks). Participants categorized as not maintaining criteria are highlighted.

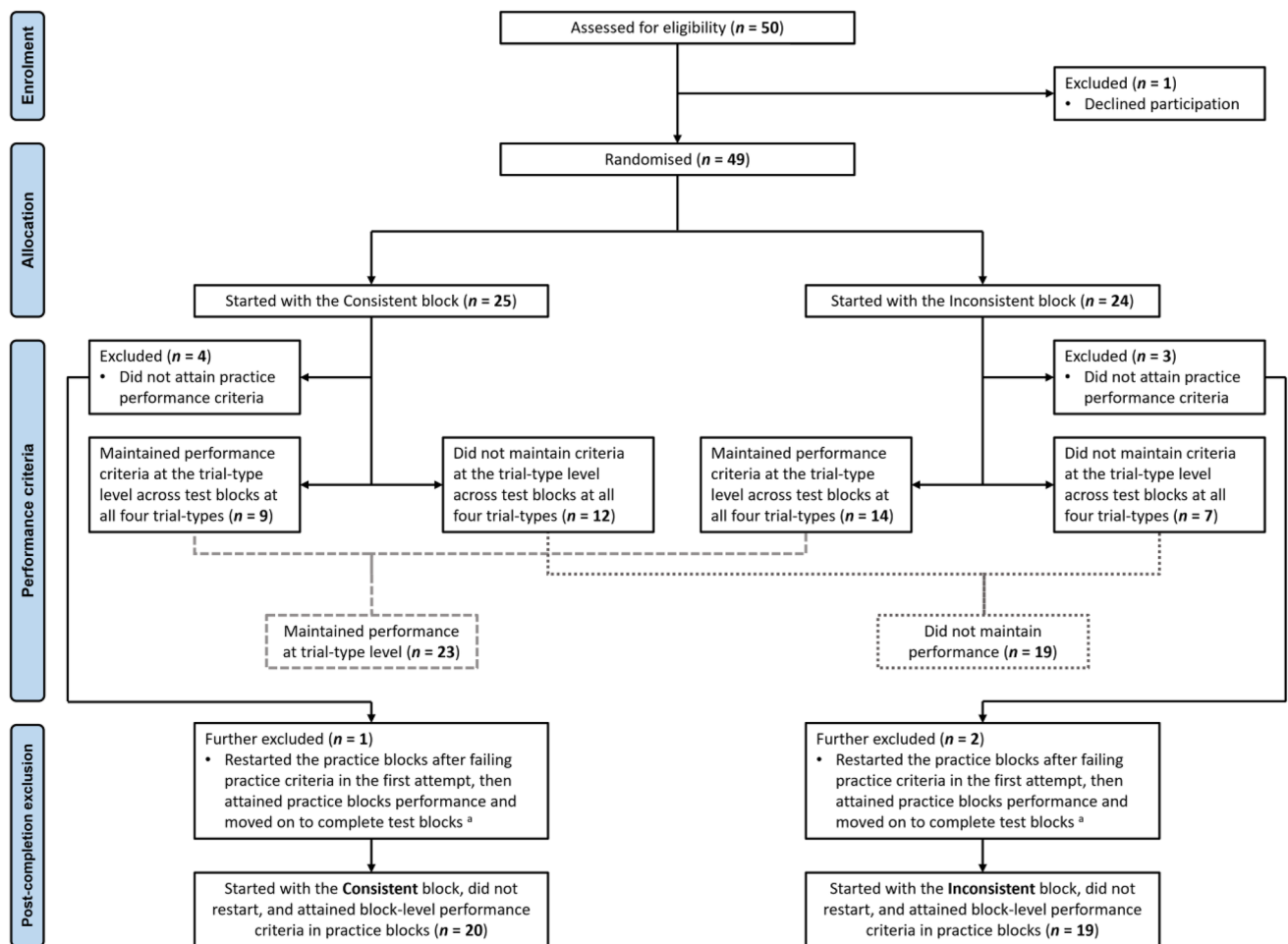


Fig. 2 Participants Flowchart. Note. Seven participants did not perform the test blocks because they exhausted their practice blocks without attaining performance criteria. Therefore, 42 participants took part in test blocks. ^a Three of the 42 valid participants had previously completed a first round of practice blocks in which they failed

to achieve the practice block-level performance criteria. Then, they chose to restart the session, after which they achieved the practice performance criteria. Because these three participants restarted the practice blocks, they were excluded from analyses based on starting block, but were included in the other analyses

2 minus trial-type 3 (again, we will use this generic label for both effects).

In Fig. 3, participants were grouped according to two variables: whether they started with the consistent or the inconsistent block (left-hand side panels), or whether or not they maintained both accuracy and latency performance criteria at the trial-type level (right-hand side panels). The top panels display the D_{IRAP} scores differences between trial-types 1 and 4 under the “STTDE” label. The bottom panels display the D_{IRAP} scores differences between trial-types 2 and 3 under the “DTTTE” label. Because this data plotting strategy was pioneered by Finn et al. (2019), these graphs will be henceforth referred to as “Finn Graphs.”

The top-left panel of Fig. 3 shows the STTDE according to starting block. Out of the 20 participants who started with the consistent block (darker bars), 16 showed a “happiness superiority effect” (i.e., bars above zero), and 4 showed a

“fear superiority effect” (i.e., bars below zero). Out of the 19 participants who started with the inconsistent block (lighter bars), 13 showed a “happiness superiority effect,” and 6 showed a “fear superiority effect.” The contingency table is presented in the top-right corner of the panel. A chi-squared analysis indicated that the occurrence of happiness or fear superiority effects was not associated with block order (Pearson’s $\chi^2_{(1)} = 0.69, p = 0.408$).

The top-right panel of Fig. 3 shows the contingency table for the STTDE and maintaining performance criteria in test blocks at the trial-type level. A chi-squared analysis indicated that the STTDE was not associated with maintaining criteria (Pearson’s $\chi^2_{(1)} = 1.94, p = 0.163$). Given that the superiority effects were not associated with either block order or maintaining criteria, an exact binomial test was applied to all 42 participants, with 31 showing a happiness superiority effect, and 11 showing a fear superiority

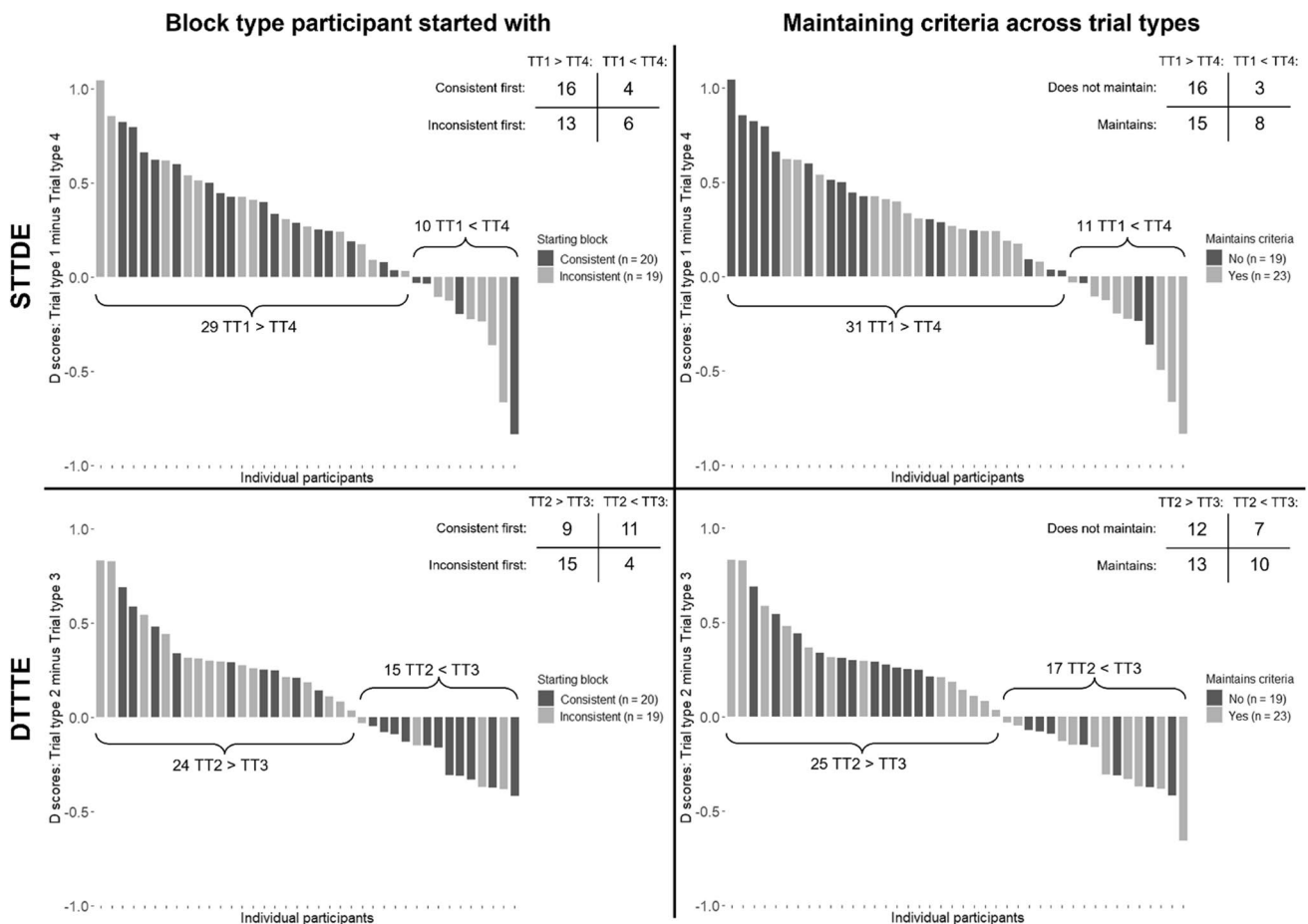


Fig. 3 Finn Graphs Showing Trial-type Differences per Participant According to Starting Block and Maintaining Trial-Type Level Performance Criteria across Test Blocks. *Note.* This figure shows the difference between trial-types 1 and 4 (STTDE; upper panels), and between trial-types 2 and 3 (DTTTE; lower panels), grouping participants according to the rule they started with (left-hand side panels),

as well as whether or not they maintained both performance criteria at the trial-type level across test blocks in all four trial-types (right-hand side panels). The upper-right corners of each panel present the contingency table that supports the χ^2 Test of Independence of the distribution of frequencies in question

effect ($p=0.003$; one-sample proportions test: $\chi^2_{(1)}=8.60$, $p=0.003$). A Bayesian binomial test with a prior beta distribution with parameters $\alpha=1$ and $\beta=1$ provided $BF_{10}=23.89$ (i.e., strong evidence for H_1), where H_1 : proportion of Happy Superiority Effect \neq proportion of Fear Superiority Effect, across the entire sample.

The bottom-left panel of Fig. 3 shows the DTTTE according to block order. Out of the 20 consistent-first participants, 9 showed larger D_{IRAP} scores for trial-type 2 relative to trial-type 3, and 11 showed D_{IRAP} scores for trial-type 3 larger than for trial-type 2. The contingency table for the DTTTE and block order is presented at the top-right corner of this panel. A chi-squared analysis indicated that the DTTTE effects were significantly associated with block order (Pearson's $\chi^2_{(1)}=4.74$, $p=0.029$). We thus proceeded

to exact binomial tests (and Bayesian binomial tests with prior beta distribution where $\alpha=1$ and $\beta=1$) on each type of block order. For participants who started with the consistent block, the DTTTE effects were equally likely ($p=0.824$; $BF_{10}=0.30$). However, for participants who started with the inconsistent block, these effects were significantly different ($p=0.019$; $BF_{10}=6.76$; $BF_{+0}=13.45$, i.e., strong evidence for H_1 : proportion of trial-type 2 bigger than trial-type 3 > proportion of trial-type 2 smaller than trial-type 3). In other words, among participants who started with the inconsistent block, a significant majority showed higher D_{IRAP} scores for trial-type 2 relative to trial-type 3.

The bottom-right panel of Fig. 3 shows the DTTTE according to maintaining practice criteria in test blocks across the four trial-types. A chi-squared analysis indicated

that the DTTTE effects were not associated with maintaining criteria (Pearson's $\chi^2_{(1)} = 0.19, p = 0.663$). Overall, we observed 25 participants showing larger D scores in trial-type 2 relative to trial-type 3 (i.e., bars above zero), and 17 participants showing the opposite pattern (i.e., bars below zero), yielding a difference in proportions which proved to be nonsignificant (exact binomial test: $p = 0.280$, one-sample proportions test: $\chi^2_{(1)} = 1.17, p = 0.280$; Bayesian binomial test with prior beta distribution where $\alpha = 1$ and $\beta = 1$, $BF_{10} = 0.42$, i.e., no evidence for either hypothesis).

In summary, a significant majority of participants produced a happiness superiority effect (across the entire sample), which was not associated with either block order or maintaining performance criteria at the trial-type level. With respect to the DTTTE, there was no overall significant difference between the number of participants who produced a larger trial-type 2 versus 3, or trial-type 3 versus 2 pattern. No significant association between the performance criteria and DTTTE was obtained. However, block order was significantly associated with the DTTTE. In particular, a participant was significantly more likely to produce a trial-type 2 larger than 3 pattern when starting with an inconsistent block, but unlikely to show differences between these trial-types when starting with a consistent block.

Model-Based Comparisons of DAARRE Effects

The foregoing analyses are dichotomous and thus are not sensitive to the relative sizes of the differences among the trial-types. In an effort to capture these relative magnitude differences, we conducted four 2×2 mixed-effects repeated-measures ANOVAs,⁸ each of which corresponded to one of the frequency-based analyses previously presented in Fig. 3. A graphical representation of these four sets of analyses is presented in Fig. 4. Pairwise comparisons were conducted for the pairs of interest (i.e., between trial-types 1 and 4, and between trial-types 2 and 3), for each level of the two factors of interest: block order (starting with either the consistent or inconsistent block) and performance maintenance

across test blocks (maintains or does not maintain criteria). Two types of pairwise comparisons were examined: Bonferroni-adjusted comparisons and a Bayesian version of paired-samples Wilcoxon signed-rank tests based on a data augmentation algorithm with 5 chains of 10,000 iterations, employing a Cauchy prior distribution with a scale of 0.707.

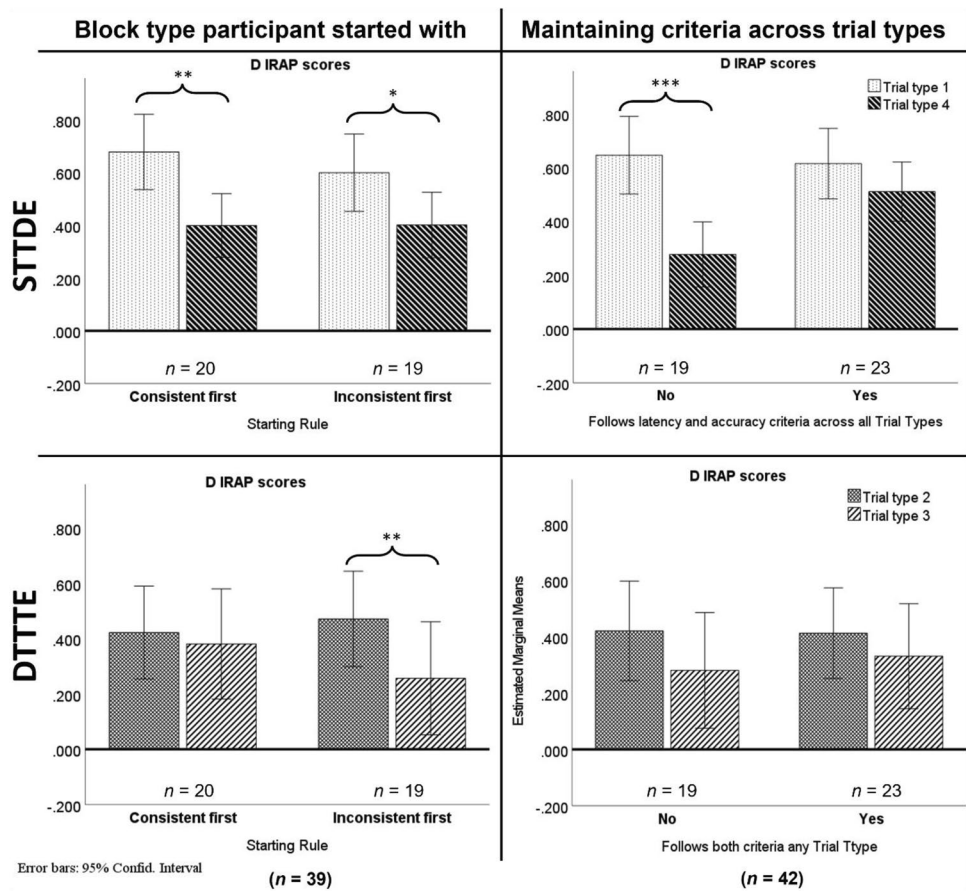
The 2×2 ANOVA conducted on the data presented in the upper-left panel indicated a significant main effect for the STTDE (i.e., trial-types 1×4) ($F_{(1, 37)} = 13.41, p < 0.001, \eta^2_p = 0.266$). That is, across both levels of the between-subjects factor (i.e., starting block), the contrast between trial-types 1 and 4 was significant, with no interaction with block order ($F_{(1, 37)} = 0.32, p = 0.572, \eta^2_p = 0.009$). Bonferroni-adjusted pairwise comparisons estimated the mean difference (ΔM) between trial-types 1 and 4 for each level of block order. Both the consistent-first ($\Delta M = 0.279, F_{(1, 37)} = 9.40, p = 0.004, \eta^2_p = 0.203$) and the inconsistent-first ($\Delta M = 0.198, F_{(1, 37)} = 4.15, p = 0.040, \eta^2_p = 0.109$) participants showed significant mean differences between trial-types 1 and 4. Corresponding to this, the Bayesian analyses indicated strong evidence for trial-type 1 bigger than trial-type 4 when participants started with the consistent block ($BF_{+0} = 18.18$) and moderate evidence when they started with the inconsistent block ($BF_{+0} = 3.44$).

The 2×2 ANOVA conducted on the data presented in the upper-right panel of Fig. 4 indicated a significant main effect for the STTDE (i.e., trial-types 1×4) ($F_{(1, 40)} = 15.51, p < 0.001, \eta^2_p = 0.279$). That is, across the between-subjects factor (i.e., maintaining performance criteria), the contrast between trial-types 1 and 4 was significant. In this case, the STTDE interacted with maintaining performance criteria ($F_{(1, 40)} = 4.86, p = 0.033, \eta^2_p = 0.108$). Bonferroni-adjusted pairwise comparisons estimated that the mean difference (ΔM) between trial-types 1 and 4 was significant for participants who *did not maintain* the performance criteria ($\Delta M = 0.306, F_{(1, 40)} = 17.23, p < 0.001, \eta^2_p = 0.301$). However, this difference was *not significant* for participants who *maintained* both performance criteria across all four trial-types ($\Delta M = 0.104, F_{(1, 40)} = 1.66, p = 0.205, \eta^2_p = 0.040$). The Bayesian analyses corroborated the pairwise comparisons conclusions: participants who did not maintain at least one of the performance criteria at any trial-type across test blocks showed extremely strong evidence for trial-type 1 bigger than trial-type 4 ($BF_{+0} = 354.73$), whereas those who maintained all performance criteria across all trial-types showed no conclusive evidence for either hypothesis ($BF_{+0} = 0.85$).

The 2×2 ANOVA conducted on the data presented in the lower-left panel indicated a significant main effect for the DTTTE (i.e., trial-types 2×3) ($F_{(1, 37)} = 5.89, p = 0.020, \eta^2_p = 0.137$). That it, across both levels of the between-subjects factor (i.e., starting block), the contrast between trial-types 2 and 3 was significant; and this effect

⁸ The distribution of D_{IRAP} scores in each of the four trial-types followed normality ($p > .150$ for all Kolmogorov–Smirnov and Shapiro–Wilk tests). The sphericity assumption could not be tested within each of the four repeated measures ANOVAs (Mauchly's tests' $df = 0$), and thus Greenhouse–Geisser estimators were used whenever possible. Nevertheless, an overreaching repeated-measures ANOVA with the four trial-types' D_{IRAP} scores met the sphericity assumption (Mauchly's $W = .812, \chi^2_{(5)} = 8.26, p = .142$ without between-subjects factors, and $W = .760, \chi^2_{(5)} = 9.27, p = .099$ with both factors). In addition, across these four mixed-effects ANOVAs, Box's tests demonstrated equality of D_{IRAP} scores covariance matrices across groups ($.299 < p < .715$), and Levene's tests demonstrated equality of trial-types' error variances based on means across groups ($.050 < p < .984$).

Fig. 4 D_{IRAP} Scores Means and 95% Confidence Intervals of Trial-Types per Block Order and Criteria Maintenance. *Notes.* STTDE: Single Trial-Type Dominance Effect (TT1 vs. TT4). DTTTE: Dissonant-Target Trial-Type Effect (TT2 vs. TT3). Asterisks refer only to Bonferroni-adjusted pairwise comparisons between trial-types (i.e., the within-subjects factor) at each level of the between-subjects factor, in each 2×2 mixed-effects ANOVA. * $p < .05$. ** $p < .01$. *** $p < .001$



did not interact with block order ($F_{(1, 37)} = 2.70$, $p = 0.109$, $\eta^2_p = 0.068$). However, the categorical analyses indicated an association between DTTTE and block order, so we proceeded with the pairwise comparisons. Bonferroni-adjusted pairwise comparisons estimated the mean difference between trial-types 2 and 3 for each level of block order. The difference was not significant for participants who started with the consistent block ($\Delta M = 0.041$, $F_{(1, 37)} = 0.31$, $p = 0.579$, $\eta^2_p = 0.008$). However, the inconsistent-first participants showed significantly larger D_{IRAP} scores in trial-type 2 relative to trial-type 3 ($\Delta M = 0.214$, $F_{(1, 37)} = 8.08$, $p = 0.007$, $\eta^2_p = 0.179$). The Bayesian analyses again corroborated the pairwise results: for participants who started with the consistent block, there was no evidence for either hypothesis ($BF_{+0} = 0.40$), but for those who started with the inconsistent block, there was strong evidence for trial-type 2 bigger than trial-type 3 ($BF_{+0} = 12.28$).

The 2×2 ANOVA conducted on the data presented in the lower-right panel indicated a significant main effect for the DTTTE (i.e., trial-types 2×3) ($F_{(1, 40)} = 4.12$, $p = 0.049$, $\eta^2_p = 0.093$). That is, across both levels of the between-subjects factor (i.e., maintaining criteria), the contrast between trial-types 2 and 3 was significant, with

no significant interaction with this factor ($F_{(1, 40)} = 0.28$, $p = 0.597$, $\eta^2_p = 0.007$). Bonferroni-adjusted pairwise comparisons estimated the mean difference between trial-types 2 and 3 for each level of maintaining criteria. In neither case were significant differences observed (did not maintain criteria: $\Delta M = 0.140$, $F_{(1, 40)} = 3.00$, $p = 0.091$, $\eta^2_p = 0.070$; maintained the criteria: $\Delta M = 0.082$, $F_{(1, 40)} = 1.24$, $p = 0.272$, $\eta^2_p = 0.030$). The Bayesian analyses indicated weak evidence for trial-type 2 bigger than trial-type 3 for participants who did not maintain the performance criteria ($BF_{+0} = 2.14$), and no conclusive evidence for either hypothesis when participants maintained the criteria ($BF_{+0} = 0.55$).

In summary, the analyses indicated that the difference between trial-types 1 and 4 (STTDE) was affected by maintaining the performance criteria. In particular, those participants who failed to maintain the criteria produced a significantly larger D_{IRAP} score on trial-type 1 relative to trial-type 4; the difference for those who maintained the criteria was nonsignificant. For the difference between trial-types 2 and 3 (DTTTE), the block order affected the relative size of the trial-type D_{IRAP} scores. In particular, only the group who started with the inconsistent block produced a significant difference

between trial-types 2 and 3, which appears to have driven the main effect.

Discussion

In the current study, 42 participants completed an IRAP with happy and fearful faces and words. Participants were divided in terms of whether they commenced with a history-consistent or a history-inconsistent block of trials. In addition, although participants were required to achieve the latency and accuracy performance criteria at the block level, participants were grouped according to whether or not they maintained the criteria at the trial-type level (over test blocks). Consistent with previous findings and the DAARRE model, there was evidence of both a STTDE and a DTTTE. However, both of these effects were moderated by other variables. In particular, the difference between trial-types 1 and 4 was strongly significant for the group of participants who failed to maintain the criteria at the trial-type level, but it was not significant only for the group of participants who managed to maintain such performance across all trial-types. Furthermore, the difference between trial-types 2 and 3 was only significant for the group of participants who commenced the IRAP with a history-inconsistent block. Although the findings were broadly consistent with the DAARRE model, the impact of block order and maintaining criteria at the trial-type level suggests that the model needs to be extended to accommodate these effects.

Let us first consider the impact of maintaining the criteria at the trial-type level on the STTDE. There was limited evidence for the STTDE for the group who maintained the criteria at the trial-type level, but a clear STTDE for the group who did not. As noted previously, the majority of IRAP studies have focused on performance criteria at the block level alone, and thus the current finding could be important in terms of informing future IRAP research. At this point, it seems important to consider a possible explanation for the impact of maintaining trial-type level criteria. As noted earlier, RFT distinguishes between C_{rel} and C_{func} properties of stimuli and the DAARRE model, and in particular focusses on this distinction in the elements presented within an IRAP. The C_{rel} properties refer to the “semantic” relationships between the label and target stimuli, whereas the C_{func} properties refer to the functional properties of the stimuli, such as valence or attentional effects.

If a participant’s performance on an IRAP was controlled largely by the C_{rel} properties, then the IRAP would simply involve categorizing the stimuli according to their semantic relations; indeed, all participants successfully did so when rating the stimuli without performance constraints

(Table 1). Assuming that a participant has an appropriate history to complete such categorizing, each trial-type might be considered broadly equal in difficulty. In contrast, if the C_{func} properties of the elements in an IRAP are involved, they may “interfere” with the control by the C_{rel} properties and thus generate the observed differences between trial-types 1 and 4. If this view is correct, then maintaining the performance criteria at the trial-type level may be seen as indicating dominant control by the C_{rel} properties of the IRAP. Or, more informally, participants who maintained the criteria were simply categorizing happy faces with happy words, and fearful faces with fearful words, and this responding did not seem to be heavily influenced by the valence or emotional functions of stimuli, thereby producing roughly equal IRAP effects in trial-types 1 and 4 (i.e., they were more under C_{rel} than C_{func} control). In contrast, failing to maintain the criteria at the trial-type level may indicate that the C_{func} properties interfered with C_{rel} control. In other words, participants who failed to maintain one or more performance criteria may have found it easier to respond positively (i.e., “true”) to two positive stimuli (maximal coherence) than to two negative stimuli (reduced coherence), thereby producing differential trial-type effects. This may be explained by the *dynamic* influence of the two types of contextual control that can take place differently for different people. Similar interpretations have been applied to explain clinical processes in the context of human psychological suffering (for a detailed discussion, see Harte et al., 2022).

Let us now consider the impact of block order on the DTTTE. Although D_{IRAP} scores for trial-type 2 were overall higher than for trial-type 3, this difference was significant only for participants who started with the inconsistent block (i.e., the main effect was significant, but the only paired comparison that showed significance was for inconsistent-first participants). It is difficult to explain this difference based on C_{func} coherence alone, because the number of stimuli bearing positive and negative C_{func} properties is the same across these two trial-types. They both encompass one positive stimulus (as label in trial-type 2 and as target in trial-type 3) and one negative stimulus (as label in trial-type 3 and as target in trial-type 2). Also note that the C_{func} for both response options is negative in history-consistent blocks and positive in history-inconsistent blocks. In an attempt to explain the significant difference between trial-types 2 and 3 for participants who started with the inconsistent block, we will first draw on Kavanagh et al.’s (2019) explanation, which focused on the spatial contiguity between the targets and the correct response options in each trial-type.

Kavanagh et al. (2019) noted that, in history-consistent blocks, for trial-type 2, the correct response option (e.g., “false”) bears a negative C_{func} , which is *coherent* with the *negative target* (e.g., “fearful”); but for trial-type 3, it is

incoherent with the *positive target* (e.g., “cheerful”). As such, responding correctly to trial-type 3 relative to 2 may be more difficult in the history-consistent block. In contrast, in history-inconsistent blocks, for trial-type 2, the correct response option (e.g., “true”) bears a positive C_{func} which is *incoherent* with the *negative target*; but for trial-type 3, it is *coherent* with the *positive target*. Hence, responding correctly to trial-type 2 may be more difficult relative to trial-type 3 during the history-inconsistent block. Overall, therefore, responding correctly to trial-type 2 may be easier during consistent blocks and more difficult during inconsistent blocks, but the opposite is the case for trial-type 3 (i.e., more difficult during consistent blocks, but easier during inconsistent-blocks). If this analysis is correct, it would explain why the D_{IRAP} score for trial-type 3 is less than the score for trial-type 2. That is, for trial-type 2, the functional overlap between target and correct response option is coherent during consistent blocks and incoherent during inconsistent blocks (thereby increasing the D_{IRAP} score); but, for trial-type 3, there is incoherence between target and correct response option during consistent blocks, and coherence during inconsistent blocks (thereby decreasing the D_{IRAP} score). The current finding indicates that this differential trial-type effect (the DTTTE) was only significant⁹ when participants commenced the IRAP with an inconsistent block, an effect that was not explored (and therefore not reported) by Kavanagh et al.

How might we explain this effect for block order? When participants commenced the IRAP with a history-consistent block, the overall context of the procedure, in terms of the C_{rel} control, would be coherent with the participants’ previous verbal history. Their first contact with the procedure may therefore establish relatively strong control by both the label and target stimuli in terms of C_{rel} properties. In contrast, for those participants who commenced with the history-inconsistent block, responding coherently with the C_{rel} properties would have been punished and therefore other sources of coherence may have affected their performance (assuming that, in general, participants had an extensive history of responding coherently in their natural environment). One such source of coherence would have been the overlap in the C_{func} properties of the spatially contiguous target and correct response option for trial-type 3 (in a history-inconsistent block). Given the current findings, it seems that this initial source of coherence in the first block of the IRAP continued to affect performance throughout the procedure. In other words, participants who started with the history-inconsistent

block may have developed a type of “nearest C_{func} coherence” response bias for trial-type 3, which only favored performance in history-inconsistent blocks. This bias would be punished during subsequent history-consistent blocks for trial-type 3, which may explain the reduction in the D_{IRAP} scores for trial-type 3 relative to trial-type 2. We are assuming here that the “nearest C_{func} coherence” response bias resurged during all subsequent history-inconsistent¹⁰ blocks for participants who started with this type of block.

Of course, this interpretation is post hoc and speculative, but it could be tested in subsequent research. For example, it would be interesting to examine the impact of specific instructions on such effects. In the current study, participants were required to learn to relate the stimuli through trial-and-error, because no specific instruction on how to relate the stimuli was provided. Perhaps providing detailed instructions explaining that the first IRAP block would require responding in a manner that was incorrect (i.e., incoherent with prior verbal history) might undermine the order effect observed here. That is, such an instruction would render a history-inconsistent block coherent with the prior instructions to respond incoherently, thereby attenuating the “nearest C_{func} coherence” bias (see Finn et al., 2016).

One might assume that using words as targets and response options may have contributed to the coherence between these stimuli. However, Kavanagh et al. (2019) observed a similar DTTTE to the one reported here, even though the IRAP configuration employed by Kavanagh et al. used pictures as targets (and words as response options). As noted previously, Kavanagh et al. did not analyze their data in terms of block order, but the overall pattern was similar (trial-type 2 larger than trial-type 3). Furthermore, it is also worth noting Kavanagh et al. employed relatively neutral stimuli (i.e., pens and nonemotional faces), and thus the differential trial-type effects they obtained might have been based more on orienting (e.g., attentional) rather than evoking (e.g., valence) functions (see Finn et al., 2018). As such, the pattern observed here appears to be generally consistent with previous studies (cf. Hussey & Drake, 2020).

Only two previously published studies, to our knowledge, have focused on both the STTDE and the DTTTE (Kavanagh et al., 2019; Schmidt et al., 2021). The STTDE, in particular, has been consistently replicated across different domains, such as color–color over shape–shape (Finn et al., 2018), face–face over pen–pen (Kavanagh et al., 2019), happy-symbol–positive-word over fearful-symbol–negative-word (Perez et al., 2019);

⁹ This was the only *pairwise comparison* that showed a statistically significant difference, which alone seems to have driven the main effect for DTTTE observed for both ANOVAs: one with block order (fixed) and the other with performance criteria maintenance as the between-subjects factor.

¹⁰ We reiterate that by “history-inconsistent” or “history-consistent” we are referring to coherence with participants’ likely verbal history *prior to the experiment*. This should not be confused with the history established during the IRAP session itself.

happy-face–happy-symbol over negative-emotion–negative-symbol (Bortoloti et al., 2019; see also Bortoloti et al., 2020), happy-face–preferred-icon over angry-face–indifferent-icon (Pinto et al., 2020); and opposing patterns of in-group over out-group positivity bias (Hughes et al., 2017). However, the current study was only the second to predict a priori both the STTDE and the DTTTE (the first being Schmidt et al., 2021), and these effects were indeed observed. Furthermore, the current study is the first to analyze the data at the individual level and, it is interesting to note, not all the participants showed these two effects. In this context, it is important to understand that the type of C_{func} properties involved in a participant's performance could, in principle, change the pattern of responding on the IRAP. For example, if the valence (i.e., evoking properties) of the faces tended to control responding, then a happiness superiority effect would be more likely; in contrast, if the salience (i.e., orienting functions) of specific features of the faces (e.g., wide-open eyes in fearful faces) tended to control responding, then a fear superiority effect would be more likely. In other words, a fear superiority effect for a particular individual should not be interpreted as “preferring” or “liking” fearful more than happy faces (see Hughes et al., 2018). According to the DAARRE model, it is the specific C_{func} properties of the stimuli (orienting versus evoking) that play an important role in determining the overall response patterns observed on the IRAP. These stimulus properties may well differ between participants based on their idiosyncratic preexperimental histories. However, consistent with previous research (Bortoloti et al., 2019; Craig et al., 2014; Lepänen & Hietanen, 2004), a happiness superiority effect, *in general*, was observed in the current study, although this was moderated by maintaining the performance criteria at the trial-type level, which is a novel finding in the IRAP literature.

In line with Finn et al. (2019), the present study reported individual-participant data, thus supporting their call for such analyses in IRAP research (p. 434). The potential advantage in doing so may be exemplified when considering the impact of maintaining criteria on the STTDE. In particular, for individual participants (Fig. 3), the majority of those who did not maintain the criteria showed a larger D_{IRAP} score for trial-type 1 relative to 4. Furthermore, only 3 of these 19 participants showed the opposite effect (trial-type 4 larger than 1). Although the frequency-based analyses did not indicate an association between maintaining criteria and the STTDE, the variance-based analyses did support a significant interaction. One likely reason is because the frequency-based analyses did not capture the relative magnitudes of the D_{IRAP} scores, thus failing to reflect the impact of relatively large effects such as the five participants with the highest trial-type-1-minus-4 differences, who were all

in the criteria-failing group. In effect, when participants do not maintain the performance criteria, their differential trial-type effects may tend to be larger (in the predicted direction) than when they do maintain the criteria. Analyzing IRAP data both in terms of frequency and variance allows us to see such effects.

The current findings are broadly consistent with previous research that has drawn on the DAARRE model in highlighting the combined role of C_{rel} and C_{func} properties in determining differential trial-type effects on the IRAP. This increasingly well-established finding seems to warrant a change in focus within RFT (Harte & Barnes-Holmes, 2024). In traditional RFT research, the relationship between entailed relations and the transformation of functions has frequently focused on the extent to which establishing specific entailed relations allows for specific changes in the functional properties of the stimuli participating within those relations. For example, in many RFT studies a relational network may be established in which three stimuli are related to each other, for instance when A is trained as more than B, and B is trained as more than C. A specific functional property may then be established for one of the stimuli, such as an aversive (e.g., electric shock) function for C. The A stimulus may subsequently evoke an aversive reaction stronger than the reaction that was observed for the C stimulus, because A is derived as more than C (see, for example, Dougher et al., 2007). This approach focuses on the impact of the stimulus relations (C_{rel}) on the functional properties (C_{func}) of the stimuli in a given network. In the current and related research, however, there is a greater focus on the impact of the functional properties of the stimuli on relating. In particular, the coherence among the C_{func} properties of the stimuli within the network seems to interact with their C_{rel} properties in a way that explains the responding patterns observed on the IRAP. On balance, recognizing the impact of C_{func} properties on C_{rel} properties does not mean that the “traditional” approach (i.e., from relation to function) was incorrect; both approaches can supplement each other (Harte & Barnes-Holmes, 2024).

We should also acknowledge that previous research has reported effects consistent with a function-to-relation approach, in which functional classes were shown to generate equivalence relations (e.g. Sidman et al., 1989; Smeets et al., 1997). Nevertheless, most of this research focused on demonstrating that functional classes may generate equivalence relations, but they did not indicate that the functional properties of the stimuli could affect the accuracy and latency of relational responding itself. The DAARRE model addresses the relative coherence among the C_{func} and C_{rel} properties of the stimuli within a network on the IRAP, and this supports a more thorough analysis of the dynamic interplay between the relational and the functional properties of stimuli (Harte & Barnes-Holmes, 2024).

Adopting a relatively balanced focus between C_{rel} and C_{func} properties, as we have suggested here, may serve to highlight potential commonalities with research on language in other domains. For example, semantics, as a subset of linguistics, has traditionally tended to focus on explanations of the formal structure of language, even when addressing the referencing relationships between linguistic expressions and concepts or categories (e.g., cf. Glynn, 2015; Rakhilina et al., 2022; Rosch, 1973). Broadly speaking, in RFT terms, this focus would be on the C_{rel} properties of a semantic network. However, when language is viewed as an evolutionary feature of the human species (Pinker & Bloom, 1990), this calls for a focus on the psychological learning history of language across human development (i.e., how the specific properties of language are selected by the environment, both across and within generations), and thus on the pragmatics of language learning. The present findings emphasize the importance of the transformation of historically established psychological functions (C_{func}) of stimuli in arbitrarily applicable relational responding (under C_{rel} control), as studied by RFT, in conceptualizing meaning and categorization, for it shows that the former affects upon the latter.

In using merely *descriptive* adjectives (e.g., happy, fearful) with facial expressions assumed to be generally described as such (cf. Barrett et al., 2019), the task itself simply required participants to categorize attributes into their domains based on literal semantic equivalences. These mappings referred to emotional concepts that are presumed not socially meaningful (e.g., in terms of race or gender), but the differences in the bare emotional value of stimuli influenced the emergence of differential trial-type effects. These effects may possibly be applied in research on socially sensitive issues, such as in-group and out-group biases in the context of social identities.

For instance, using *evaluative* adjectives (e.g., honest, nice, friendly; nasty, aggressive, hostile), Hughes et al. (2017) asked Northern Irish participants from either community (Catholic or Protestant) to respond to four trial-types: Catholic–good, Catholic–bad, Protestant–good, Protestant–bad, using “True” and “False” as response options. On average, trial-types with a positive evaluative adjective were all significantly different from zero in the direction where the positive adjective had to be responded to as “True,” for both groups (i.e., both groups assessed their in-group and out-group as “good”). However, Catholic participants showed D_{IRAP} scores in the Catholic–good trial-type that were higher than the Protestant–good trial-type, whereas Protestant participants showed D_{IRAP} scores in the Protestant–good trial-type higher than the Catholic–good trial-type. Hughes et al. concluded that the IRAP revealed in-group favoritism in the absence of out-group derogation. One direction for future research would be to replicate the current study using socially loaded stimuli like those employed by Hughes et al. Would the effects reported by those researchers be moderated by the two variables found to be relevant in the current study?

Summary and Concluding Remarks

We henceforth abridge our analyses and key findings. The significant main effect for STTDE did not interact with block order, as a significant “happiness superiority effect” (trial-type 1 bigger than trial-type 4) was significant regardless of block order. STTDE also had a significant main effect when performance maintenance was concerned, but in that case it interacted significantly with whether or not participants maintained the performance criteria at the trial-type level. In particular, there was strong evidence for higher D_{IRAP} scores in trial-type 1 relative to 4 for participants who failed to maintain at least one of the performance criteria in at least one trial-type, but there was no evidence for such difference for participants who maintained both performance criteria across all trial-types. This interaction was not reflected in the chi-square test because this analysis considered STTDE in a dichotomous way (i.e., $1 > 4$ or $4 > 1$), and therefore it could not capture the magnitudes of the scores’ differences. A visual inspection of the Finn graphs in Fig. 3 (upper-right panel) illustrates that point.

A significant main effect for DTTTE (trial-type 2 bigger than trial-type 3) did not interact with block order, but the categorical analyses captured a significant association whereby DTTTE effects depended on block order. This was due to a significantly higher proportion of trial-type 2 bigger than trial-type 3 for the participants who started with the inconsistent block, contrasted with an absence of difference for those who started with the consistent block. Comparisons of mean D_{IRAP} scores corroborated this conclusion: for participants who started with the inconsistent block, there was strong evidence for trial-type 2 bigger than trial-type 3, but not for those who started with the consistent block. Finally, DTTTE had a significant main effect regardless of whether participants did or did not maintain the performance criteria at the trial-type level, with no interaction and no specific difference for those two groups.

These results highlight the importance of studying the influence of functional properties of stimuli. In other words, it emphasizes the relevance of including a function-to-relation approach by using the DAARRE framework, and indeed RFT, to interpret behavioral data, in addition to the traditional relation-to-function approach. We have relied on a “nearest C_{func} coherence” response bias explanation for the DTTTE, and we have considered the extent to which maintaining performance criteria could reduce the STTDE. We have considered how future research on socially sensitive, evaluative responding may be informed by our findings. We conclude that the DAARRE model, expanded to accommodate block order and trial-type performance as relevant factors, is a promising framework for analyzing the dynamics of arbitrarily applicable relational responding in the environment of the IRAP.

Notes. There are no IRAP test-blocks data for seven participants who did not attain practice-blocks performance criteria and chose not to restart the session, but we kept their participant numbers because we hold their practice-blocks data, as well as DASS and rating scale responses. We kept no records of the participant who declined participation (no participant number). The three participants who chose to take a second IRAP session after using all practice blocks in the first IRAP session kept both participant numbers.

^a Participant did not attain practice-blocks performance criteria in order to move to test-blocks. ^b Participant restarted IRAP after first complete round of practice-blocks. ^c Average time in milliseconds spent until correct response.

Appendix B

The focus of the present article is on the experimental results reported. To analyze the data, we used a first version of a Python code we developed to identify IRAP performances at the trial-type level (the GO-IRAP software provides outputs with block-level performances). The original version of this code has been substantially expanded (and can be accessed at <https://policing-irap.rd.psc.br/>) and its full explanation warrants a separate article. The source code for the original version can be found at https://github.com/rodva/TrialType_Criterion_Police.

The expanded version consists of an HTML code with chunk of Python code in it. It can also be found at the foregoing GitHub link. This version identifies the trials that fail time or accuracy criteria according to the user's input, which is an interactive option that the local source code does not provide. This means that, in the online version, the user can customize the time and accuracy criteria to find the trials that violate them. A full explanation of those codes shall be the topic of a separate text, but the reader is invited to contact the authors for queries regarding these algorithms.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40732-024-00604-1>.

Author Contribution Rodrigo de Almeida: Software, formal analysis, investigation, data curation, writing—original draft, visualization, funding acquisition. Dermot Barnes-Holmes: Conceptualization, methodology, resources, writing—original draft, supervision, project administration, Funding acquisition. Julian C. Leslie: Resources, writing—reviewing and editing, supervision, project administration.

Funding The only source of funding for this research is Ulster University Vice-Chancellor's Research Studentship (Department for the Economy) awarded to the first author to fund his PhD and the research program it involves.

Data Availability The data presented in this article is available at the Open Science Framework repository and can be found at <https://osf.io/nj3hk/> or using the DOI <https://doi.org/10.17605/OSF.IO/NJ3HK>.

Materials and Code Availability The code developed for this study is available at https://github.com/rodva/TrialType_Criterion_Police and can be used at <https://policing-irap.rd.psc.br/>. The GO-IRAP installer may be found at <https://balc-i.net/software-and-materials/>.

Declarations

Ethics Approval This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the School of Psychology staff and Postgraduate Research Ethics Filter Committee (FCPSY-21-035-B).

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent to Publish No participant's image is presented in this article. The pictures on Fig. 1 were obtained from the FACES database at <https://faces.mpdl.mpg.de/imeji/> (Ebner et al., 2010). All participants agreed with the following statements, among others: "I understand that my analyzed data can be anonymously published and presented in scientific media" and "I understand that the data can be shared with other researchers using open science repositories."

Competing Interest On behalf of all authors, the corresponding author states that there is no conflict of interest. There are no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Barnes, D., Lawlor, H., Smeets, P. M., & Roche, B. (1996). Stimulus equivalence and academic self-concept among mildly mentally handicapped and nonhandicapped children. *The Psychological Record*, 46, 87–107. <https://doi.org/10.1007/BF03395165>
- Barnes-Holmes, D., & Harte, C. (2022). The IRAP as a measure of implicit cognition: A case of Frankenstein's monster. *Perspectives on Behavior Science*, 45(3), 559–578. <https://doi.org/10.1007/s40614-022-00352-z>
- Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010a). A sketch of the implicit relational assessment procedure (IRAP) and the relational elaboration and coherence (REC) model. *The Psychological Record*, 60(3), 527–542. <https://doi.org/10.1007/bf03395726>
- Barnes-Holmes, D., Finn, M., McEntegart, C., & Barnes-Holmes, Y. (2017). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *Perspectives on Behavior Science*, 41(1), 155–173. <https://doi.org/10.1007/s40614-017-0124-7>
- Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The implicit relational assessment procedure (IRAP) as a

- response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record*, 58, 497–516. <https://doi.org/10.1007/BF03395634>
- Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010b). The implicit relational assessment procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-Black stereotyping among white Irish individuals. *The Psychological Record*, 60(1), 57–79. <https://doi.org/10.1007/bf03395694>
- Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010c). Using the implicit association test and the implicit relational assessment procedure to measure attitudes toward meat and vegetables in vegetarians and meat-eaters. *The Psychological Record*, 60(2), 287–305. <https://doi.org/10.1007/bf03395708>
- Barnes-Holmes, D., Staunton, C., Whelan, R., Barnes-Holmes, Y., Commins, S., Walsh, D., Stewart, I., Smeets, P. M., & Dymond, S. (2005). Derived stimulus relations, semantic priming, and event-related potentials: Testing a behavioral theory of semantic networks. *Journal of the Experimental Analysis of Behavior*, 84(3), 417–433. <https://doi.org/10.1901/jeab.2005.78-04>
- Barnes-Holmes, D., Waldron, D., Barnes-Holmes, Y., & Stewart, I. (2009). Testing the validity of the implicit relational assessment procedure and the implicit association test: Measuring attitudes toward Dublin and country life in Ireland. *The Psychological Record*, 59, 389–406. <https://doi.org/10.1007/BF03395671>
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1). <https://doi.org/10.1177/1529100619832930>
- Bortoloti, R., de Almeida, R. V., de Almeida, J. H., & de Rose, J. C. (2019). Emotional faces in symbolic relations: A happiness superiority effect involving the equivalence paradigm. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.00954>
- Bortoloti, R., de Almeida, R. V., de Almeida, J. H., & de Rose, J. C. (2020). A commentary on the dynamics of arbitrarily applicable relational responding involving positive valenced stimuli and its implications for the IRAP research. *The Psychological Record*, 71(3), 481–486. <https://doi.org/10.1007/s40732-020-00413-2>
- Bortoloti, R., Pimentel, N., & de Rose, J. C. (2014). Electrophysiological investigation of the functional overlap between semantic and equivalence relations. *Psychology & Neuroscience*, 7(2), 183–191. <https://doi.org/10.3922/j.psns.2014.015>
- Craig, B. M., Becker, S. I., & Lipp, O. V. (2014). Different faces in the crowd: A happiness superiority effect for schematic faces in heterogeneous backgrounds. *Emotion*, 14(4), 794–803. <https://doi.org/10.1037/a0036043>
- Devany, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language-disabled children. *Journal of the Experimental Analysis of Behavior*, 46(3), 243–257. <https://doi.org/10.1901/jeab.1986.46-243>
- Dias, G. C. B., Silveira, M. V., Bortoloti, R., & Huziwar, E. M. (2020). Electrophysiological analysis of stimulus variables in equivalence relations. *Journal of the Experimental Analysis of Behavior*, 115(1), 296–308. <https://doi.org/10.1002/jeab.664>
- Dixon, M. R., Rehfeldt, R. A., Zlomke, K. M., & Robinson, A. (2006). Exploring the development and dismantling of equivalence classes involving terrorist stimuli. *The Psychological Record*, 56, 83–103. <https://doi.org/10.1007/BF03395539>
- Dougher, M. J. (2020). Murray Sidman's contributions to clinical behavior analysis. *Journal of the Experimental Analysis of Behavior*, 115(1), 36–43. <https://doi.org/10.1002/jeab.644>
- Dougher, M. J., Hamilton, D. A., Fink, B. C., & Harrington, J. (2007). Transformation of the discriminative and eliciting functions of generalized relational stimuli. *Journal of the Experimental Analysis of Behavior*, 88(2), 179–197. <https://doi.org/10.1901/jeab.2007.45-05>
- Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1), 351–362. <https://doi.org/10.3758/BRM.42.1.351>
- Finn, M., Barnes-Holmes, D., & McEntegart, 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, 11–25. <https://doi.org/10.1007/s40732-017-0262-z>
- Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016). Exploring the behavioral dynamics of the implicit relational assessment procedure: The impact of three types of introductory rules. *The Psychological Record*, 66(2), 309–321. <https://doi.org/10.1007/s40732-016-0173-4>
- Finn, M., Barnes-Holmes, D., McEntegart, C., & Kavanagh, D. (2019). Predicting and influencing the single-trial-type-dominance-effect: The first study. *The Psychological Record*, 69(3), 425–435. <https://doi.org/10.1007/s40732-019-00347-4>
- Glynn, D. (2015). Semasiology and onomasiology: Empirical questions between meaning, naming and context. In J. Daems, E. Zenner, K. Heylen, D. Speelman, & H. Cuyckens (Eds.), *Change of paradigms—new paradoxes: Recontextualizing language and linguistics* (pp. 47–79). Mouton de Gruyter. <https://doi.org/10.1515/9783110435597-004>
- Gomes, C. T., Perez, W. F., de Almeida, J. H., Ribeiro, A., de Rose, J. C., & Barnes-Holmes, D. (2019). Assessing a derived transformation of functions using the implicit relational assessment procedure under three motivative conditions. *The Psychological Record*, 69(4), 487–497. <https://doi.org/10.1007/s40732-019-00353-6>
- Harte, C., & Barnes-Holmes, D. (2024). Recent developments in RFT encourage interbehavioral field-based views of human language and cognition: A preliminary analysis. *Perspectives on Behavior Science*. <https://doi.org/10.1007/s40614-024-00407-3>
- Harte, C., Barnes-Holmes, D., de Rose, J. C., Perez, W. F., & de Almeida, J. H. (2022). Grappling with the complexity of behavioral processes in human psychological suffering: Some potential insights from relational frame theory. *Perspectives on Behavior Science*, 46, 237–259. <https://doi.org/10.1007/s40614-022-00363-w>
- Haydu, V. B., Aquino, C. T. de, Gaça, L. B., & Tomanari, G. Y. (2019). Funções de Estímulos Pré-experimentais na Formação de Classes de Equivalência [Pre-experimental Stimuli Functions in Equivalence Class Formation]. *Psicologia: Teoria e Pesquisa*, 35, e3524. <https://doi.org/10.1590/0102.3772e3524>
- Haydu, V. B., Camargo, J., & Bayer, H. (2015). Effects of preexperimental history on the formation of stimulus equivalence classes: A study with supporters of Brazilian soccer clubs. *Psychology & Neuroscience*, 8(3), 385–396. <https://doi.org/10.1037/h0101276>
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Kluwer Academic/Plenum.
- Hughes, S., Barnes-Holmes, D., & Smyth, S. (2017). Implicit cross-community biases revisited: Evidence for ingroup favoritism in the absence of outgroup derogation in Northern Ireland. *The Psychological Record*, 67(1), 97–107. <https://doi.org/10.1007/s40732-016-0210-3>
- Hughes, S., Barnes-Holmes, D., Van Dessel, P., de Almeida, J. H., Stewart, I., & De Houwer, J. (2018). On the symbolic generalization of likes and dislikes. *Journal of Experimental Social Psychology*, 79, 365–377. <https://doi.org/10.1016/j.jesp.2018.09.002>
- Hussey, I., & Drake, C. E. (2020). The implicit relational assessment procedure is not very sensitive to the attitudes and learning

- histories it is used to assess [Unreviewed manuscript]. PsyArXiv. <https://doi.org/10.31234/osf.io/sp6jx>
- Hussey, I., Thompson, M., McEnteggart, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). Interpreting and inverting with less cursing: A guide to interpreting IRAP data. *Journal of Contextual Behavioral Science*, 4(3), 157–162. <https://doi.org/10.1016/j.jcbs.2015.05.001>
- 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(2), 163–176. <https://doi.org/10.1007/s40732-018-0276-1>
- Kavanagh, D., Matthyssen, N., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Vastano, D. (2019). Exploring the use of pictures of self and other in the IRAP: Reflecting upon the emergence of differential trial-types effects. *International Journal of Psychology & Psychological Therapy*, 19(3), 323–336.
- Kelly, A., & Barnes-Holmes, D. (2013). Implicit attitudes towards children with autism versus normally developing children as predictors of professional burnout and psychopathology. *Research in Developmental Disabilities*, 34(1), 17–28. <https://doi.org/10.1016/j.ridd.2012.07.018>
- Leppänen, J. M., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research Psychologische Forschung*, 69, 22–29. <https://doi.org/10.1007/s00426-003-0157-2>
- Leslie, J. C., Tierney, K. J., Robinson, C. P., Keenan, M., Watt, A., & Barnes, D. (1993). Differences between clinically anxious and non-anxious subjects in a stimulus equivalence training task involving threat words. *The Psychological Record*, 43(1), 153–161.
- Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behaviour Research & Therapy*, 33, 335–343. [https://doi.org/10.1016/0005-7967\(94\)00075-U](https://doi.org/10.1016/0005-7967(94)00075-U)
- McKenna, I. M., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2007). Testing the fake-ability of the implicit relational assessment procedure (IRAP): The first study. *International Journal of Psychology & Psychological Therapy*, 7(2), 253–268.
- Perez, W. F., de Almeida, J. H., de Rose, J. C., Dorigon, A. H., de Vasconcellos, E. L., da Silva, M. A., Lima, N. D. P., de Almeida, R. B. M., Montan, R. N. M., & Barnes-Holmes, D. (2019). Implicit and explicit measures of transformation of function from facial expressions of fear and of happiness via equivalence relations. *The Psychological Record*, 69(1), 13–24. <https://doi.org/10.1007/s40732-018-0304-1>
- Pidgeon, A., McEnteggart, C., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2021). Four self-related IRAPs: Analyzing and interpreting effects in light of the DAARRE model. *The Psychological Record*, 71(3), 397–409. <https://doi.org/10.1007/s40732-020-00428-9>
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral & Brain Sciences*, 13(4), 707–727. <https://doi.org/10.1017/S0140525X00081061>
- Pinto, J. A. R., de Almeida, R. V., & Bortoloti, R. (2020). The stimulus' orienting function may play an important role in IRAP performance: Supportive evidence from an eye-tracking study of brands. *The Psychological Record*, 70(2), 257–266. <https://doi.org/10.1007/s40732-020-00378-2>
- Rabelo, L. Z., Bortoloti, R., & Souza, D. H. (2014). Dolls are for girls and not for boys: Evaluating the appropriateness of the implicit relational assessment procedure for school-age children. *The Psychological Record*, 64(1), 71–77. <https://doi.org/10.1007/s40732-014-0006-2>
- Rakhilina, E., Ryzhova, D., & Badryzlova, Y. (2022). Lexical typology and semantic maps: Perspectives and challenges. *Zeitschrift Für Sprachwissenschaft*, 41(1), 231–262. <https://doi.org/10.1515/zfs-2021-2046>
- Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both? *Journal of Health Psychology*, 15(3), 416–425. <https://doi.org/10.1177/1359105309350232>
- Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). Research Foundation of the State University of New York Academic Press.
- Schmidt, M., de Rose, J. C., & Bortoloti, R. (2021). Relating, orienting and evoking functions in an IRAP study involving emotional pictographs (emojis) used in electronic messages. *Journal of Contextual Behavioral Science*, 21, 80–87. <https://doi.org/10.1016/j.jcbs.2021.06.005>
- Sereno, M., Quigley, J., & Smith, G. S. (2021). A systematic review of the use of the implicit relational assessment procedure (IRAP) to assess implicit biases held by professionals toward client populations. *Research on Social Work Practice*, 32(1), 73–91.
- Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech & Hearing Research*, 14(1), 5–13. <https://doi.org/10.1044/jshr.1401.05>
- Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Authors Cooperative.
- Sidman, M. (2018). What Is interesting about equivalence relations and behavior? *Perspectives on Behavior Science*, 41, 33–43. <https://doi.org/10.1007/s40614-018-0147-8>
- Sidman, M., Wynne, C. K., Maguire, R. W., & Barnes, T. (1989). Functional classes and equivalence relations. *Journal of the Experimental Analysis of Behavior*, 52(3), 261–274. <https://doi.org/10.1901/jeab.1989.52-261>
- Smeets, P. M., Barnes, D., & Roche, B. (1997). Functional equivalence in children: Derived stimulus–response and stimulus–stimulus relations. *Journal of Experimental Child Psychology*, 66(1), 1–17. <https://doi.org/10.1006/jecp.1997.2378>
- Timmins, L., Barnes-Holmes, D., & Cullen, C. (2016). Measuring implicit sexual response biases to nude male and female pictures in androphilic and gynephilic men. *Archives of Sexual Behavior*, 45(4), 829–841. <https://doi.org/10.1007/s10508-016-0725-3>
- Tonneau, F. (2001). Equivalence relations: A critical analysis. *European Journal of Behavior Analysis*, 2(1), 1–33. <https://doi.org/10.1080/15021149.2001.11434165>
- Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record*, 41(1), 33–50. <https://doi.org/10.1007/BF03395092>