Incidental learning of trust from identity-contingent gaze cues: boundaries, extensions and applications.

James Strachan

PhD

University of York

Psychology

September 2016

Abstract

Monitoring the trustworthiness of social interaction partners is a cornerstone of social cognition. However, the mechanics of learning about trust during online interactions as a result of a person's behaviour can be difficult to explore. The current experiments use a gaze cueing paradigm where faces provide either valid (always shift their gaze towards the location of a subsequent target), or invalid cues (always shift their gaze to a different location). Following gaze cueing, participants rate valid faces as more trustworthy than invalid faces. We show that this incidental trust learning is sensitive to the emotional expression of the face, is specific to assessments of trust, occurs outside of conscious awareness, and is driven primarily by a decrease in trust for invalid faces (Chapter 2), perhaps reflecting a cheater detection module. Memory for incidentally learned trust is surprisingly durable, is affected by the familiarity of the cueing faces (Chapter 3), and does not affect memory for the faces' physical features, nor does the trustworthiness of the face generalise to other stimuli (Chapter 4). Furthermore, learning is modulated by top-down knowledge of social group membership − when group identity is made experimentally salient, participants default to a group-level representation as a heuristic for social judgements (Chapter 5), while using naturally occurring group memberships (i.e. race) results in better learning for in-group members than out-group (Chapter 6). Finally, while there is evidence that trust learning is driven by learning about eye-gaze behaviour, this cannot be explained purely by disruptions to visuomotor fluency (Chapter 7), which suggests that this phenomenon is part of an active social monitoring framework that relies on physical changes or behaviours in a face to affect subsequent social judgements.

Contents

List of Tables

Page

List of Figures

Page

Acknowledgements

I would like to thank my supervisor Steve Tipper for his constant support and enthusiastic guidance throughout the last three years. Thanks also go to the other members of my advisory panel, Harriet Over and Nick Barraclough, for their contributions and advice. I would also like to thank the other members of my lab for being there to provide a sympathetic ear or helpful suggestions: Victoria Brattan, Tim Vestner, and Alex Kirkham in particular, who helped with a lot of the technical aspects of getting experiments running and helping me to analyse them.

On a broader note, I would like to thank those people both in York and elsewhere who helped to keep me mostly sane during my PhD. To Beth, Adam, Phil, Debby, Damian, and all of my friends both within and outside the department, who are too numerous to list here but who were always on hand to keep my spirits up, and to my Mum, Dad, my sister Megan, and the rest of my family for their love and support through all these years, I say a huge thank you. Finally, thanks go to the Golden Fleece on Pavement for letting me turn up and set up camp in their beer garden, for keeping me supplied with Copper Dragon through the really difficult times, and particularly to Steph for making me feel like part of the family.

Author Declaration

This thesis is the original work of James Strachan. All purely behavioural data were collected by James Strachan, and electromyographic (EMG) data were collected by James Strachan and Dr. Alexander Kirkham. No part of this thesis has been submitted for a degree or any other qualification at this University or any other institution. The work in this thesis is not published anywhere else unless stated here:

- Experiments 2.1, 2.2, 2.3, 7.1, and 7.2 are published in *Journal of Experimental Psychology: Learning Memory and Cognition* as:

  Strachan, J.W.A., Kirkham, A.J., Manssuer, L.R. & Tipper, S.P. (2016). Incidental Learning of Trust: Examining the Role of Emotion and Visuomotor Fluency. *Journal of Experimental Psychology: Learning Memory and Cognition. 42*(11) 1759-1773.

- The first four experiments in Chapter 3 (Experiments 3.1, 3.2, 3.3, and 3.4) are published in *The Quarterly Journal of Experimental Psychology* as:

  Strachan, J.W.A. & Tipper, S.P. (2016). Examining the durability of incidentally learned trust from gaze cues. *Quarterly Journal of Experimental Psychology.* 1-16.

Chapter 1. Introduction and literature review

Navigating the social world requires a huge amount of processing power. Every new person with whom we interact is quickly evaluated according to a range of social dimensions, such as how friendly they may be, how attractive, how trustworthy, how interested they are in us, and how dangerous they may be if interactions sour. These are important judgements because they can colour every subsequent social decision that we make − whether or not to continue a conversation, or accept a drink, or invest whatever time and resources may be necessary to build and maintain a relationship with that person.

These early decisions are made quickly and automatically. Willis and Todorov (2006) found that people can make decisions about the attractiveness, likeability, trustworthiness, aggressiveness, and competence of a face image with as little as 100ms exposure, and these judgements correlated strongly with judgements made in the absence of time constraints. This indicates that 100ms exposure to a face image is enough to make reliable social decisions about the person.

The underlying properties of these early social decision mechanisms have been well explored. First impressions of faces can be explained using three dimensions: trustworthiness, dominance, and youthful-attractiveness (Sutherland et al., 2013). Decisions about trustworthiness also relate to cues to emotion, to the point where trustworthy faces (faces displaying a physiognomic architecture identified as trustworthy by Todorov, Baron & Oosterhof, 2008) that express smiles are rated as expressing more intense happiness than smiling untrustworthy faces, while the reverse is true for angry expressions (Oosterhof & Todorov, 2009). This suggests that emotion and trustworthiness have a shared perceptual basis, an interpretation supported by the finding that they also share neural mechanisms (Engell, Todorov & Haxby, 2010).

First impressions of faces that load on these dimensions are heavily image-dependent: as images rated as more trustworthy are more likely to include certain configurations of features that might be similar to emotional expressions, this means that ambient (varying naturally in lighting, viewpoint, expression, hairstyle, age, etc.) images of individuals can generate widely varied social decision ratings (Jenkins, White, Van Montfort & Mike Burton, 2011; Todorov & Porter, 2014). In fact, computational analysis of image properties can account for 58% of variance in human raters' impressions of previously unseen ambient images (Vernon, Sutherland, Young & Hartley, 2014).

Given that much of this literature into first impressions focuses on decisions that appear to be highly motivated by image-level properties rather than in vivo social decisions about people, this research is particularly applicable in a world of social media, where selection of profile pictures and avatars to represent us online is a key way of communicating with people who have not met us in person. It also has far reaching and potentially catastrophic implications in the justice system − Wilson and Rule (2015) found that naïve participants' ratings of inmate photographs predicted the likelihood that a convicted murderer would be sentenced to death (as opposed to life imprisonment), and that this relationship was true even of wrongfully convicted people who were later exonerated, indicating that physical appearance can increase the likelihood of a wrongfully severe punishment.

However, there are some important caveats about generalising from this research to other real-world social decisions. When interacting with people in the real world our experiences are much more varied and dynamic than the use of static images allow, and research on social judgements comparing dynamic with static stimuli is scarce, inconclusive, and focuses primarily on judgements of attractiveness over any other social

dimensions (Rubenstein, 2005; Roberts et al., 2009; Kościński, 2013). There is also inconsistent evidence on the validity and accuracy of these judgements (Porter, England, Juodis, Ten Brinke & Wilson, 2008; Kramer & Ward, 2010; Stirrat & Perrett, 2010; Carré & McCormick, 2008; Efferson & Vogt, 2013; Gómez-Valdés et al., 2013).

Research into first impressions has yielded some important insights, but there are other techniques to study social judgements. One way is to use third-party information through short descriptions of a character's behaviour (what one might consider a laboratory analogue of gossip). These behavioural vignettes are typically crafted to elicit particular trait judgements (e.g. "Bob gives a lot of money to charity" could indicate that Bob is generous, while, "Adam likes to make jokes at other people's expense" would indicate that Adam is unlikeable).

The advantage of measuring attitudes elicited by these behavioural vignettes rather than physical facial features is that the same identities can be compared when they embody different traits. This is particularly useful in neuroimaging, where third-party behavioural descriptions can be used to dissociate higher order social information from low-level visual properties, and have helped to show representations of social information in visuotemporal (Verosky, Todorov & Turk-Browne, 2013) and posterior cingulate cortex (Kuzmanovic et al., 2012).

Behavioural descriptions have also been used to examine some behavioural properties of social learning. Falvello, Vinson, Ferrari and Todorov (2015) investigated the set capacity of learning from positive, negative, or neutral behavioural descriptions by showing participants 500 image-description pairs, a subset of which (either 100, 200, 300, or 400 images) were faces and the rest places. Participants were instructed to read the behavioural descriptions and form impressions about the face presented with it, and

after learning they rated each face on a scale of trustworthiness from 1 (untrustworthy) to 9 (trustworthy). They found that people rated faces paired with socially positive behaviours as more trustworthy than those paired with negative behaviours, and this effect was equally strong for a face set of 400 faces as for a set of 100 faces, which suggests that the set capacity for this social learning is impressively large.

Rule, Slepian and Ambady (2012) found that information about trustworthiness and deception can have a knock-on effect to downstream cognitive processes such as memory: untrustworthy faces are remembered better than trustworthy faces, and this effect emerges whether faces physically appear as trustworthy or untrustworthy, or if they are described as such using third-party information. This is also something that can be seen to a degree in the results of Falvello et al. (2015) − ratings for faces associated with positive behaviours were rated on average about 0.2 points higher on the 9-point trustworthiness scale than were faces associated with neutral behaviours, while faces associated with negative behaviours were rated about 0.75 points *lower* than neutral faces, which suggests that that effect was primarily driven by better memory for untrustworthy identities.

It is important to recognise some limitations when generalising from research that uses behavioural vignettes, and to acknowledge that these types of third-party information cannot tell us everything about how we form social judgements and learn about the trustworthiness of others. For example, Kuzmanovic et al. (2012), who found posterior cingulate activation in response to trait-diagnostic vignettes, found no such activity if trait evaluations were made on the basis of nonverbal expressive behaviours, and these were instead associated with amygdala activation. The fact that separate brain regions are recruited for these different tasks suggests that they arrive at the same

end goal (a social evaluation) through different means, and therefore that this posterior cingulate activity relates more to whether learned information is verbal rather than whether it is social.

Rather than a specifically social mechanism for evaluating other people, research using behavioural vignettes seems to recruit modality-independent affective mechanisms that are not specifically tailored for social information. This is shown by the fact that Falvello et al. (2015) found similar learning capacity for place-description associations as for face-description associations. This points to an interpretation of results from studies that use third-party information as showing fundamentally more general learning mechanisms than we might employ naturally within a given social context.

Imagine the social world as a geographical place, and navigating through it as a literal navigational task. In this case, first impressions from physical appearance and physiognomy are like a compass. They provide useful information that, for short journeys (i.e. short interactions with little chance of future consequences) may well suffice. You may also sample this information multiple times in the early stages to calibrate and get your bearings with the journey. Conversely, third-party information is like street signs − information from the external world that structures and guides your expectations and more often than not will serve as useful sources of information that guide your behaviour.

However, there is a crucial third source of information that has been less explored in the literature, and that is one's own experiences. To use the geography analogy, this may be noticing information (e.g. the terrain, the visibility, potential shortcuts) not offered by signs and compasses. In a social sense, this kind of information is direct experience of a person's behaviour, a source that should arguably be privileged over third-party information and lead to stronger representations as it pertains to the self

(Ham & van den Bos, 2006, as cited in Uleman, Adil Saribay & Gonzalez, 2008) as it has a fundamental basis in episodic memory. This information can also be used to calibrate first impressions – if a person looks particularly aggressive but consistently behaves in an agreeable and friendly manner, then the initial impression should be updated in line with new information.

Learning directly from experience can be a powerful cue to a person's character and, through the use of social dilemma and economic games, trustworthiness in particular can be both manipulated and measured. An example of an economic trust game might be a situation where a participant and a confederate are taking turns investing in a central pot. If both players cooperate, the participant will invest their money (and so take a risk), the confederate will reciprocate, and the total sum of money will be split between them, which is a fair arrangement. Alternatively, if the participant cooperates and the confederate betrays their trust, the confederate gets all of the money and the participant gets nothing. If both players attempt to cheat their partner then both players get nothing. You can therefore manipulate trust by manipulating the behaviour of the confederate, and measure it by monitoring how much the participant is willing to invest. Other variations of the task might use different sequences, risks, or strategies to interrogate specific features of participants' trust.

These types of games have been used to good effect to investigate realistic in-the-moment negotiations of trust. For example, research has shown that women are more forgiving of trust transgressions than are men (Haselhuhn, Kennedy, Kray, Van Zant & Schweitzer, 2015). There is also evidence that enhanced memory for untrustworthy individuals may be driven by expectations. Buchner, Bell, Mehl and Musch (2009) found that participants who engaged in an economic game with

cooperative and cheating faces showed better source memory for cheaters in a surprise memory test, but Bell, Buchner and Musch (2010) found better old-new recognition and source memory for both cooperators and cheaters over neutral control faces, and what better predicted source memory of faces was the comparative rarity of the face's behaviour. Later studies (Bell, Buchner, Erdfelder et al., 2012; Bell, Buchner, Kroneisen & Giang, 2012) confirm that this enhanced memory for cheaters appears to be driven more by expectancy violations − that is, people remember surprising or unexpected events (in a style of learning similar to a Rescorla & Wagner, 1972, model, where learning is better for unexpected than expected events) rather than using memory specifically tailored for deceivers.

Economic games are a powerful tool for measuring and manipulating trust as they rely on direct experience of trustworthy or untrustworthy behaviour that participants use to make decisions. Compared with trust judgements elicited by faces' physical appearance, these techniques have the advantage of being able to make the same identities as both trustworthy and untrustworthy within an experimental design. On the other hand, these techniques also have the advantage over third-party information as the deceptive behaviour participants are using to judge the faces is personally relevant and has a foundation within episodic memory, which allows researchers to examine deception and betrayal in a more ecologically valid context. Using combinations of these different methods dimensions can be manipulated orthogonally to investigate what happens when different signals about a person's trustworthiness compete (e.g. Suzuki & Suga, 2010; Rezlescu, Duchaine, Olivola & Chater, 2012).

However, a disadvantage of economic trust games is that they are an overt, explicit manipulation that is exogenously driven. Participants know that the behaviour of their

partners is a crucial manipulation in a lot of these games, and the mere context of a competitive game may change how they behave towards cooperators and cheaters. On the other hand, in the real world we do not typically operate in such a competitive context, and deceptive or cooperative behaviour may therefore be treated differently than it is in game situations. Indeed, in real social interactions our decisions may be driven by behavioural cues outside of our conscious awareness − resulting in so-called 'gut' decisions.

There is some evidence that unconscious processing is a powerful and useful tool, and in some cases it may be necessary for processing of information to be implicit. In a recent review ten Brinke, Vohs and Carney (2016) describe research into lie detection, and describe a trend where participants perform poorly when asked to explicitly judge whether a partner has told a lie or the truth. However, in more implicit scenarios, such as when participants do not know that lying is a part of the experimental manipulation, participants often demonstrate behaviours that suggest suspicion (for example, some report feeling less comfortable with their partner and are less likely to engage in economic games with that person, despite reporting no explicit awareness). This counter-intuitive trend, that people are better at detecting or learning deception when distracted by another task, points to an unconscious social monitoring system. This could be advantageous in that an unconscious system would be less susceptible to interference from other factors, such as the potential social repercussions of falsely accusing somebody of lying (which may make explicit judgements more conservative).

Kaunitz, Rowe and Tsuchiya (2016) have also found that incidental learning can impact conscious memory for faces. In a visual search task, participants were asked to search arrays of between 20 and 55 faces for a particular target. Using eye-tracking, the

authors monitored which non-target faces they had fixated during their search, and then judge whether those faces had been fixated and rejected during the search. Despite these faces being incidental to the task (meaning there is no strategic value in remembering them), participants could remember up to seven irrelevant non-target faces. This effect was specific to faces in a natural configuration, as when the same images were inverted participants' memory capacity was reduced to only three identities. As such, incidental learning can occur outside of awareness, but it can also intrude on explicit and conscious memory processes.

The majority of studies looking at social learning use explicit manipulations to change individuals' social judgements of another person. One outstanding question that this thesis looks to address is how social learning (in particular social learning of trust, as this is a powerful and fundamental tenet of social cognition) can occur via *incidental* learning processes. That is, the experiments presented here use a technique where the faces and their behaviour are irrelevant and ignored while participants focus on a different task. The proposal is that such information is nonetheless monitored and used to make social decisions, and that this may take place without explicit awareness of the social cues. The paradigm that we use to examine this manipulates trust using gaze cueing.

1.1 Gaze Cueing

Eye gaze in humans is a uniquely powerful social cue, and can be used to alert others to potentially interesting, relevant or dangerous stimuli in the environment, as well as encouraging interpersonal engagement in social interactions. It has been proposed that the eye itself has evolved to subserve these interactive features; the strong contrast

between the iris and sclera that is unique to humans enables easier decoding of eye gaze direction, and humans are more sensitive to the subtle information provided by these physical features than non-human primates (the Cooperative-Eye Hypothesis; Tomasello, Hare, Lehmann & Call, 2007).

When we see a person's gaze change direction, we automatically reorient our own attention to where they are looking (Friesen & Kingstone, 1998). This is similar to basic directional cueing that can be generated with arrow stimuli (Tipples, 2002), in that any object that cues the spatial location of a target improves subsequent processing of the target when it appears. Despite similar levels of cueing to arrows in laboratory settings, however, eyes are a unique cueing stimulus and research with eye-tracking has shown that people will orient their eyes towards the heads and particularly eyes of another person over and above any other type of stimulus, including arrows, in natural scenes (Birmingham, Bischof & Kingstone, 2009).

Gaze cues cannot be ignored completely (Driver et al., 1999), and susceptibility to them develops early in life (Hood, Willen & Driver, 1998; Reid, Striano, Kaufman & Johnson, 2004). Gaze cueing has been demonstrated in a variety of different paradigms, and has been shown during face-to-face interactions (Lachat, Conty, Hugueville & George, 2012), using subliminally presented stimuli (Xu, Zhang & Geng, 2011; Chen & Yeh, 2012), and even using illusory faces (Takahashi & Watanabe, 2013). Kuhn, Tatler and Cole (2009) have found that the direction of eye gaze is the most powerful biological tool to misdirect participants' attention during a magic trick. Gaze cueing effects are reliable, robust, and difficult to change, and evidence that cueing effects are susceptible to features such as the trustworthiness or identity of the face is inconsistent (Süßenbach & Schönbrodt, 2014; Hungr & Hunt, 2012; Strachan, Kirkham & Tipper, n.d.; Frischen

& Tipper, 2004).

We know that gaze cues can be used to deceive or cooperate with partners, and so if a partner directs or misdirects us to a target we can extrapolate information about that person's personality and intentions. We also learn early in life to monitor the behaviour of those around us to measure their reliability − infants have been shown to prefer to follow the gaze of adults who reliably look towards the location of a reward at 14 (Chow, Poulin-Dubois & Lewis, 2008) and 8 months old (Tummeltshammer, Mareschal & Kirkham, 2014). Extracting such information becomes more sophisticated as we grow older.

1.1.1 Incidental social learning from gaze cues

Bayliss and Tipper (2006) used a gaze-cueing paradigm where participants had to perform an object categorisation task on images that appeared on either side of the screen while an unrelated face appeared in the centre and either always cued the correct location (valid) or always cued the incorrect location (invalid). See Figure 1.1 for an outline of the trial sequence. When subsequently asked to judge which faces appeared more trustworthy in a 2-alternative forced-choice (2AFC) paradigm, participants chose faces that consistently looked towards the targets (valid-cueing) over those that always looked away from targets (invalid cueing).

Interestingly, when asked which of the face pairs appeared more frequently during the experiment, participants chose the invalid face over the valid. This could reflect better learning of invalid or untrustworthy identities, as they are associated with the unexpected feeling of being deceived in a non-competitive context, in line with previous research (Rule et al., 2012; Buchner et al., 2009; Bell et al., 2010; Bell, Buchner,

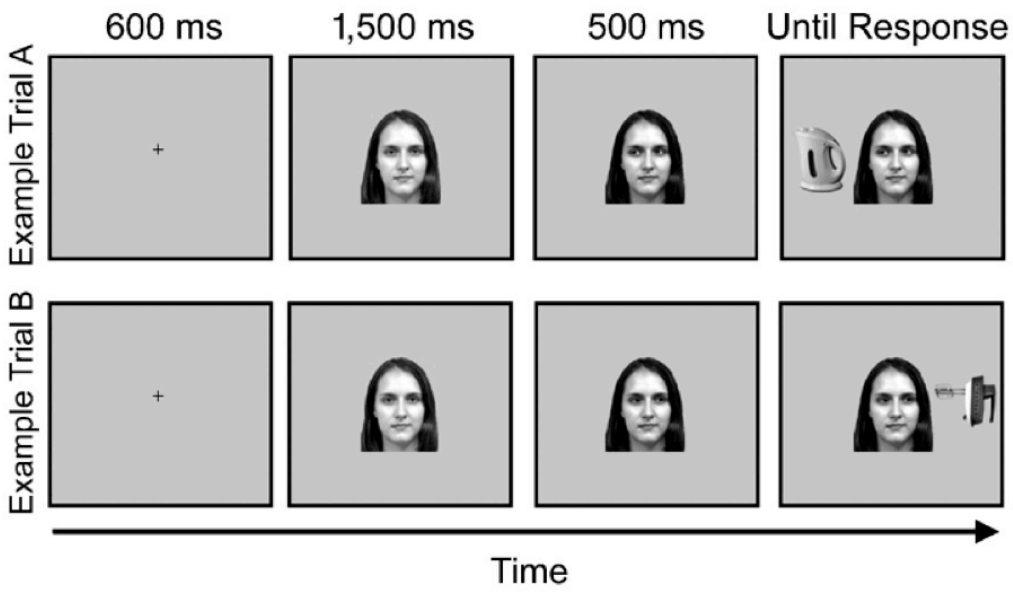

Figure 1.1: Example of valid-cueing face trials used in Bayliss and Tipper (2006). A face would appear in the centre of the screen and cue either left or right, then the object would appear. Participants had to categorise the object as either a kitchen or garage item. Faces would consistently cue either validly or invalidly throughout the experiment. Faces were matched in pairs and the pair member that was valid was counterbalanced across participants.

Erdfelder et al., 2012; Bell, Buchner, Kroneisen & Giang, 2012).

It is important to note that participants were instructed to focus on quickly and accurately classifying the peripheral targets during the gaze-cueing procedure and to ignore the central face as irrelevant, which means that this learning via gaze cueing behaviour was incidental. When interviewed after the fact, participants also expressed little suspicion of the critical manipulation of gaze behaviour. Overall, this finding reflects a form of incidental learning about the identities involved in this experiment. Although participants did not report consciously remembering the gaze behaviour of the faces, their choices as to which of a pair was more trustworthy consistently reflected that person's cueing validity.

Since this initial study there have been several replications of this finding. Bayliss, Griffiths and Tipper (2009) replicated the study using faces that varied in their emotional expression. They found that this learning (i.e. the likelihood of selecting the valid face as the trustworthier of the pair) was strongest when faces expressed a smile, was only

marginal for neutral faces, and was absent when faces were posing angry expressions. The fact that this effect is affected by higher order social information such as emotion, which conveys a person's internal state and motivations, suggests that this learning of trust from gaze contingencies reflects a uniquely social learning mechanism. This is further supported by the fact that similar cueing from individual but non-social arrow stimuli does not result in acquired attitude changes to these stimuli (i.e. people do not like valid arrows more than invalid arrows; Manssuer, Pawling, Hayes & Tipper, 2015).

The results of Bayliss and Tipper (2006) raise several questions. While there is a hint that there may be differences between how valid and invalid faces are remembered, and this would be in line with previous research, it is impossible to tell whether such asymmetries are present when using 2AFC as a measure, as the responses to valid and invalid faces are inherently co-dependent (that is, one cannot select the valid face as being trustworthier than the invalid face without also selecting the invalid faces as being untrustworthier). Rogers et al. (2014) replicated the original experiment but introduced an economic game to measure how gaze cueing behaviour might impact participants' real-world decisions about the face identities. They found that participants invested more money with valid than invalid faces, even at their own expense (that is, even when there was no chance of reciprocity), which suggests that their learned knowledge of faces' cueing behaviour can have real-world social consequences.

Manssuer, Roberts and Tipper (2015) and Manssuer, Pawling et al. (2015) have updated the original paradigm by asking participants to rate each face on trustworthiness using scalar ratings at the beginning and end of the experiment. This way, they can measure not only how individual faces relate to each other, but also how the crucial manipulation of gaze behaviour *changes* the apparent trustworthiness of a face over the

course of the experiment. This technique also has the advantage of being much easier to apply than conventional economic games such as those used in Rogers et al. (2014).

1.1.2 A model of incidental social learning from gaze cues

The aim of the current body of work is to explore the mechanisms behind how people can incidentally learn social information from task-irrelevant behaviours such as gaze direction. Imagine a model of face processing similar to that posed by Bruce and Young (1986) or Haxby, Hoffman and Gobbini (2000), where there are two streams of information processing about faces. One stream is thought to process invariant aspects of faces such as their identity, gender, age, and race, and is subserved by ventral brain areas in the fusiform gyrus. The other stream processes more transient information that can change from moment to moment, such as eye gaze, and are subserved by more dorsal temporal areas such as superior temporal sulcus. Incidental social learning from gaze cues points to some communication between these streams of information, where gaze behaviour feeds into stored identity representations (which also receive input from the invariant stream) to inform later trustworthiness judgements.

There are many potential ways in which these streams could communicate. Consider the simple model shown in Figure 1.2. This shows a two-stream model for information from faces, leading from the earliest stages of face processing (where the structural configuration is recognised as a face, likely subserved by face-selective regions in occipital cortex) through to a unit that stores information about a particular identity (likely subserved by anterior temporal areas that relate to semantic and representational memory; Haxby et al., 2000), which can then be used for higher order social tasks such as mindreading. Labelled are examples of how processing of information can be distorted

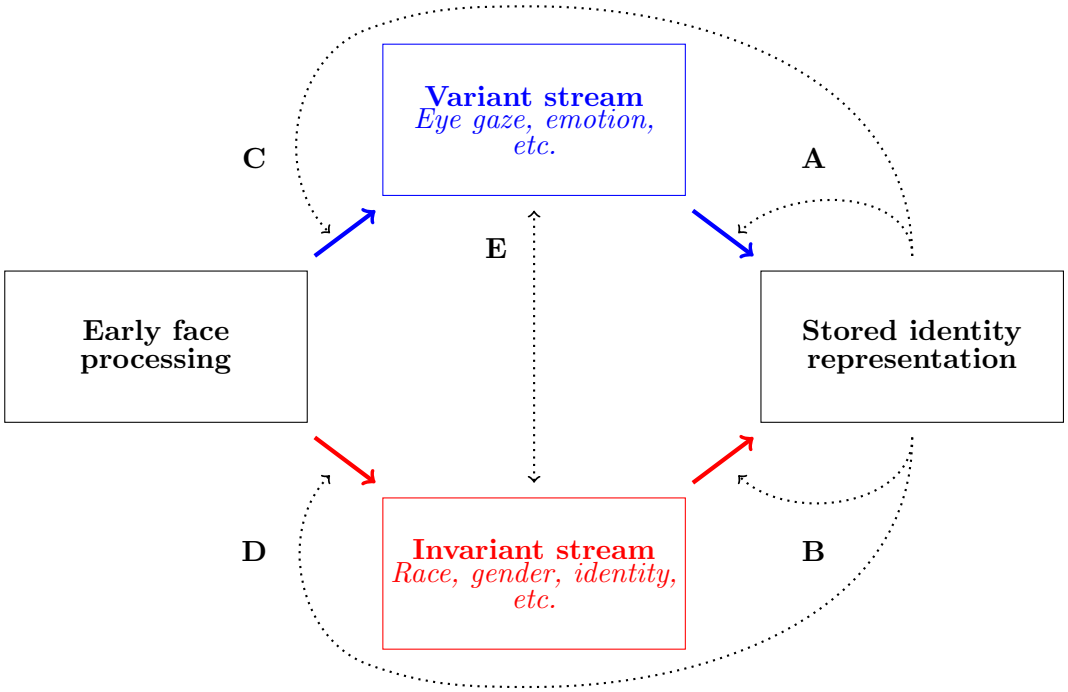

Figure 1.2: A model of incidental social learning from gaze cues. Visual information enters the model from the left, through early face processing systems that identify the face-like configuration and structure. Information is then processed by separate streams: an invariant stream, in red, which processes information that is unlikely to change over the course of an interaction (e.g. identity), and a variant stream, in blue, which processes information that is likely to change (e.g. gaze direction). These streams then feed into a stored representation of the individual's identity, which can be used later to process incoming information. Some examples of feedback communications are shown: A. and B. processing of variant (A.) and invariant (B.) information is not affected by person knowledge, but the integration of this information is affected by what is already known about that identity; C. person knowledge affects processing of variant information such as eye gaze or emotion; D. person knowledge affects processing of invariant information such as gender or race; E. either variant or invariant information is affected by the content of the other.

or affected by other modules in the system. In this model, processes A. and B. would serve as filters for incoming information − evaluating information from both streams about an individual so that it can be integrated with the representation for that identity. These processes likely depend on the nature of the stored identity representation, or person knowledge, as well as the surrounding context. For example, if the identity representation is sparse (i.e. the person is unfamiliar) certain types of information may be privileged over others − for example, if the person is unfamiliar then negative information may be privileged over positive (a 'stranger danger' monitoring system).

On the other hand, processes C. and D. show what might happen when stored person knowledge is clear enough that it can interfere with earlier processing of incoming

information. For example, Süßenbach and Schönbrodt (2014) used famously trustworthy or untrustworthy characters from popular films (e.g. Viggo Mortensen's Aragorn from *Lord of the Rings* as a trustworthy example; Anthony Hopkins' Hannibal Lecter from *Silence of the Lambs* as an untrustworthy example) as gaze cueing stimuli. The authors found that gaze cueing from untrustworthy individuals was inhibited, and this was potentially moderated by participants' levels of trait anxiety. On the other hand, Manssuer (2015) found no effect of trustworthiness on gaze cueing when trustworthiness was based on the physical appearance of individuals. This suggests that stored person knowledge can interfere with gaze following, but physical cues to trustworthiness (which are closely related to emotion and therefore likely processed along the variant stream also) appear to be processed in parallel, with little interference. With this in mind, we might therefore expect to see the magnitude of cueing costs reduce as participants learn more about the cueing validity of particular faces.

The final communication channel displayed in Figure 1.2 is a channel between the variant and invariant processing pathways, process E. Information in one processing stream may be able to influence information in the other. An example of such an effect might be the Own-Race bias in emotion processing, where information about a face's race (processed in the invariant stream) affects the efficiency of recognising emotional expressions (processed in the variant stream). Evidence for this effect in emotion has been shown repeatedly (e.g. Elfenbein & Ambady, 2002b, 2002a; Hu, Wang, Han, Weare & Fu, 2015; Beaupré & Hess, 2006), but there is comparatively little research examining the processing of information such as eye gaze. Some evidence suggests that eye gaze can have an effect in the other direction − memory deficits for other-race faces appear to be related to direct eye gaze, for example (Adams, Pauker & Weisbuch, 2010). However,

further exploration of this pathway, and how it might affect the assimilation of variant information into the stored identity representation, remain relatively unexplored.

This model is dramatically oversimplified − none of the modules presented here are straightforward enough to be encompassed in a single box, as the two streams likely include subordinate, parallel streams for different types of information, and the stored identity representation covers a range of processes including mentalising, long-term memory, and other top-down processes such as information from third-party gossip. However, even this simplified model paints a complex picture of social learning. The aim of this thesis is to explore how this model might behave in a particular set of circumstances: that is, how do we build an identity representation on the basis of information from the variant stream − eye gaze − while controlling for information from other sources. By controlling the invariant information (identity) and top-down knowledge, we can explore some of the key features and limitations of learning about trustworthiness from observing a person's gaze behaviour.

For example, we can explore whether, as participants are exposed to more and more information and use that to build a representation of an identity, acquired knowledge of cue validity affects earlier processing of gaze cues later in the experiment (i.e. by the end of the experiment are people less susceptible to gaze cues because they have learned which faces are trustworthy; pathway C.). We can explore how information such as emotion, which is represented in the same pathway, can change the assimilation of eye gaze behaviour into a stored identity representation (pathway A.). Or we can examine whether gaze behaviour can change how information from alternative sources, such as top-down knowledge of the social group membership or race of the face, is incorporated or represented (i.e. can knowledge of out-group members' helpful behaviour change how

we make judgements based on that information; pathway D.). We can even ask whether eye gaze behaviour can inform trustworthiness judgements in the absence of conscious awareness. The gaze cueing paradigm used here is a powerful tool for investigating the intricacies of social learning from behaviour.

1.2 Scope of this thesis

The current thesis aims to explore key questions and features associated with this incidental learning of trust. A key underlying theme of all experiments is to monitor the incidental learning of gaze contingencies associated with each face stimulus, and to explore how these contingencies then manifest as social decisions about trust. Over six experimental chapters and eighteen separate experiments, we aim to describe the mechanisms subserving this incidental social learning.

Chapter 2 explores certain key features of this incidental learning, including some important replications of previous results. As well as exploring the role of emotional expression in trust learning, this chapter also addresses whether this effect is specific to trust as a social judgement, and whether or not this learning is a result of conscious awareness of gaze contingencies. In Chapter 3, we look at how durable the effect is and how initial levels of familiarity with the faces may lead to more stable representations of individual identities over time. This chapter also addresses whether the effect is strong enough to survive reversing learned individual gaze behaviours unexpectedly.

Chapter 4 looks at whether there are other ways of capturing these learned representations of trustworthiness without directly asking participants to explicitly rate the faces, while chapters 5 and 6 examine how the identity of the cueing face − as a member of either the participant's own social group or an out-group, using both minimal

and real-world (racial) social groups – influences learning of trust. These latter chapters also address some differences between using implicitly salient real-world groups and using explicitly instructed laboratory-based groups in social research.

Finally, Chapter 7 examines the role of visuomotor fluency in this learning, and questions whether the same learning of trust can be seen in the absence of gaze cueing – without any apparent behaviour from the faces, this chapter explore whether simply associating some faces with disruptions to fluent visuomotor processing is enough to lead to changes in trustworthiness.

Chapter 2. Key boundaries of incidental trust learning

This chapter explores some of the key features of incidental learning of trust from gaze cues. This effect has been shown previously using a 2-alternative forced choice (2AFC) rating procedure, where participants are shown pairs of faces (one that has provided valid cues throughout gaze cueing and the other invalid cues) and asked to select which they feel is the more trustworthy of the two. However, this does not explain how these changes come about: whether this effect is driven by an increase in trustworthiness for valid faces, a decrease for invalid faces, or a bidirectional mix of the two. To further investigate the specific nature of changes in trust ratings two scalar ratings of trustworthiness are used in the current studies, one at the beginning and one at the end of the experiment, to track changes in trustworthiness for both valid and invalid faces (c.f., Manssuer, Pawling et al., 2015). This more sensitive measure provides the ideal approach to further investigate key boundary conditions for the understanding of the processes mediating incidental learning of trust.

Based on findings by Manssuer, Roberts and Tipper (2015) that event-related potentials (ERPs) to valid and invalid cueing identities are characterised by a late positive potential (LPP) component that is associated with stimulus valence and that rises only in response to invalid faces, we expect that this learning of trust may be specialised to detect identities that are likely to deceive. As such, we predict that the behavioural pattern of results in this experiment would be primarily characterised by a decrease in trust for invalid faces.

One outstanding issue where this new measure may be benefial concerns the role of facial emotion. Bayliss et al. (2009) found that gaze-contingent trust effects appear to rely on a positive social context, as they found no trust effects when the faces expressed

anger and a reliable effect only when the faces smiled. However, the neutral expression condition was somewhat ambiguous, as participants were only slightly more likely to select the valid face as the more trustworthy of a matched pair in a 2AFC paradigm. This previous work using forced choice between valid and invalid cueing individuals creates a somewhat blunt measure; we can see that valid faces are preferred over invalid, but as each judgement made about a face is also inseparable from the judgement made about its pair (i.e. in a 2AFC, you cannot judge one face as trustworthy without judging the other to be untrustworthy), we cannot tell what the underlying mechanisms might be. We also do not know how exactly the emotional expression of a face might change the pattern of results using forced-choice measures of trustworthiness. It still could be the case that neutral faces are not sufficient to elicit learning of trust −there is a wealth of evidence that smiling faces are treated differently from neutral faces in various social interactions, both when measured by trustworthiness judgements (Hehman, Flake & Freeman, 2015) and by more implicit measures (Wang, Ranganath & Yonelinas, 2014) − or it may be that we can detect trust learning with neutral faces using this new, more sensitive measure.

Therefore we examined the role of emotion in the incidental learning of trust in conditions where faces expressed neutral emotions (Experiment 2.1) and when they smiled (Experiment 2.2). We aim to unequivocally identify whether incidental learning of trust from gaze cueing can be detected when faces express neutral emotions. Additionally, and more importantly, we can assess whether the pattern of learning (whether valid faces increase in trust and invalid faces decrease in trustworthiness, or whether the effects are unidirectional) is the same for both neutral and positive emotions.

The further issue this chapter investigates is whether incidental learning of eye-gaze

patterns is specific to judgements of trust, or generalises to other emotional assessments, such as liking of a person. One might assume that trust and liking will be closely related: if we trust someone, we are more likely to like them. Indeed, the two are often conflated as aspects of warmth in dual-dimension theories of social cognition (e.g. Fiske, Cuddy & Glick, 2007). However, subtle behaviours that can be used to deceive others, such as gaze shifts, could have quite specific effects on trust. For example, whether to invest money with another person is influenced by incidental learning of patterns of gaze shifts, as is the decision to be altruistic while computing the likelihood that such an act will be reciprocated in the future (Rogers et al., 2014). Such decisions might not be affected by general feelings of liking, for example we may trust a lawyer to do their utmost to preserve our freedom, but we may not like them on a personal level; the two feelings are distinct, and can be separated. To our knowledge, there is little previous work directly addressing the question of whether trust and liking are functionally similar in this way, so in Experiment 2.3 we replace the trustworthiness ratings of Experiment 2.1 with likeability ratings, to see if this incidental learning is specific to trust or if there is a broader affective spillover into other social judgements.

The final issue that this chapter addresses is that of awareness. Bayliss and Tipper (2006) found that most participants did not report that they were aware of the experimental manipulation, suggesting that they demonstrated trust learning in the absence of conscious awareness. However, the current set of experiments uses an updated paradigm where participants rate the faces for trustworthiness at the beginning of the experiment as well as the end. It is possible that this pre-experiment rating could cue participants to the nature of the experiment early on and so encourage demand characteristics that may be driving any effects. Experiment 2.4 tests this explanation by

asking participants to make explicit recollections of which faces were valid and which were invalid, with the aim of seeing whether explicit memory for the face's behaviour can explain this incidental learning of trust.

2.1 Experiment 2.1

We re-examine whether incidental learning of trust can be detected with faces expressing neutral emotion, and if such an effect is detected, what pattern of changes in trust are revealed. In Experiment 2.1 we compare the reaction times (RTs) to valid and invalid gaze cues, and the trustworthiness changes between the beginning and end of the experiment.

2.1.1 Methods

Participants

A total of 24 participants (18 female, $M_{age} = 19.96$, $s.d. = 1.49$) volunteered for the study in return for payment or course credit. All were students of the University of York. All participants provided written consent and the study was given ethical approval by the Departmental Ethics Committee of the University of York Psychology Department.

Stimuli

Target stimuli for the object classification task were the kitchen and garage object images used in Bayliss, Paul, Cannon and Tipper (2006). There were 13 unique objects in each group (kitchen/garage) that appeared in both left and right orientations. All of the stimuli were coloured in blue. In total, there were 52 individual images used in the experiment. Face stimuli were taken from the Karolinska Directed Emotional Faces

(KDEF) stimulus set (Lundqvist, Flykt & Öhman, 1998), and included sixteen images; eight male and eight female. These faces were initially selected by eye from a figure in the Supplementary Material of Oosterhof and Todorov (2008), in which the faces from this set are plotted along six judgement dimensions. The faces used were all taken from the centre (within 1SD from the intercection of all six dimensions) of this plot, so the faces used in our experiments were, compared with the rest of the KDEF set, as close to neutral trait judgements as possible.[1] These faces were split into two groups, which would appear as either valid or invalid cues in the experiment (counterbalanced across participant). The eyes of each face were manipulated using Adobe Photoshop CS6 to generate faces where the eye gaze was either straight ahead, left or right. Unaltered images were used for the trustworthiness rating sections.

The study was run on an Intel Core i5 PC with a 21.5" monitor. The experiment was presented using E-Prime 2.0 software with a white background throughout and the resolution set to 1024x768 pixels. Participants were sat approximately 60cm from the display, and during trustworthiness ratings the face stimuli had a visual angle of 19.29° horizontally and 20.97° vertically, while during gaze-cueing the face stimuli had a visual angle of 13.36° horizontally and 14.93° vertically.

Design and Procedure

Participants were told that they would be asked to perform an object categorisation task on images of objects that appeared on the left or right side of the screen, and to respond

[1] Although means and standard deviations were not retrieved from Oosterhof & Todorov's (2008) material, we can validate our assumption that these groups of faces were close to neutral by examining the pre-ratings assigned to them in Experiments 2.1 and 2.2, as well as Experiments 7.1 and 7.2 (experiments included in Strachan, Kirkham, Manssuer & Tipper, 2016). As the pre-ratings occurred before any participants had a chance to experience the faces within an experimental context, any differences could only be explained by physical cues to trustworthiness, and the combined power of these four experiments would be sufficient to detect this. We explored this and found that the pre-ratings for faces in one group (M=-2.71, $s.d.$=13.04) did not significantly differ from the other (M = 0.85, $s.d.$ = 6.82; $t(14)$ = -0.68, p = 0.506).

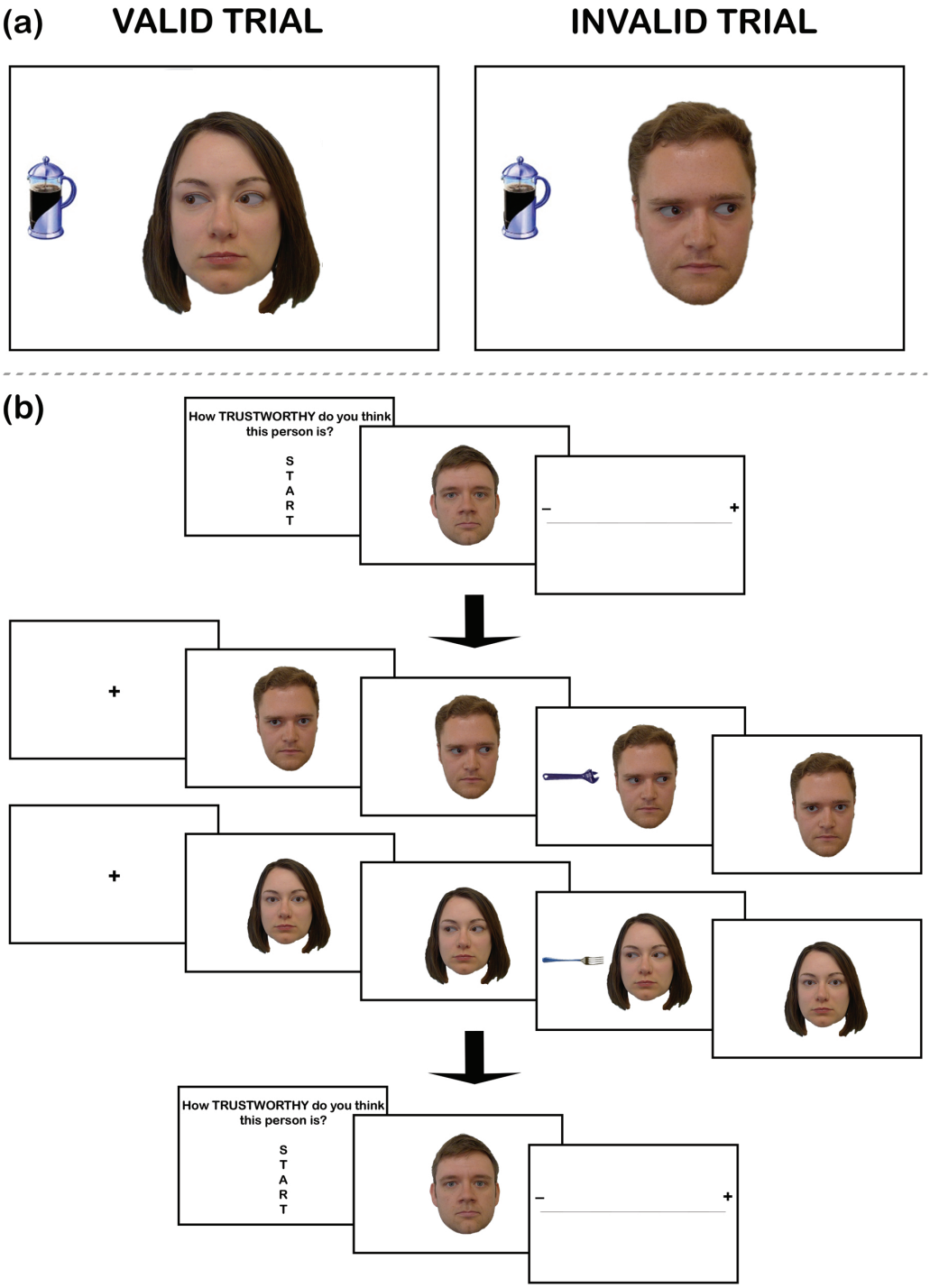

Figure 2.1: Outline of gaze-cueing procedure used in Experiment 2.1. (a) Examples of a single face on a valid (left) and invalid (right) trial. A participant would see this face in only one of the two conditions; that is, it would only ever be valid or invalid whenever it appeared throughout the experiment. (b) The trial sequence of the whole experiment. Participants made trustworthiness ratings of the faces at the beginning (top) and end (bottom) of the experiment, and in the main body participants categorised the kitchen and garage objects with key-press responses while ignoring the faces.

with whether these were garage or kitchen objects. They were also told that the central face images were irrelevant and to be ignored. Before the experiment participants were allowed to study printed versions of the kitchen/garage images, in order to familiarise themselves. This was done firstly to ensure that participants had the knowledge of what each object was, and secondly to make sure that early responses from the first trial block were not confounded by uncertainty as to the object categories of the targets.

Each trial began with a 600ms fixation cross in the centre of the screen, which was then replaced by a face showing a direct gaze for 1,500ms. The face then shifted gaze either to the left or right for 500ms before the target stimulus appeared on either the same (valid) or opposite (invalid; see Figure 2.1a) side of the gaze direction. The target stimulus remained until 2,500ms had passed, following which an error tone would sound if an incorrect response was logged and the face shifted back to direct gaze for a further 1,000ms. A blank screen followed for 500ms before the next trial began. The trial structure is shown in Figure 2.1b.

The object categorisation responses were the H key and the Space bar of a keyboard, chosen because the H key appears directly above the Space bar on QWERTY keyboards and this direction was orthogonal to the possible location of the target. Participants were instructed to respond with their index finger on the H key and thumb on the Space bar. For half of the participants, H represented kitchen objects, while for the other half it represented garage objects.

In total there were five blocks of 32 trials each, and each face appeared twice in each block, once gazing left and once right. The order of faces was randomised, as was the order of target objects, the side that the target appeared, and the order of valid and invalid trials.

At the beginning and the end of the experiment, participants rated the original unmanipulated face images used to generate the gaze cueing stimuli. Participants were shown a calibration slide where they clicked in the centre to start, and then the face images were presented for 1,000ms. Participants were then instructed to click along an uninterrupted scale that appeared on the screen at a point that conformed to how trustworthy they felt the person was. The scale recorded responses between -100 and +100, calculated by the distance from the centre of the line of the participants' mouse click − responses to the left of the centre were coded as negative, while those to the right were coded as positive (these were indicated on the screen with a − and + sign at either end of the scale). Identities were presented in a randomised order.

To ensure that participants were naïve to the primary manipulations of the experiment, and as such that their trust ratings were due to implicit factors rather than demand characteristics, participants were interviewed after the experiment to see what suspicions they held of the experimental manipulations. Some participants did report suspicion of the primary research question when asked, but this issue is addressed directly in Experiment 2.4 and so we do not discuss it further here.

Data analysis

Before data were analysed, participants' responses were filtered to remove all error trials (where participants reported the incorrect answer in the object categorisation task) and RT outliers − RTs below 250ms (too short to process the stimuli) and above 2,500ms (indicating that participants had not given a response in the allotted time). The number of remaining trials was then compared with the original number of trials to check that all participants retained at least 70% of their total trials and had not scored below 70%

total correct on any one condition during the gaze cueing task.

Linear mixed effects models Data were analysed using a linear mixed effects modelling approach. RTs, accuracy rates and trustworthiness scores could have been analysed using conventional parametric techniques such as factorial ANOVAs or t-tests. However, we felt that in this paradigm it is particularly important to control for more than just subject-level variance. That is, there is also the issue of materials- or stimulus-level variance (differences in how specific faces might be treated). While the stimuli used throughout this thesis were pre-selected to be as controlled as possible, it is nonetheless possible that particular individual identities could affect results − for example, while all faces were selected as being similar in trustworthiness, there is still a distribution in trustworthiness that may have made certain faces appear more or less trustworthy when compared with their companions.

With this in mind, we elected to use linear mixed effects models that could control for variance at both the subject and stimulus level. To do this we used the `lme4` package in `R`. To give a conceptual introduction to how we used these models, below is an example of a maximum random structure or null model for modelling trustworthiness ratings:

```
> model.null <- lmer(Rating ~
+ (1|Subject) + (1|Identity) + (0 + Time|Identity) +
+ (0 + Time|Subject) +  (0 + Validity|Subject),
+ data = dataframe)
```

In this example, Rating is the outcome variable (trustworthiness rating) that we are measuring. 1|Subject and 1|Identity are the random intercept terms for the subject- and

stimulus-level random effects. The other terms refer to the slopes of those factors that are repeated within one of the random factors (e.g. each identity is shown at both times - pre- and post-experiment - and each subject experiences both times, so these are both included as random slope terms). When exploring the role of fixed factors in this effect, we include additional fixed parameters in the model:

```
> model.time <- lmer(Rating ~ Time +
+ (1|Subject) + (1|Identity) + (0 + Time|Identity) +
+ (0 + Time|Subject) +  (0 + Validity|Subject),
+ data = dataframe)
> model.valid <- lmer(Rating ~ Validity +
+ (1|Subject) + (1|Identity) + (0 + Time|Identity) +
+ (0 + Time|Subject) +  (0 + Validity|Subject),
+ data = dataframe)
```

These models give beta estimates and standard error values that we report here. In order to get a test of significance – that is, to see whether either of these models explains a significantly greater proportion of the variance in the data than does the null model, we compare each of these in turn with the null model using the `anova` function in `R` (c.f. Baayen, Davidson & Bates, 2008). This is conceptually similar to testing for a main effect in a factorial ANOVA. To test for interactions, we generate models with both factors defined: one that includes an interaction and one that does not, and these can then be compared in the same way:

```
> model.2factor <- lmer(Rating ~ Time + Validity +
+ (1|Subject) + (1|Identity) + (0 + Time|Identity) +
```

```
+ (0 + Time|Subject) +  (0 + Validity|Subject),
+ data = dataframe)
> model.interact <- lmer(Rating ~ Time * Validity +
+ (1|Subject) + (1|Identity) + (0 + Time|Identity) +
+ (0 + Time|Subject) +  (0 + Validity|Subject),
+ data = dataframe)
```

When analysing RTs in Experiment 2.1, we then generated models for each fixed factor individually (block-only and validity-only models), which were compared with the maximum random structure using the `anova` function. To measure the interaction of these factors, we modelled a block x validity interaction and compared this with a block + validity model, where both factors were fixed but there was no interaction. For RTs the maximum random structure would not converge with all random slopes defined, which indicates that the model is overfitting the data. In this case it is customary to remove random slopes until it fits, which in this case only happened once the block | identity and validity | subject terms were removed. We removed these terms from the null, block-only and validity-only models to allow for direct comparison. The two-factor and interaction models both converged with the block | subject and block | identity slope terms removed.

For accuracy rates, responses were averaged across subject, validity and block and calculated as percentage correct. For accuracy rates, stimulus effects were not modelled as this would mean each value could only be an average of two trials, which would be too much constraint on the variance. As such, each model was calculated with a 1 | Subject intercept term. Validity-only and block-only models were generated, as well as block + validity and block * validity, and these were all compared using the same `anova`

function as were RTs. The null and single-factor models would not all converge with the same random terms, and so all terms were removed from these models to allow for direct comparison. The two-factor and interaction models would not converge until the block | subject term was removed.

For trustworthiness ratings, the process was largely similar. We again generated a maximum random structure then generated models for each fixed factor individually (time-only and validity-only). These were compared with the maximum random structure using the `anova` function. A time x validity interaction model was then compared with a time + validity model with no interaction but both factors included. No trustworthiness models would converge until both the time | subject and time | identity error terms were removed.

For comparison, we also report more standard analyses in the form of factorial ANOVAs for each experiment (where appropriate) throughout this thesis. The results of these analyses are collated in Appendix A.

2.1.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 2.1 are shown in Figure 2.2. RTs were aggregated across subject, identity, block, and validity, and these were analysed using linear mixed-effects modelling. Adding block as a fixed factor significantly improved the fit of the maximum random structure model (β = -89.13, SE = 22.82, $\chi^2(4) = 20.84$, p <.001), as did including validity as a fixed factor (β = 43.50, SE = 7.39, $\chi^2(1) = 34.35$, p <.001). Comparison of the two-fixed-factor models (block + validity and block * validity) found no evidence of an interaction (β = -136.64, SE = 17.47, $\chi^2(4) = 5.27$, p

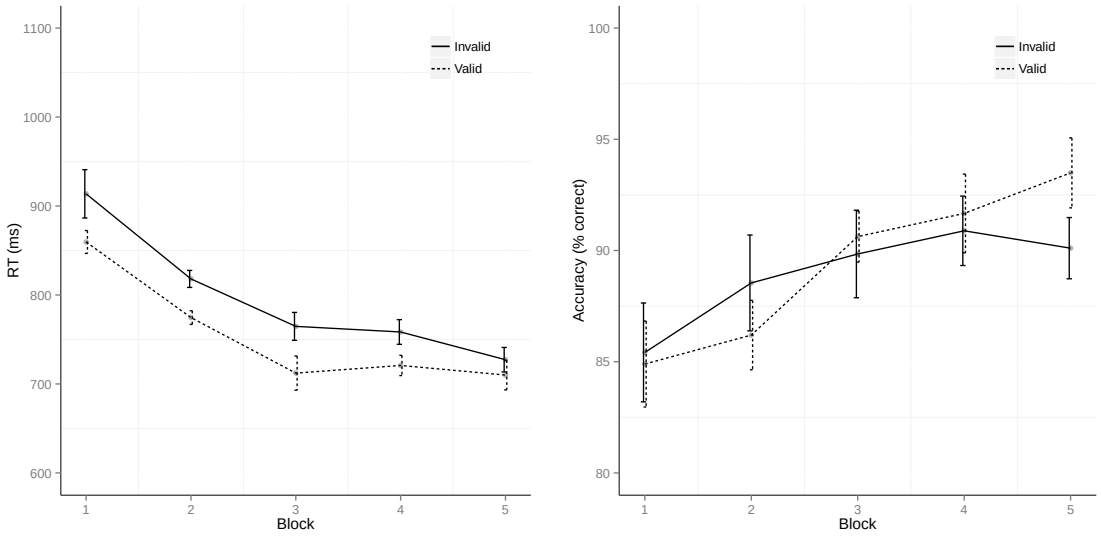

Figure 2.2: Timecourse of reaction times in milliseconds (left plot) and accuracy rates (percent correct; right plot) across all five blocks in Experiment 2.1 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

= 0.260).

The analysis of accuracy rates found that adding block to the null model significantly improved the fit ($\beta = 1.56$, $SE = 1.35$, $\chi^2(4) = 24.08$, $p < .001$), but including validity did not ($\beta = 0.07$, $SE = 0.85$, $\chi^2(1) = 0.01$, $p = 0.933$). Comparison of the two-fixed-factor models (block + validity and block * validity) found no evidence of an interaction ($\beta = 4.50$, $SE = 1.89$, $\chi^2(4) = 3.56$, $p = 0.468$).

There was no evidence of an interaction of block and validity over the course of the experiment, indicating that participants did not seem to adapt to and predict invalid gaze cues to anticipate the true location of a subsequent target. Although this is just one experiment, there is evidence from the wider literature that supports the idea that participants do not show diminished cueing costs over time (Manssuer, Roberts & Tipper, 2015; Manssuer, Pawling et al., 2015). Throughout the experiments in this thesis we find no evidence that cueing costs reduce over time as evidenced by an interaction of block and validity (this is reflected in the 2x5 ANOVAs of RTs and accuracy rates listed in Appendix A, along with figures similar to those presented here, with RTs and accuracy

broken down across blocks. In these analyses, only one experiment shows an interaction of block and validity in RTs (Experiment 3.1: $p = 0.045$), and one more is marginal (Experiment 2.4: $p = 0.097$), but no other analyses approach significance, indicating that these are likely spurious). Therefore, to be as concise as possible in the main text we collapse across blocks and look only at whether validity improves the model fit (that is, look for evidence of a gaze-cueing cost) throughout the main body of the thesis.

Trustworthiness ratings

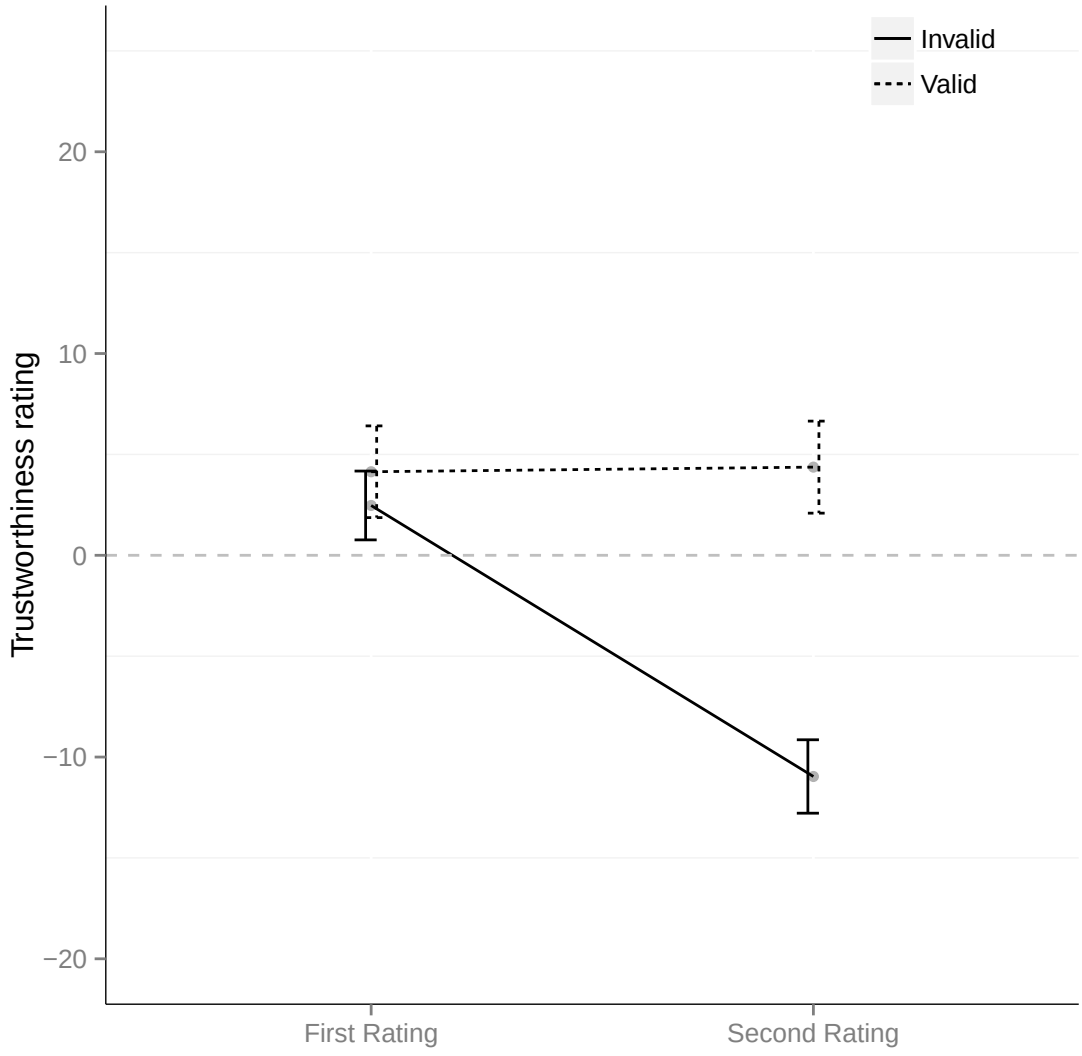

Figure 2.3: Time course of trustworthiness ratings over Experiment 2.1 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

The results of the trustworthiness ratings for Experiment 2.1 are shown in Figure 2.3. Adding time to the null model significantly improved the model fit (β = -6.59, SE = 1.70, $\chi^2(1) = 14.91$, p <.001), as did including validity (β = -8.60, SE = 2.40, $\chi^2(1) = 11.20$, p <.001). Finally, the interaction model (time x validity) fit the data significantly better than did the full model (time + validity), where both factors were modelled but without an interaction (β = -13.85, SE = 3.37, $\chi^2(1) = 16.74$, p <.001).

The results of Experiment 2.1 conform to the pattern one would expect based on previous studies. Observing somebody's gaze movements automatically triggers a shift in attention to the same location and resulting in faster processing at that location than somewhere uncued. This association of face identity and cueing behaviour appears to be incidentally encoded; even though the face was irrelevant to the task, individuals who looked away from the target object (invalid cues) were trusted less. This was supported by subsequent analyses, where we generated separate models of the valid (models converged once the time | subject term was removed) and invalid rating data separately (models would not converge with any error terms) and compared a null model with a model with time as a fixed factor (that is, to see if ratings changed significantly over time) and found that time improved the model fit for invalid faces (β = -13.52, SE = 2.42, $\chi^2(1) = 30.01$, p <.001) but not for valid faces (β = 0.33, SE = 2.32, $\chi^2(1) = 0.02$, $p = 0.885$), indicating that this effect was primarily driven by a decrease in trust to invalid faces.

A significant gaze-cueing effect was found in the RT data and was reflected in the form of decreased trust over the course of the experiment in response to misleading or invalid faces.

2.2 Experiment 2.2

This experiment aimed to explore how emotion affects this incidental learning of trust. This replicates all details of Experiment 2.1, but used smiling rather than neutral face images as the cueing and rating stimuli. Note that Bayliss et al. (2009) demonstrated significant learning of trust when the faces expressed positive emotions with a smile. However, when the faces expressed a neutral emotion the same pattern of trust was observed, but it was of marginal significance. Experiment 2.1 has shown that it is possible to detect significant learning of trust when faces are neutral, however the effect was asymmetrical, as invalid faces declined in trust and valid faces did not change. Whether faces expressing positive emotions produce this same pattern is the key question for Experiment 2.2.

2.2.1 Methods

Participants

24 participants (24 female, $M_{age} = 20.46$, $s.d. = 4.05$). volunteered for this study in return for a mixture of course credit and payment.

Stimuli, Design and Procedure

This experiment was identical to Experiment 2.1 in every way except that the KDEF faces used were frontal-view smiling faces in place of neutral (see Figure 2.4). All other details were identical.

Experiment 2.1: Neutral

Experiment 2.2: Smiling

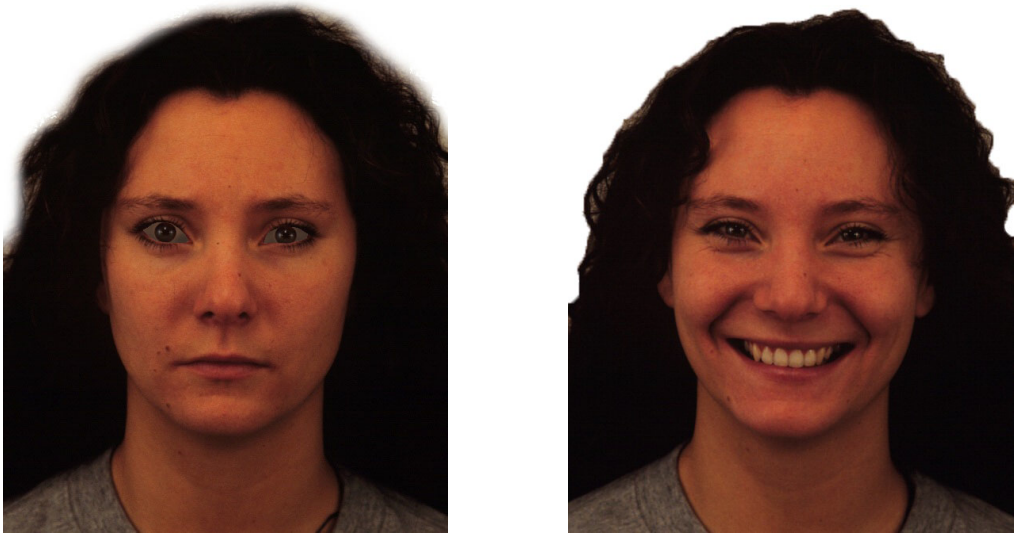

Figure 2.4: Examples of the neutral (left) and smiling (right) stimuli used in Experiments 2.1 and 2.2, respectively.

Data analysis

RT filters were applied in the same way as in Experiment 2.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials. Data were analysed in the same way as outlined in Experiment 2.1 Data Analysis section, with the exception that when analysing RTs and accuracy rates we dropped block (blocks 1-5, but see Appendix A) as a fixed factor in our models and instead compared a null with a validity-only model with validity | subject as the only random slope term. Models of accuracy would not converge with any random terms defined.

No models of trustworthiness would converge until both the time | identity and time | subject terms were removed.

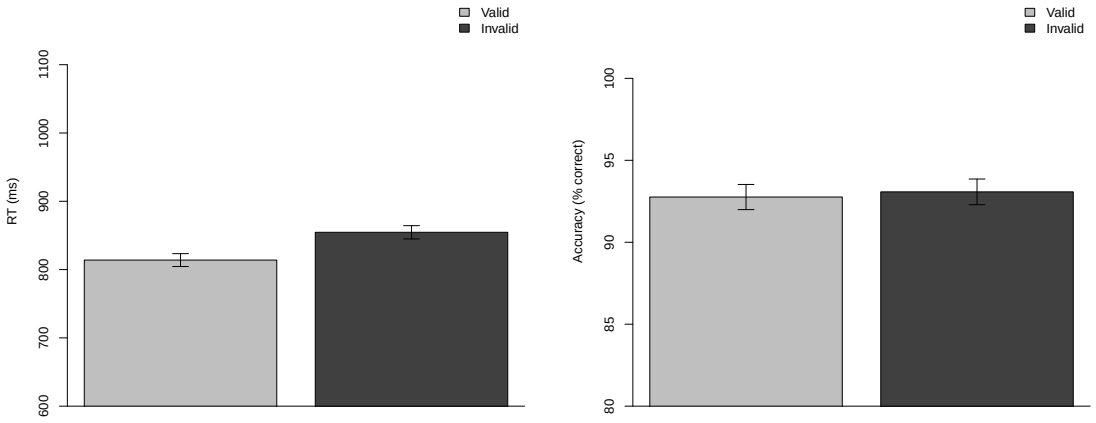

Figure 2.5: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 2.2 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

2.2.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 2.2 are shown in Figure 2.5. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -41.28, SE = 10.08, $\chi^2(1)$ = 15.19, p <.001). This improvement was not seen for accuracy scores (β = -0.01, SE = 0.02, $\chi^2(1)$ = 0.16, p = 0.687).

Trustworthiness ratings

The changes in trustworthiness ratings for the faces in Experiment 2.2 are shown in Figure 2.6. In this experiment, participants found invalid faces less trustworthy after the experiment than before, but this time there was a noticeable increase in trust for valid faces. Adding time to the null model did not significantly improve the model fit (β = -0.56, SE = 2.06, $\chi^2(1)$ = 0.07, p = 0.785), but including validity did (β = -8.36, SE = 2.63, $\chi^2(1)$ = 9.42, p = 0.002). Finally, the interaction model (time x validity) fit the data significantly better than did the full model (time + validity), where both factors

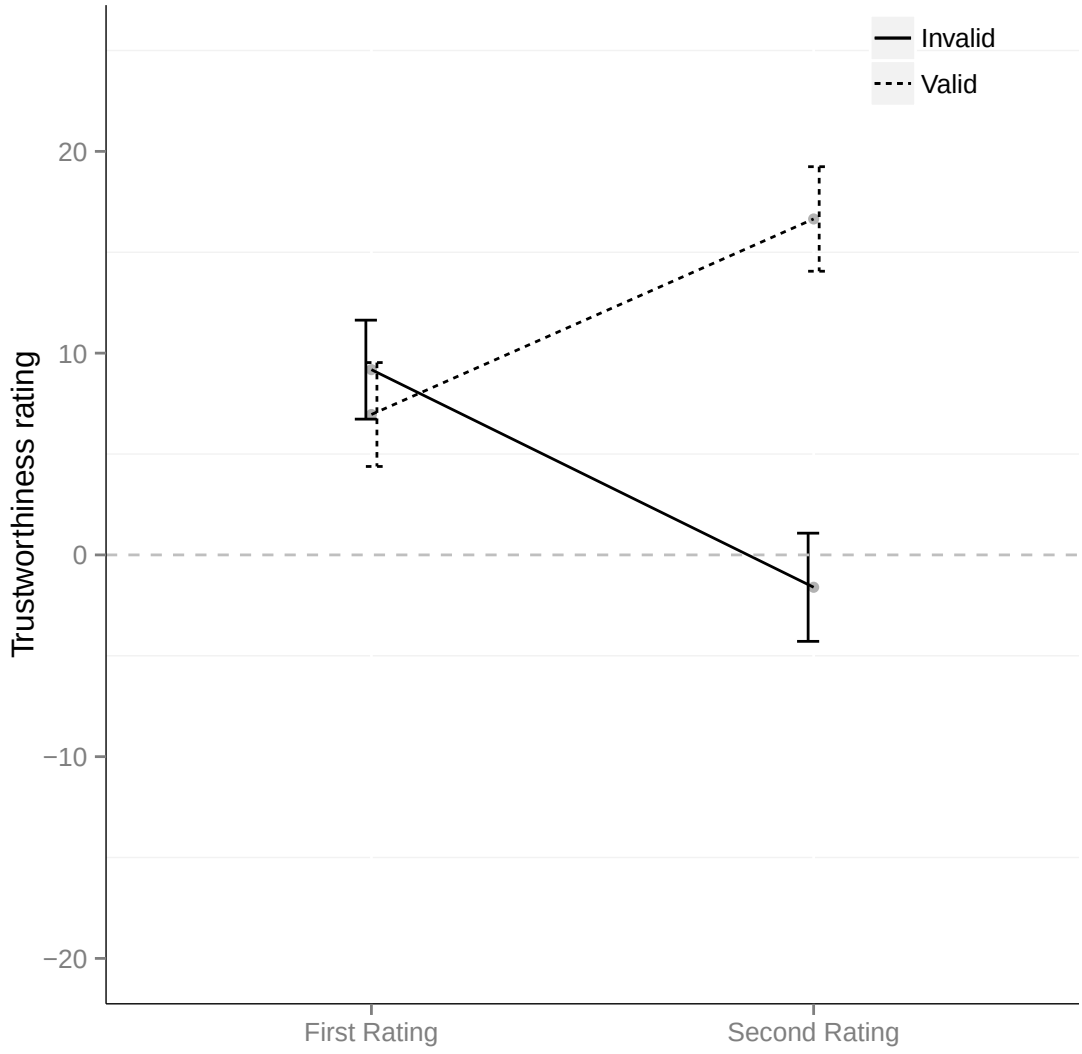

Figure 2.6: Time course of trustworthiness ratings over Experiment 2.2 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

were modelled but without an interaction (β = -21.17, SE = 4.05, $\chi^2(1)$ = 26.86, p <.001).

Smiling faces elicited a clearer trust effect than neutral faces, and the pattern of data is quite telling; when neutral faces were used, valid cueing faces did not change their trustworthiness over the course of the experiment, while invalid cueing faces showed a clear devaluation of trustworthiness. Further analysis of the changes in trustworthiness as a function of time for valid and invalid faces (for which no models would converge with any random terms) separately found that time improved the model fit for invalid

faces as in Experiment 2.1, and this was again significant (β = -11.15, SE = 2.89, $\chi^2(1)$ = 14.57, p <.001) but this time we also saw a significant improvement for valid faces (β = 10.02, SE = 2.72, $\chi^2(1)$ = 13.34, p <.001).

While it has been shown before that smiling faces elicit stronger trust effects than neutral faces (Bayliss et al., 2009), this is the first direct manipulation that can show differences in how valid and invalid faces are processed differently.

2.2.3 Cross-Experiment Analysis

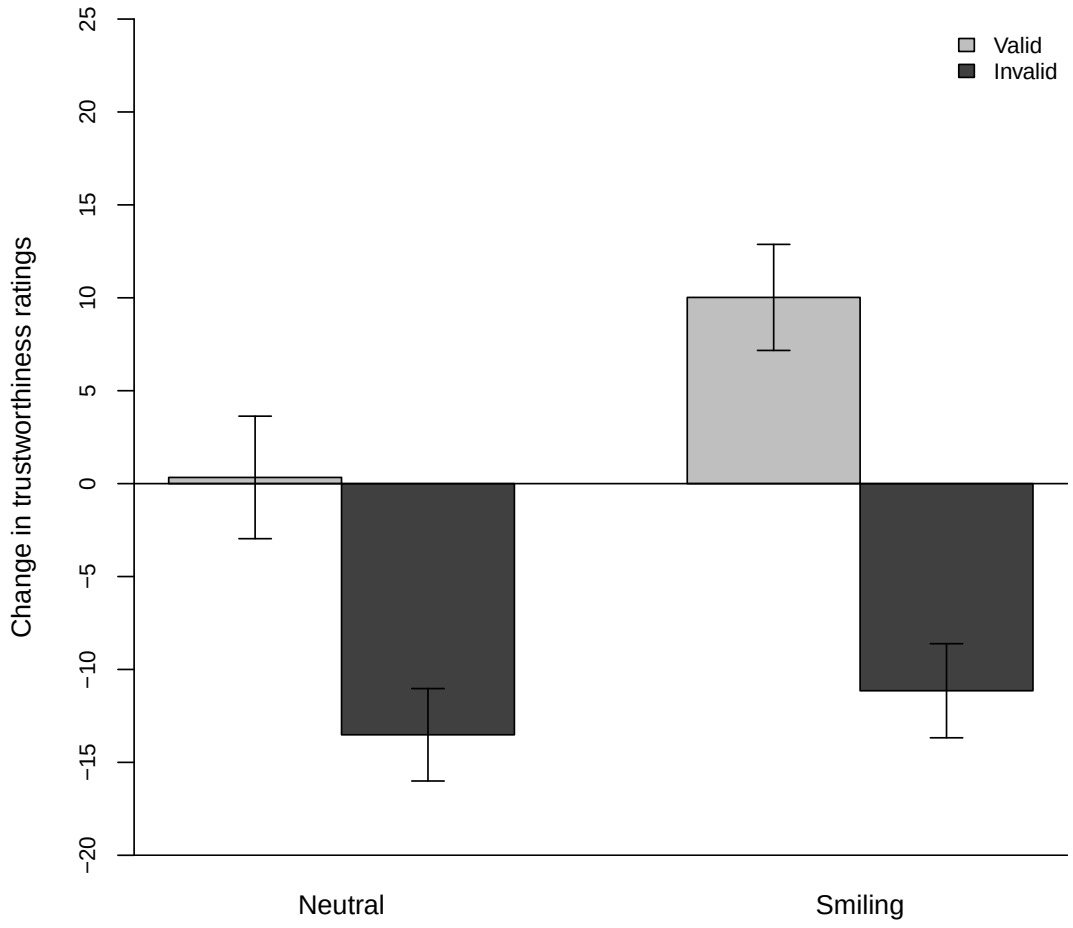

Figure 2.7: Changes in face ratings in Experiments 2.1 (left; neutral faces) and 2.2 (right; smiling faces) for valid (light grey) and invalid faces (solid line). Error bars show standard error.

The combined results of Experiments 2.1 and 2.2 are shown in Figure 2.7. As the only difference between the two experiments was expression of the face stimuli, the two changes in trustworthiness across both experiments were combined into a single dataset. In this analysis, the first trustworthiness ratings were subtracted from the second ratings to give a trustworthiness change score − that is, a value for the magnitude and direction of change over the experiment − and these change scores were compared using linear mixed effects models that included expression and validity as fixed factors. No models would converge with any slope terms included.

Adding validity to the null model significantly improved the model fit ($\beta = 17.51$, $SE = 2.45$, $\chi^2(1) = 49.43$, $p < .001$), which is not particularly surprising given that both experiments showed clear incidental trust learning. Including expression also significantly improved the fit ($\beta = 6.03$, $SE = 2.52$, $\chi^2(1) = 5.69$, $p = 0.017$), as smiling faces generally changed slightly more positively over the course of the experiment than did neutral faces. However, the interaction model (validity x expression) did not fit the data significantly better than the full model (validity + expression), where both factors were modelled but without an interaction ($\beta = 7.32$, $SE = 4.88$, $\chi^2(1) = 2.26$, $p = 0.133$).

Closer examination of the differences between valid and invalid faces (for valid faces all error terms had to be removed, while for models of invalid faces only the validity | subject term had to be removed) found that including expression in the model significantly improved the fit for valid faces ($\beta = 9.69$, $SE = 3.04$, $\chi^2(1) = 10.04$, $p = 0.002$), apparently driven by the fact that smiling valid faces were more likely to show an increase in trustworthiness over the experiment than their neutral counterparts. However, while this was not seen for invalid faces ($\beta = 2.37$, $SE = 4.42$, $\chi^2(1) = 0.30$, $p = 0.586$), there was no evidence that this effect resulted in a two-way interaction.

It is important to err on the side of caution when interpreting these between-experiment analyses, as these differ not only in terms of posed expression of the faces but also in terms of when the experiment was carried out, therefore the condition (or experiment) to which participants were assigned was not truly random. It could be that uncontrolled factors related to the timing of the data collection (e.g. time of year, amount of sunlight, ambient temperature, or proximity of university deadlines) might have affected participants' social learning in ways that we cannot determine. As such, results are presented but are meant to be indicative of interesting results.

2.3 Experiment 2.3

This experiment explores the possibility that incidental learning from gaze cues is not specific to trustworthiness per se but rather reflects a broader valence judgement associated with the faces. If this is true, then the effect should generalise to other judgements that have a moral or affective component. As such, Experiment 2.3 changes the question that participants are asked from one of trustworthiness to one of likeability.

2.3.1 Methods

Participants

27 participants volunteered for this study in return for a mixture of course credit and payment. Two participants' data were not collected due to runtime errors and a further one participant had to be removed following RT filters, so the final number available for analysis was 24 (24 female, $M_{age} = 18.96$, $s.d. = 1.06$).

Stimuli, Design and Procedure

This experiment was identical to Experiment 2.1 in every way except that at the beginning and the end of the experiment, participants were asked, ‘‘How LIKEABLE is this face?’’ rather than ‘‘How TRUSTWORTHY is this face?’’ and all mentions of trustworthiness on consent forms and instructions were replaced with the words likeable or likeability (dependent on context).

Data analysis

RT filters were applied in the same way as in Experiment 2.1, and in this Experiment one participant had to be removed for retaining less than 70% of their original trials. Data were analysed in the same way as outlined in Experiment 2.2. In this experiment, neither the RT or accuracy models would converge with any random terms defined, and so these were removed.

In this experiment, no models of likeability ratings would converge until the time | subject random term was removed.

2.3.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 2.3 are shown in Figure 2.8. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -32.04, SE = 8.85, $\chi^2(1)$ = 13.08, p <.001). This improvement was not seen for accuracy scores (β = 0.86, SE = 0.77, $\chi^2(1)$ = 1.23, p = 0.267).

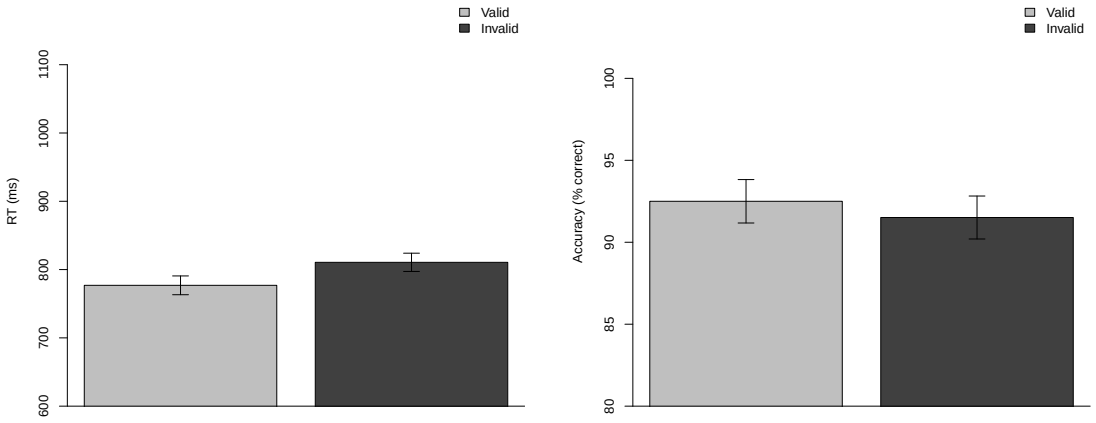

Figure 2.8: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 2.3 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

Likeability ratings

The changes in likeability ratings for the faces in Experiment 2.3 are shown in Figure 2.9. In this experiment, participants found invalid faces less likeable after the experiment than before and their valid counterparts slightly more, but this difference was quite small. Adding time to the null model did not significantly improve the model fit (β = -0.85, SE = 1.59, $\chi^2(1) = 0.29$, $p = 0.591$), nor did including validity (β = -0.46, SE = 1.60, $\chi^2(1) = 0.08$, $p = 0.771$), and the interaction model (time x validity) did not fit the data significantly better than the full model (time + validity), where both factors were modelled but without an interaction (β = -4.77, SE = 3.11, $\chi^2(1) = 2.36$, $p = 0.124$).

We compared the results of Experiment 2.3 with Experiment 2.1 by calculating trust change scores (the difference between the second and first ratings to determine how participants changed their decisions about trust across the experiment) and compared these with linear mixed effects models with validity and question (trustworthiness vs. likeability) as factors. Adding validity to the model significantly improved the fit (β = -9.31, SE = 2.06, $\chi^2(1) = 19.96$, $p < .001$), and including experiment marginally

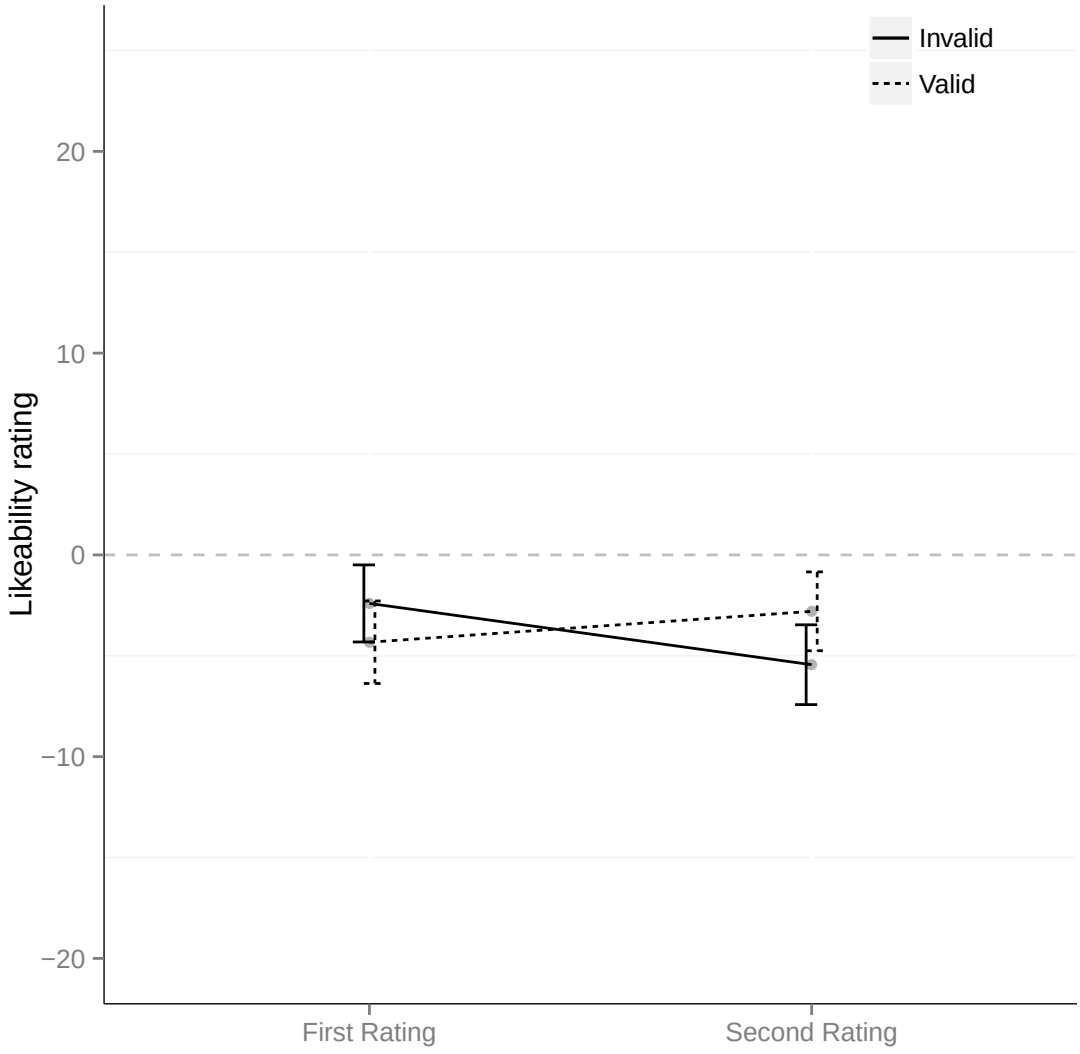

Figure 2.9: Time course of likeability ratings over Experiment 2.3 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

improved the fit (β = -4.77, SE = 42.80, $\chi^2(1) = 2.77$, $p = 0.096$), but the key finding was that an interaction model of validity x question fit significantly better than did a model with both factors without an interaction (β = -9.08, SE = 4.08, $\chi^2(1) = 4.94$, $p = 0.026$), which indicated that trust learning was significantly different from likeability learning. This is likely driven by the absence of any decrease in likeability for invalid faces, where this is seen when asking about trustworthiness.

2.4 Experiment 2.4

This experiment explores whether incidentally learned trustworthiness can be explained by demand characteristics, as participants could feasibly pick up on the nature of the experiment without actually experiencing changes in social attitudes. That is, they become explicitly aware of which faces did and did not repeatedly look towards targets, and used this explicit knowledge to make judgements on the trust scale. To explore this, Experiment 2.4 replaced the final trustworthiness rating with an explicit recollection check where participants were asked to categorise whether faces had previously looked towards (valid) or looked away from targets (invalid).

2.4.1 Methods

Participants

31 participants volunteered for this study in return for a mixture of course credit and payment. One participant's data were not collected due to a runtime error, so the total number available for analysis was 30 (M_{age} = 21.79, *s.d.* = 3.90).

Stimuli, Design and Procedure

This experiment was identical to Experiment 2.1 in every way up until the end of the gaze cueing. The final trustworthiness rating procedure was replaced with a 2AFC procedure where participants were asked to explicitly recall whether each face was valid or invalid. That is, after the gaze cueing task, an instruction screen appeared explaining that each face during the experiment had either always looked towards or away from where the object was about to appear. They were then told that the next procedure involved them having to recall whether each face had looked towards or away from the

object.

In a trial in the awareness check procedure, a face appeared in the centre of the screen with the question, ''Did this face look TOWARDS or AWAY from the object?'' and response key reminders on either side of the screen. Participants were instructed to press Z if they felt the face had looked towards where the object had been about to appear, M if they felt the face had looked away, and the SPACE bar if they could not remember. Each face was shown once in a randomised order, then the experiment ended.

Data analysis

RT filters were applied in the same way as in Experiment 2.1, and in this experiment one participant had to be removed for retaining less than 70% of their original trials. Mean RTs and percentage accuracy scores were calculated for each participant for both valid and invalid trials. RT and accuracy data were analysed in the same way as outlined in Experiment 2.2. In this experiment, RT models would not converge with any random terms defined and so these were removed.

For awareness results, participants' data was marked as incorrect if the participant chose the wrong cueing behaviour or if they pressed the SPACE bar to indicate that they did not know. As there were 16 faces, each participant could score a total number correct out of 16. Chance level (50% correct) was 8 out of 16, and binomial tests indicated that 12 was the threshold at which recall could be considered significantly above chance. As such, participants scoring 12/16 correct or above were considered aware of the manipulation and face cueing behaviour, while those scoring below this were considered naïve to the manipulation.

We categorised participants on this individual basis, but we also calculated the total

number of successes across all participants and tested whether these differed significantly from chance accuracy using a binomial test, to see if evidence of awareness emerged at the population level where it might not at the individual level.

2.4.2 Results and Discussion

Gaze-cueing

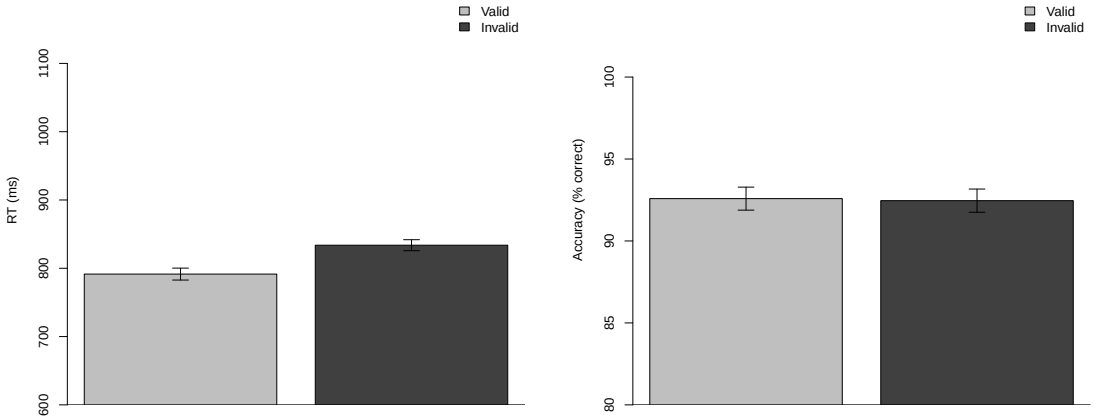

Figure 2.10: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 2.4 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

The RT and accuracy results of Experiment 2.4 are shown in Figure 2.10. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -43.06, SE = 8.26, $\chi^2(1)$ = 27.03, p <.001). This improvement was not seen for accuracy scores (β = 0.20, SE = 0.73, $\chi^2(1)$ = 0.07, p = 0.789).

Awareness check

The results of the awareness check are shown in Figure 2.11. Out of 30 participants, only 4 scored significantly above chance accuracy when asked to explicitly recall which faces were valid and which were invalid.

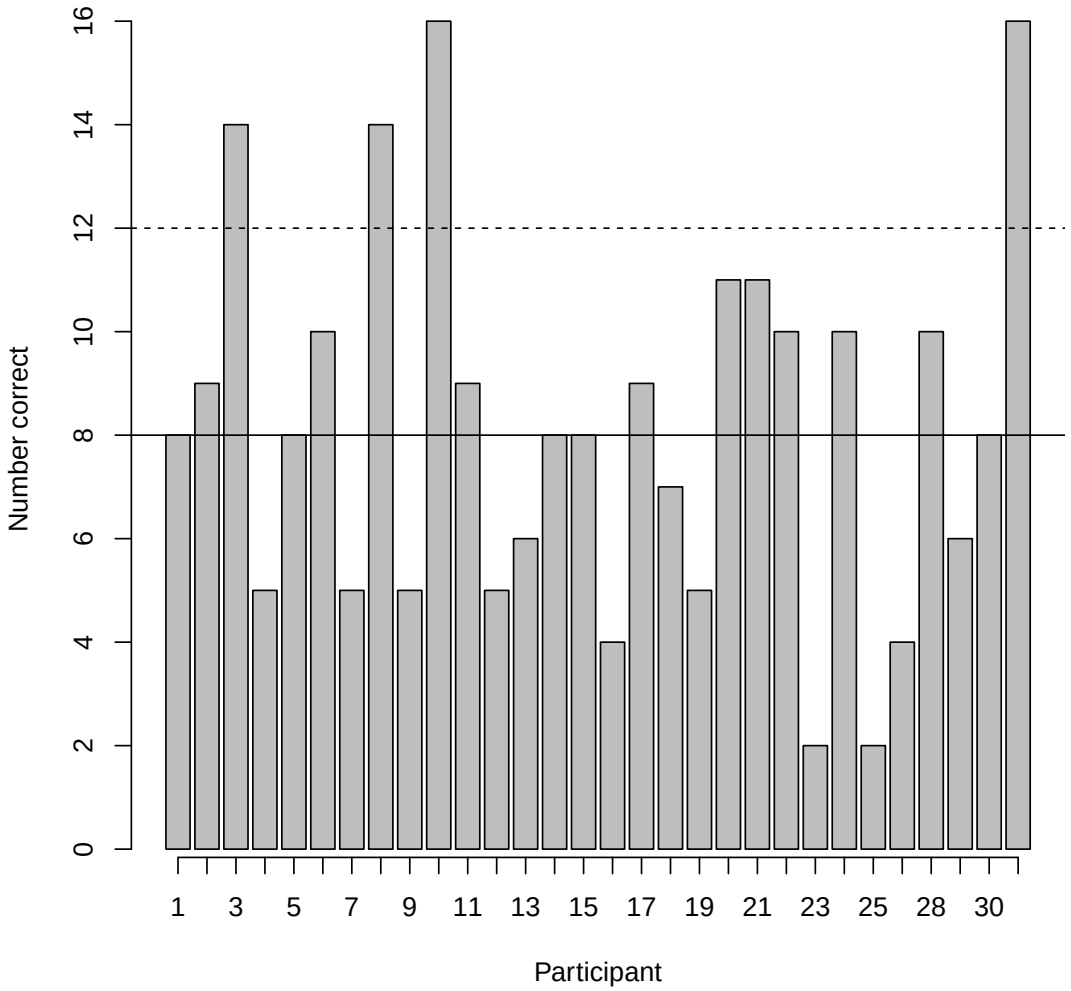

Figure 2.11: Total number of faces correctly identified for each participant out of 16 in Experiment 2.4. The solid horizontal line denotes the chance level of 50%. The dashed horizontal line designates the threshold above which performance was considered significantly above chance level.

Across all participants, average recall was close to 50% levels ($M = 8.17$, $s.d. = 3.69$), and the proportion of accurate recall was not significantly above chance ($p = 0.681$).

2.5 Chapter Discussion

There were several aims to this chapter. The first was to replicate previous research showing incidental trust learning from gaze cueing using a separate stimulus set from previous experiments and using a pre- and post-rating design that allowed for tracking

changes in trust over the experiment. Experiment 2.1 shows that with neutral faces, changes in trustworthiness ratings primarily manifest as a unidirectional decrease for invalid faces, while valid faces mostly remain unchanged.

Experiment 2.1 demonstrated that learning of trust is possible even when faces express neutral emotions. Previous work highlighted the role of emotion in these gaze-trust effects. Bayliss et al. (2009) demonstrated significant trust learning effects when faces smiled, no effects when they frowned, and marginal effects when the faces were neutral. Experiment 2.1 has revealed that significant trust effects can be obtained with neutral faces. It is unclear whether the failure to detect effects in the Bayliss et al., 2009 study was a Type II error, or whether the changes to the procedure were of critical importance. In the previous work a 2AFC task was employed where pairs of faces that had consistently looked towards targets (valid) or had looked away from targets (invalid), were presented and participants selected the one who they felt was more trustworthy. In contrast, the current study requires assessment of trust for each individual face and it measures changes in trust ratings from the start to the end of the experiment.

This new approach is a more sensitive and robust means of measuring trust. Furthermore it provides important information concerning where the effect may lie. That is, 2AFC can only reveal that faces that previously looked towards targets tend to be selected as more trustworthy, not whether valid faces are trusted more, invalid faces trusted less, or both. The results of Experiment 2.1 suggest an asymmetry, where the effect is only observed in the decline in trust of invalid faces that looked away from targets, whereas there is no change in trust rating for the valid faces that always looked towards targets.

Experiment 2.2 demonstrates that there appears to be a change in the pattern of

trust learning when the faces smile; that is, in contrast to Experiment 2.1 where effects were only detected in a decline in trust for invalid faces. When the faces smile a bi-directional effect is observed, where invalid faces again show a decrease in trust, while valid faces now produce a significant increase in trust. This latter bi-directional effect with smiling faces has also been demonstrated by Manssuer, Roberts and Tipper (2015), Manssuer, Pawling et al. (2015).

There are multiple potential explanations for the difference in the pattern of results between Experiments 2.1 and 2.2 that future research should investigate. For example, the default learning mechanism might be to detect deception. Certainly in terms of memory for faces, this is better for faces that deceive (Bayliss et al., 2009; Bayliss & Tipper, 2006; Bell, Buchner, Erdfelder et al., 2012; Buchner et al., 2009), hence learning of trust is only evident in invalid faces that deceive and look away from targets. In contrast, when the faces all express positive emotion, this combines with the positive signal of joint attention evoked by valid cueing faces, hence increasing trust of these faces. Alternatively, the positive social context elicited by a smiling expression motivates participants to try to remember the faces better − since invalid faces are already remembered well regardless of emotion (c.f. Experiment 2.1), this advantage would primarily affect valid faces. However, it is important to be cautious with this interpretation, as later results (particularly those of Experiment 3.1 in Chapter 3) may call this stance into question.

Experiment 2.3 replaced the question of trustworthiness that participants were asked with a question of likeability; simply by changing a single word in the design the effect was abolished. This lack of effect with liking judgements suggests that this incidental learning is highly specific to trust − a fact that makes sense if one considers

that trust as a trait judgement serves much more heavily as a predictive model of behaviour than does liking: we decide how much to trust someone based on how we expect them to behave, whereas liking is a more subjective and affective judgement, and one less based in statistical contingencies. For example, incidental learning of gaze contingencies will influence economic decisions to invest in another person (e.g. Rogers et al., 2014). It is possible that effects of liking might emerge if participants assimilated these deceptive cues in a more personal way − perhaps if we manipulated beliefs about intentions (as the knowledge that someone intentionally deceived you carries important implications for social behaviour) or a sense of competition with the face (such that the face is deceiving you to maximise their own chances at a reward) we would see a change in how participants like the faces, but at its basic level this effect appears to be specific to monitoring the trustworthiness of interactants.

The final aim of this chapter was to explore whether the inclusion of trust ratings at the beginning might influence the explicit awareness of gaze behaviour. Note that in previous work that showed no awareness of gaze cueing behaviour (Bayliss et al., 2006; Rogers et al., 2014) participants were not asked about trust until the end of the experiment. Hence up to that trust measure at the end of the experiment the faces had been irrelevant and to-be-ignored. However, in the current task the faces are rated for trustworthiness at the start of the experiment. It is possible that this trust rating could cue the participants to the relevance of the faces, and facilitate learning of gaze behaviour, resulting in explicit/conscious knowledge of which faces consistently looked towards and away from targets.

It should be noted that such an explanation based on demand characteristics was unlikely given that Experiment 2.3 yielded null results despite that experiment arguably

being as easy to interpret and predict as Experiments 2.1 and 2.2. For peace of mind, however, Experiment 2.4 explored this question and found that explicit memory for the faces' cueing behaviour could not explain this effect, as only 4 participants out of 30 scored significantly above chance when asked to explicitly categorise the faces as valid or invalid.

This distinction between explicit memory for cueing behaviour in Experiment 2.4 and more implicit memory in the form of trustworthiness ratings in Experiments 2.1 and 2.2 offers an intriguing insight into the nature of these learned representations. Indeed, it suggests that these effects occur outside of conscious awareness, and in theory it is possible that incidental trust learning might *only* occur outside of conscious awareness. While it is not possible to examine this using the current data, it is somewhat telling that the proportion of participants who demonstrated explicit awareness (4/30) is approximately similar to the proportion of participants in Experiment 2.1 who did not show the typical trust learning effect (changes to valid faces more positive than changes to invalid: 7 out of 24). This is not conclusive, but it is suggestive and may be an avenue for future research to investigate whether these two populations are similar.

The aim of this chapter was to explore some of the key features and boundaries of incidental trust learning from gaze cues. We demonstrated that using dual trustworthiness rating scales at either end of the experiment was a more sensitive measure of trust than 2AFC used in previous experiments, as we found a clear pattern of trust learning with neutral faces (Experiment 2.1), and that this pattern changed when faces smiled (Experiment 2.2), a finding that would be impossible with more blunt methods. We also found that this effect is specific to trustworthiness and does not reflect a general valence impression of the face (Experiment 2.3) and occurs outside of explicit,

conscious awareness (Experiment 2.4).

Chapter 3. Examining the durability of incidentally learned trust

Incidental learning of trust from gaze cues has been replicated in several different experiments, including now those in Chapter 2. However despite these replications of the original effect, one question that has never been explored directly concerns the stability of this learned representation. That is, no studies to date have examined how long incidentally learned trustworthiness lasts, despite the fact that this question carries important implications for interpreting this effect. If this incidental trust learning is short-lived and easily disrupted by an intervening task, then this suggests that it reflects an active, online monitoring of in-the-moment statistical contingencies, concerned only with short-term interactions. If, on the other hand, this effect can survive interference, it suggests a mechanism that is actively feeding such short-term monitoring into durable, long-term representations of interaction partners.

Consider the Haxby et al. (2000) model of face processing, and the tentative model of incidental trust learning posited in Chapter 1, which proposes two separate streams of information when viewing faces; one through the fusiform gyrus and anterior temporal lobe that appears to encode invariant stable features of the faces such as identity and physical appearance, and a second that projects more dorsally through the superior temporal sulcus (STS) and processes more variant aspects of the face such as expression and gaze direction. That there is some communication between these two streams is evident from our previous research. That is, the specific property of face identity is associated with particular patterns of gaze behaviour, resulting in changes of face trustworthiness.

Trustworthiness can be considered a stable property of person identity, and hence one might predict it will be stable over time. That is, once a person is encoded as less

trustworthy, such information should be available for future encounters with that person. However, note that the learning of the association between face identity and patterns of gaze takes place while the face is irrelevant and ignored while participants undertake a different task. Previous work has shown little awareness of this learning (Bayliss et al., 2006; Rogers et al., 2014), and in Experiment 2.4 participants could not explicitly recall whether particular face identities had always looked towards or away from targets. Hence the lack of explicit awareness of the face-gaze relationship might reflect weak and transient memories.

With the aim of exploring these questions, Experiment 3.1 offers a direct replication of Experiment 2.1 and replicates the original effect. Then, in Experiment 3.2, we recreate the baseline experiment but add in a short interference task where participants watched videos of reaching motions in an attempt to distract them from their memories of the faces in the experiment.

To further test the possibility for stable representations of trust, we can also manipulate participants' initial experience with the faces. As noted, incidental learning during gaze cueing relies not only on attention orienting evoked by eye gaze but also on trial-invariant aspects of the face such as identity recognition. When learning about the face identities from their gaze cueing behaviour, these distributed systems must share information such that untrustworthy behaviour can be linked to the individual expressing it. In previous studies, the faces used have been unfamiliar, and so the identity representations that serve as the anchors for these trust representations are less stable, which may compromise the strength of these associations. More familiar faces have been shown to have more stable neural representations (Eger, Schweinberger, Dolan & Henson, 2005) and that these neural representations are related to better behavioural

performance in face identification tasks (Weibert & Andrews, 2015). As such, we propose that the association between the face identity and patterns of gaze behaviour incidentally learned while ignoring the face will be facilitated if the face representation is more familiar. Experiment 3.3 explores whether using more familiar faces will have an effect on memory for incidentally learned trust by replicating Experiment 3.2 and including a short face-matching familiarisation task at the beginning. We predict that increased familiarity of the face stimuli will produce more durable memories.

The motivation of Experiment 3.4 is to extend the interference to see if the learning of trust can endure in the longer term. In this experiment we replace the filler task in Experiment 3.3 and introduce an hour-long gap where participants are sent away from the lab between the gaze cueing and the final trustworthiness rating portions of the experiment. We explore whether this effect can last over the course of an hour of real-world interference (i.e. real faces and interactions, changes in context and environment, etc.)

The final aim of this chapter is to see if learning can survive exposure to counter-typical behaviours. That is, as these representations are driven by participants learning that certain faces provide valid cues and others invalid cues, are these representations strong enough to withstand instances where the faces demonstrate the opposite behaviour (i.e. valid becomes invalid; invalid becomes valid). In Experiment 3.5 we replace the interference tasks used before with a sixth block of gaze cueing where faces switch their gaze behaviour. As such, participants experience two instances where the face's behaviour does not match the stored representation of that identity. We examine whether trust learning can survive this interference.

This also offers an opportunity to examine the effect in a different way. The crux of

this incidental learning is that participants pick up on the underlying patterns of gaze behaviour and use these to inform subsequent trustworthiness judgements. While there is little evidence that participants can adapt to these invalid cues and learn to inhibit gaze following for invalid faces (c.f. Manssuer, Roberts & Tipper, 2015; Manssuer, Pawling et al., 2015, and Appendix A), it is nonetheless possible that if this pattern were disrupted that participants would experience an error signal and show a cost in RTs, accuracy, or both. A popular measure of implicit learning of statistical patterns in presented stimuli is to change the pattern after a certain training block. Although participants do not necessarily report explicit knowledge of the pattern, (and may not show a reduction in associated RT costs; Heuer, Schmidtke & Kleinsorge, 2001), they nonetheless show a broad increase in RTs when the underlying pattern changes (Destrebecqz & Cleeremans, 2001). As such, changing the gaze cueing behaviour of faces could yield an alternative measure of this incidental learning of gaze contingencies.

However, the key issue of this chapter is the stability of these learned representations. A related issue that could influence the stability of memory is how the learning may be affected by the nature of the information to be remembered. In Chapter 2 we showed that a decrease in trust for invalid faces was a more robust result than was an increase in trust for valid faces, and this supports previous research demonstrating memory advantages for cheaters over co-operators (Bell, Buchner, Erdfelder et al., 2012; Buchner et al., 2009). As an example, Bayliss and Tipper (2006) showed that after incidental learning of trust via patterns of eye-gaze, participants subsequently reported that the invalidly cueing low-trust faces had been presented more often. Hence we might expect to observe more stable memory for invalid faces in the current studies.

3.1 Experiment 3.1

This experiment is a direct replication of the baseline experiment detailed in Experiment 2.1.

3.1.1 Methods

Participants

32 participants volunteered for this study. One participant made entirely inaccurate responses in the first block of trials, suggesting that they had misunderstood the instructions, and another participant was removed after RT filters were applied for retaining <70% of their original trials, and so the final number of participants included in analysis was 30 (21 female; $M_{age} = 20.09$, $s.d. = 1.81$).

Stimuli, Design and Procedure

This experiment was identical to Experiment 2.1. However, as this chapter deals with pushing the limits of the incidental trust learning effect, we wanted to ensure that any null results could not be attributed to lack of power and would actually reflect a disruption of memory. For this reason, we increased power by raising the n of all experiments in this chapter to 30 instead of 24.

Data analysis

As in Experiment 2.1, before data were analysed participants' responses were filtered to remove all error trials (where participants reported the incorrect answer) and RT outliers – RTs below 250ms (too short to process the stimuli) and above 2,500ms (indicating that participants had not given a response in the allotted time). The number of

remaining trials was then compared with the original number of trials to check that all participants retained at least 70% of their total trials and had not scored below 70% total correct on any one condition.

Data were analysed in the same way as described in Experiment 2.1 Data Analysis section, but collapsed across block to look only for a cost associated with gaze cueing. As such, each analysis (both RTs and accuracy rates) compared a null model with a validity-only model. Neither RT nor accuracy models would converge with validity | subject as a defined term.

When modelling trustworthiness ratings for this and all subsequent experiments, this maximum random structure (which we hereafter term the null model), would not converge when all repeated-measures factors were included, and so we removed the time | identity term from all models.

As in Chapter 2, Appendix A provides the results of more traditional factorial ANOVAs for the purposes of comparison.

3.1.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 3.2 are shown in Figure 3.1. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs, which indicates a gaze cueing effect (β = 46.41, SE = 7.28, $\chi^2(1)$ = 20.89, p <.001). This effect was not seen for accuracy scores (β = -0.03, SE = 0.09, $\chi^2(1)$ = 0.13, p = 0.721).

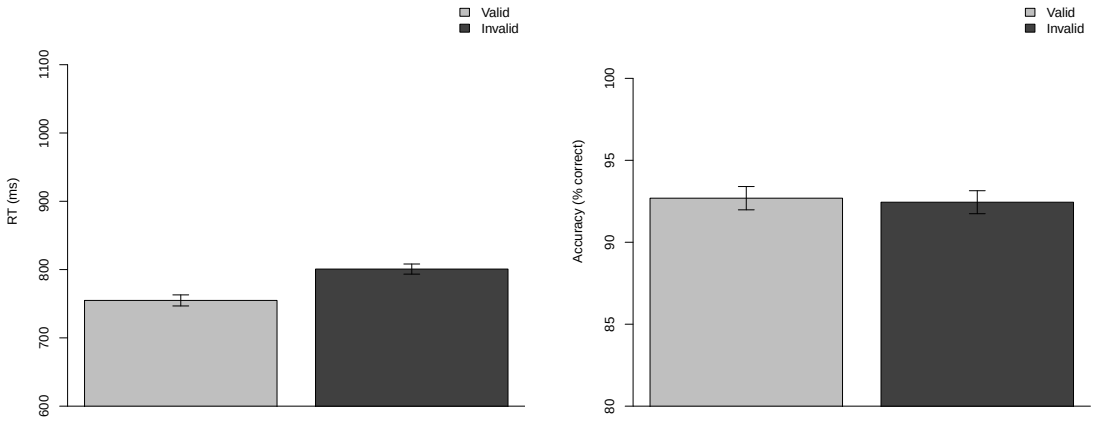

Figure 3.1: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 3.1 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

Trustworthiness ratings

The trustworthiness ratings from Experiment 3.1 are shown in Figure 3.2. Adding time to the null model did not significantly improve the model fit (β = -2.06, SE = 1.95, $\chi^2(1) = 1.13$, $p = 0.288$), but including validity did (β = -6.77, SE = 1.78, $\chi^2(1) = 12.96$, $p <.001$). Finally, the interaction model (time x validity) fit the data significantly better than did the full model (time + validity), where both factors were modelled but without an interaction (β = -11.34, SE = 3.27, $\chi^2(1) = 11.99$, $p <.001$).

We ran further analysis of the changes in trustworthiness as a function of time for valid and invalid faces separately. These models would not converge with the time | identity term included, so we removed this and found that time significantly improved the model fit for invalid faces (β = -7.99, SE = 2.35, $\chi^2(1) = 11.47$, $p <.001$) but this improvement was only marginal for valid faces ($\beta = 4.06$, SE = 2.32, $\chi^2(1) = 3.06$, $p = 0.080$), similar to Experiment 2.1.

The results of Experiment 3.1 replicate those of previous studies (Bayliss et al., 2006; Bayliss et al., 2009; Manssuer, Roberts & Tipper, 2015; Manssuer, Pawling et al., 2015).

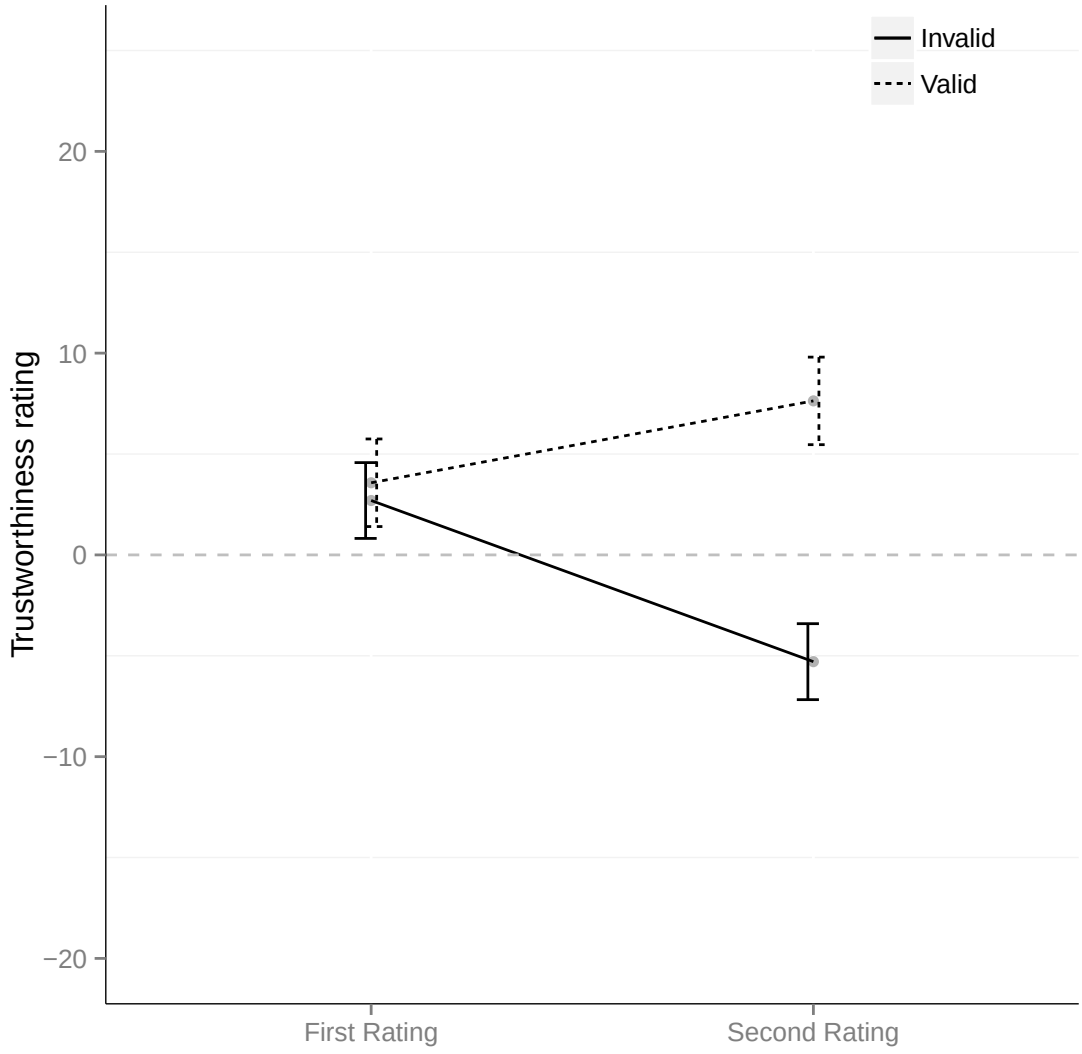

Figure 3.2: Time course of trustworthiness ratings over Experiment 3.1 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

Observing somebody's gaze movements automatically triggers a shift in attention to the same location and results in faster identification of objects at that location as compared to objects not gazed at. The association between direction of gaze (valid or invalid) and face identity appears to be encoded: even though the face was irrelevant to the task, individuals who looked away from the target object (invalid cues) were trusted less.

Having shown evidence of incidental trust learning in Experiment 3.1, we now move on to explore the key question of this chapter, which is how long this effect can survive a period of interference. In Experiment 3.2, we introduce a brief distraction task between

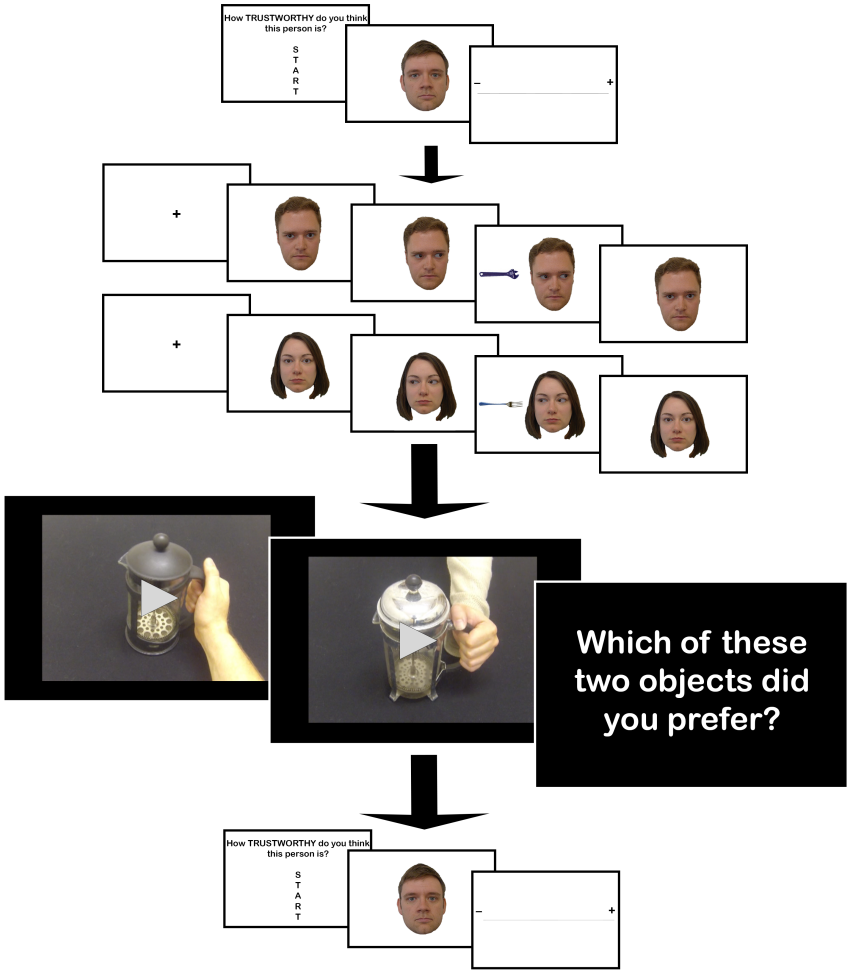

Figure 3.3: Schematic of the paradigm of Experiment 3.2, with the addition of the interference paradigm between the gaze-cueing and final trustworthiness ratings. This same interference task was used in Experiment 3.3.

the final block of gaze cueing and the final trustworthiness ratings.

3.2 Experiment 3.2

This experiment replicates the baseline experiment (Experiment 3.1) with an intervening filler task to see if the effect survives interference.

3.2.1 Methods

Participants

30 participants (21 female, $M_{age} = 20.63$, $s.d. = 1.08$) volunteered for this study.

Stimuli, Design and Procedure

This experiment was identical to Experiment 3.1 in every way except that participants performed a filler task between the gaze-cueing procedure ending and the final trustworthiness ratings (see Figure 3.3). All other details were identical. See Appendices B and C for methods and results of the main filler task. The filler task included no faces and lasted approximately 05:45 (minutes:seconds).

Data analysis

RT filters were applied in the same way as in Experiment 3.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials. All accuracy models converged when the validity | subject error term was removed. RT models converged with the maximum random structure.

In this experiment, the validity-only model would not converge with the defined random structure, and so we removed the time | subject slope from all of the models to help direct comparison.

3.2.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 3.2 are shown in Figure 3.4. RTs were faster to valid trials (M = 837.49, $s.d.$ = 159.39) than to invalid trials (M = 863.72, $s.d.$ = 155.01), and fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = 25.49, SE = 9.36, $\chi^2(1)$ = 7.41, p = 0.006). This improvement was not seen for accuracy scores (β = -0.12, SE = 0.70, $\chi^2(1)$ = 0.03, p = 0.864).

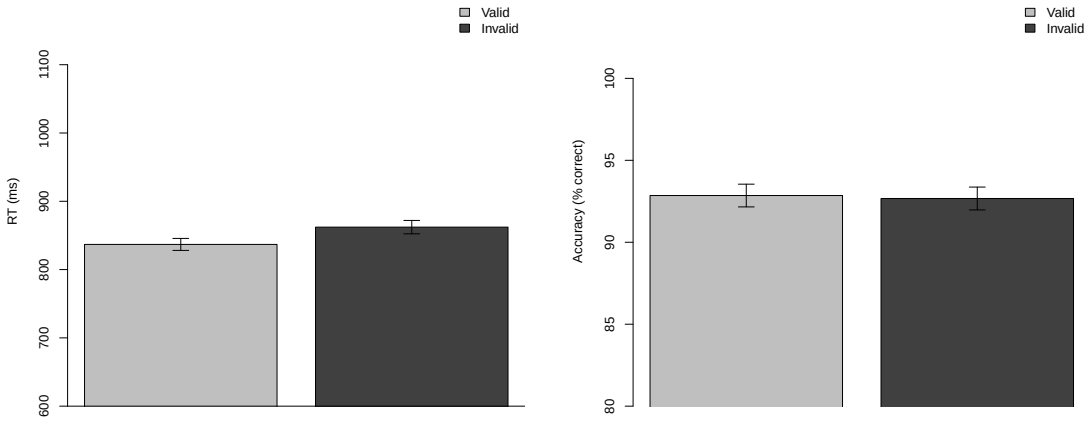

Figure 3.4: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 3.2 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

Trustworthiness ratings

The changes in trustworthiness ratings for the faces in Experiment 3.2 are shown in Figure 3.5. Adding time to the null model significantly improved the fit (β = -3.66, SE = 1.68, $\chi^2(1) = 4.74$, $p = 0.029$), as did including validity (β = -5.88, SE = 1.69, $\chi^2(1)$ = 10.29, $p = 0.001$). Finally, the interaction model (time x validity) fit the data better than did the full model (time + validity) and this approached significance (β = -6.12, SE = 3.35, $\chi^2(1) = 3.35$, $p = 0.067$).

We ran further analysis of the changes in trustworthiness as a function of time for valid and invalid faces separately. These models would not converge with the time | identity term included, so we removed this and found that time significantly improved the model fit for invalid faces (β = -6.72, SE = 2.39, $\chi^2(1) = 7.77$, $p = 0.005$) but this improvement was not seen for valid faces (β = -0.60, SE = 3.06, $\chi^2(1) = 0.04$, p = 0.839).

Experiment 3.2 aimed to explore whether a brief period of interference could disrupt the pattern of trust learning observed in Experiment 3.1. While the effect was not as

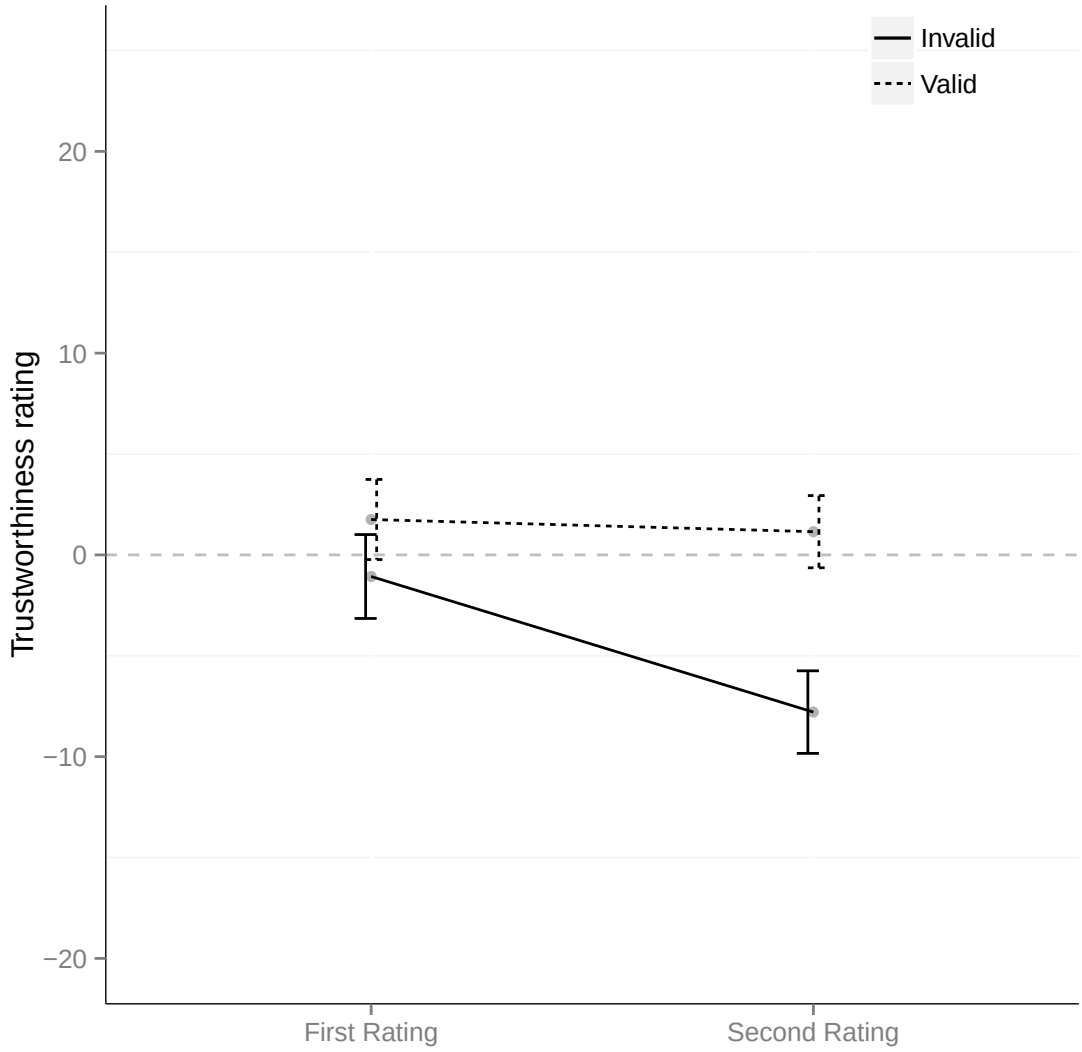

Figure 3.5: Time course of trustworthiness ratings over Experiment 3.2 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

clear as seen in previous experiments, the a priori predicted pattern of trust changes was observed.

There are trends for susceptibility to decay when an interference task is introduced. One question, then, is what is the cause of this susceptibility to decay. We propose that familiarity with the faces used as stimuli may be the deciding factor − in Experiment 3.2, the only exposure participants have to the faces is the initial pre-experiment ratings, which provides only a superficial opportunity to encode these identities. This means that when participants experience helpful or misleading gaze cues, they have to build

representations of these identities from the bottom up. It may be that this effect is more durable when cueing validity information is added to a pre-existing representation of these identities − that is, when the faces are more familiar.

In Experiment 3.3, we explore this question by including an additional task at the beginning of the experiment designed to increase participants' familiarity with the faces − we use the same faces as in Experiments 3.1 and 3.2 to avoid any confounds of different stimuli, and to avoid any pre-existing expectations of trustworthiness that might arise with using naturally familiar stimuli such as famous faces.

3.3 Experiment 3.3

Experiment 3.3 replicates Experiment 3.2 but adds an additional familiarisation task to the beginning of the procedure to trigger greater familiarity with the face stimuli used. That is, it employs procedures developed by Andrews, Jenkins, Cursiter and Burton (2015) where faces are viewed from different angles and express different emotions in a same-different face matching task. Such encoding has been shown to significantly improve face recognition performance.

3.3.1 Methods

Participants

32 participants volunteered for this study, but due to runtime errors data from two participants were not collected. This left 30 participants (21 female, M_{age} = 20.17, *s.d.* = 2.10) who received a mixture of course credit and payment.

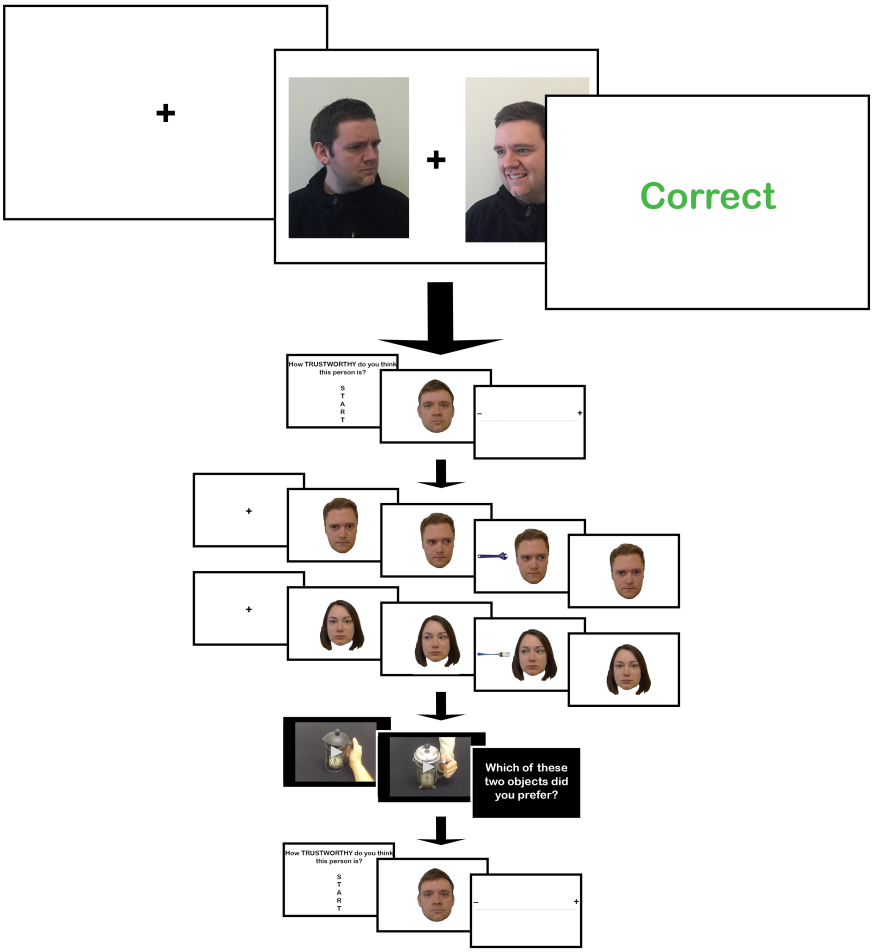

Figure 3.6: Schematic of the familiarisation task participants completed at the very beginning of Experiments 3.3 – they were shown two images of faces and asked to judge if they were the same or different identities. The paradigm was the same for Experiment 3.4 except that the 2AFC object preference interference task introduced in Experiment 3.2 was replaced with an hour away from the lab. Feedback was provided for incorrect responses.

Stimuli, Design and Procedure

This experiment was identical to Experiment 3.2 in every way except that participants also performed a face-matching task at the beginning of the experiment in order to allow for greater familiarity with the KDEF faces. Participants were shown images of all sixteen identities that varied in their head orientation (full-left, half-left, half-right or full-right) and emotion (happy, angry, disgusted, surprise, afraid or sad) – these were unaltered images from the KDEF stimulus set (Lundqvist et al., 1998) and so were presented with an off-white/brown background, rather than the plain white background that was used for the images in the gaze-cueing and trust-rating portions. Participants

made same/different judgements of the identities of face pairs, responding with a button press of S if the two images showed the same person, and D if they showed different people. Image pairs showed the same identity on 25% of trials, and a written feedback screen appeared after each trial reporting either ‘‘Correct”, ‘‘Incorrect” or ‘‘No response detected”, depending on what response was logged. During the course of a trial a fixation cross was presented for 500ms, followed by the two images either side of a fixation for 1,500ms, followed by the feedback screen for 1,000ms (see Figure 3.6).

It was expected that the variability in these images and the nature of the identity judgement task would prompt participants to encode viewpoint- and emotion-independent identity representations of the individuals. Such a task has been used before to good effect (Andrews et al., 2015), as participants reconcile within-identity variability in face photographs to develop a richer, more abstract representation of the individual.

Data analysis

RT filters were applied in the same way as in Experiment 3.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials. In this experiment, the null model for RTs would not converge with the validity | subject random slope term defined, so this was removed, but all other models converged with this included.

For trustworthiness ratings, the validity-only, time + validity, and time x validity models would not converge until the time | subject and time | identity error slope terms were removed, and so we removed these from all models to allow for direct comparison.

3.3.2 Results and Discussion

Gaze-cueing

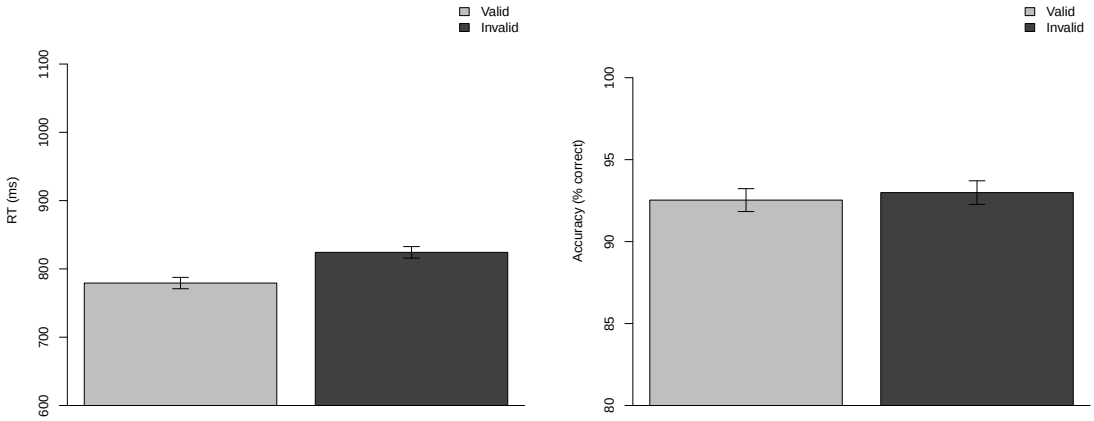

Figure 3.7: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 3.3 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

The RT and accuracy results of Experiment 3.3 are shown in Figure 3.7. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = 45.26, SE = 8.44, $\chi^2(1)$ = 28.61, p <.001). This improvement was not seen for accuracy scores (β = 0.45, SE = 0.72, $\chi^2(1)$ = 0.40, p = 0.530).

Trustworthiness ratings

The changes in trustworthiness ratings for the faces in Experiment 3.3 are shown in Figure 3.8. Adding time to the null model did not make it fit the data significantly better (β = 1.85, SE = 1.72, $\chi^2(1)$ = 1.17, p = 0.280), but it did fit significantly better when validity was included (β = -6.97, SE = 1.70, $\chi^2(1)$ = 16.42, p <.001). Finally, the interaction model fit the data significantly better than did the full model, where both factors were modelled but without an interaction (β = -13.88, SE = 3.38, $\chi^2(1)$ = 16.78, p <.001).

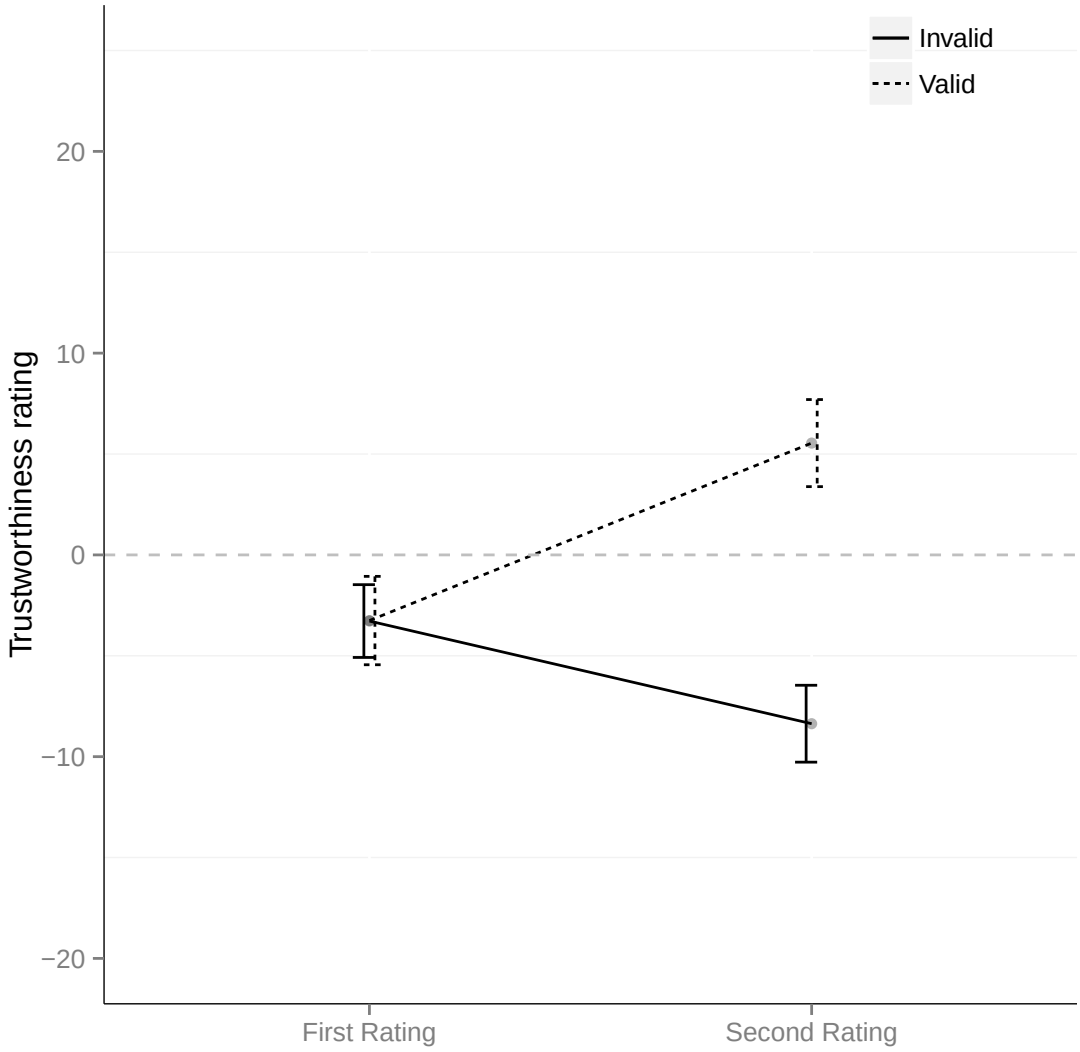

Figure 3.8: Time course of trustworthiness ratings over Experiment 3.3 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

Experiment 3.3 explored whether initial familiarity with the face identities used in the gaze cueing affected how well participants learned and retained information about individuals' trustworthiness, and it appears that familiar identities lead to durable trustworthiness representations given the clear pattern of changes in ratings that we found. Further analysis of the changes in trustworthiness as a function of time for valid and invalid faces found that time only marginally improved the fit when applied to ratings of invalid faces (β = -5.09, SE = 2.67, $\chi^2(1) = 3.47$, $p = 0.063$), but there was now a significant improvement for valid faces ($\beta = 8.80$, $SE = 2.39$, $\chi^2(1) = 11.43$, p

<.001). This is an interesting point to note in light of the results of Experiments 2.1 and 3.1, which found a significant decrease in trust for invalid faces but no significant change for valid faces. It is possible that this increase for valid faces is a result of the increased familiarity − perhaps with more familiar identities where there is a pre-existing representation upon which to build, the focus shifts from detecting deception to monitoring prosocial, cooperative behaviour. However, the current experiment can only hazard this tentatively and so this could be an avenue for future research.

These results suggest that, while with unfamiliar faces this effect appears to be somewhat susceptible to interference, increased familiarity with the faces can make this incidental learning more resilient to decay over a short period of interference. This of course raises the logical question: if this effect can now survive a brief (around 5 minute) period of interference, could it survive longer periods? To explore this, in Experiment 3.4 we replace the 5-minute filler interference task with an hour-long break during which participants were sent away from the laboratory.

3.4 Experiment 3.4

In Experiment 3.4 we replaced the 5-minute filler interference task with an hour-long break during which participants were sent away from the laboratory. This provided a naturalistic interference, as participants were given no instructions about what to do during that time, and so means that if we still see evidence of the effect after this time that this gaze cueing manipulation can lead to particularly durable changes in trustworthiness.

3.4.1 Methods

Participants

30 participants (24 female, M_{age} = 19.93, *s.d.* = 2.93) volunteered for this study.

Stimuli, Design and Procedure

This experiment was identical to Experiment 3.3 in every way except that participants performed no filler task. Instead, participants left the testing room for one hour between the gaze-cueing and the final trustworthiness rating sections.

Data analysis

RT filters were applied in the same way as in Experiment 3.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials. In this experiment, all RT models converged. The accuracy null model would not converge until validity | subject term was removed, and so this was removed from both models.

For trustworthiness ratings, the validity-only, time + validity, and time x validity models would not converge with the defined random structure, and so we removed the time | identity slope from all models to allow for direct comparison.

3.4.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 3.4 are shown in Figure 3.9. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = 27.74, SE = 8.63, $\chi^2(1)$ = 8.63, p = 0.003). This improvement was not seen for accuracy scores (β = 0.32, SE = 0.74, $\chi^2(1)$ = 0.19, p = 0.665).

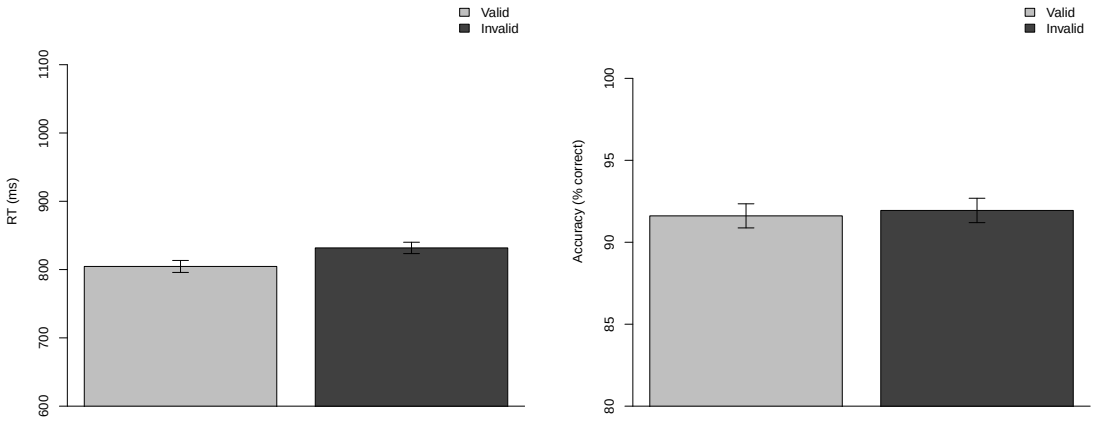

Figure 3.9: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 3.4 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

Trustworthiness ratings

The changes in trustworthiness ratings for the faces in Experiment 3.4 are shown in Figure 3.10. Adding time to the null model did not make it fit the data significantly better (β = -1.59, SE = 1.63, $\chi^2(1) = 0.95$, $p = 0.329$), nor did it predict the data significantly better when validity was included (β = -2.86, SE = 1.87, $\chi^2(1) = 2.34$, $p = 0.126$). However, the interaction model fit the data significantly better than did the full model, where both factors were modelled but without an interaction (β = -6.77, SE = 3.23, $\chi^2(1) = 4.40$, $p = 0.036$).

We ran further analysis of the changes in trustworthiness as a function of time for valid and invalid faces separately, but in this experiment including time as a fixed factor did not significantly improve the model fit for valid faces ($\beta = 1.79$, $SE = 2.29$, $\chi^2(1) = 0.61$, $p = 0.433$) but it did for invalid faces (β = -4.97, $SE = 2.21$, $\chi^2(1) = 4.63$, $p = 0.031$).

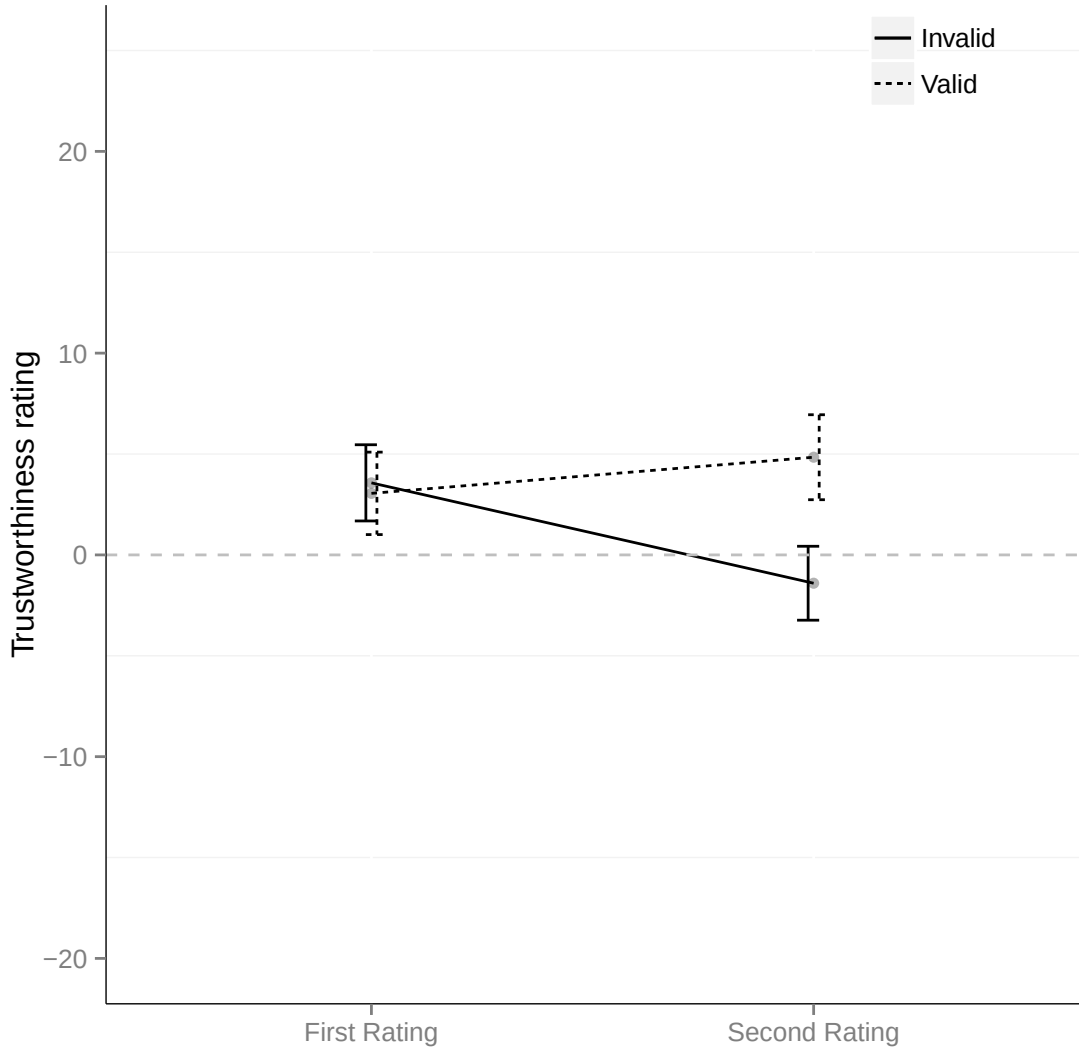

Figure 3.10: Time course of trustworthiness ratings over Experiment 3.4 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

3.5 Experiment 3.5

In Experiment 3.5 we replace the interference used in previous experiments with a sixth block of gaze cueing. In this additional block, we reversed the cueing behaviours of the faces such that previously valid faces became invalid and vice versa. The aim here was twofold: firstly, to see whether learned representations of trust can survive exposure to counter-typical gaze behaviours (i.e. whether trust learning was extinguished as the learned information, "Person A is always valid, Person B is always invalid", is no longer true).

The second aim was to investigate whether trust learning could be explored using RTs. Given that participants completed the typical five blocks of gaze cueing before they reached the reversed sixth block, they would have learned the same information that they gathered throughout previous experiments by the time they reached the final block. If participants are implicitly learning the patterns of gaze cueing they might then show a cost in RTs as the underlying pattern of gaze behaviour changes. The logic of this design is based on the sequence learning paradigm used by Knopman and Nissen (1991) and Reed and Johnson (1994), where changes from an implicitly learned sequence impairs response times.

3.5.1 Methods

Participants

34 participants volunteered for this study, but one participant's data were not collected due to a runtime error and three were removed following RT filters, which left 30 for analysis ($M_{age} = 19.93$, $s.d. = 2.93$).

Stimuli, Design and Procedure

This experiment was identical to Experiment 3.1 with the addition of an extra block of gaze cueing. As such, this experiment did not include the familiarisation task in Experiments 3.3 and 3.4. During this additional block, the cueing behaviour of the faces was reversed, such that previously valid faces now provided invalid cues, while previously invalid faces now provided valid cues.

Trustworthiness ratings were collected at the beginning and the end of the experiment to explore whether learning could survive exposure to inconsistent

information (i.e. if the fact that faces changed their cueing behaviour for the last two appearances could override the learned trustworthiness impressions from earlier in the experiment).

Data analysis

RT filters were applied in the same way as in Experiment 3.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials.

Given that there was a possibility that the change in the final additional block would lead to an increase in reaction time, we analysed cueing data in the same way as in Experiment 2.1 − that is, broken down by block as well as validity. Neither RT nor accuracy models would not converge until the validity | subject term was removed. We also investigated the effect of changing the gaze behaviour on cueing effects further by examining only the last two blocks − block 5, where the pattern has been learned fully, and block 6 where it changes. We also had to remove the validity | subject term from these models before they would converge.

No models of trustworthiness ratings would converge with the maximum random structure, and so we removed the time | subject and validity | subject slopes from all models.

3.5.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 3.5 are shown in Figure 3.11. RTs were aggregated across subject, identity, block, and validity, and these were analysed using linear mixed-effects modelling. Adding block as a fixed factor significantly improved the

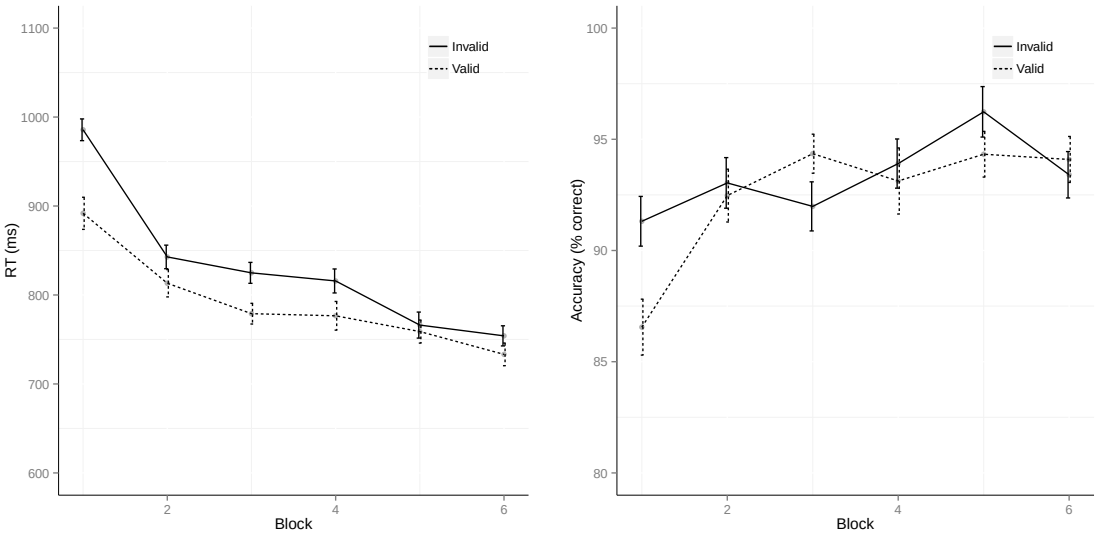

Figure 3.11: Timecourse of reaction times in milliseconds (left plot) and accuracy rates (percent correct; right plot) across all six blocks in Experiment 3.5 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

fit of the maximum random structure model (β = -33.78, SE = 5.71, $\chi^2(1)$ = 23.83, p <.001), as did including validity as a fixed factor (β = -38.80, SE = 7.46, $\chi^2(1)$ = 26.94, p <.001). Comparison of the two-fixed-factor models (block + validity and block * validity) did find evidence for an interaction of the two (β = 13.03, SE = 4.37, $\chi^2(1)$ = 8.88, p = 0.003). However, our primary point of interest was the last two blocks, where the change in gaze behaviour could cause participants to experience an error signal. Closer examination of the last two blocks found a marginal effect of block (β = -20.21, SE = 10.98, $\chi^2(1)$ = 3.38, p = 0.066), but no evidence for an effect of validity (β = -12.81, SE = 10.92, $\chi^2(1)$ = 1.38, p = 0.241).[1] The effect of block appears to be because RTs in block 6 were generally faster than in block 5, which is not consistent with an error signal. There was also no interaction between these two blocks (β = -10.82, SE = 21.82, $\chi^2(1)$ = 0.25, p = 0.619).

The analysis of accuracy rates found that adding block to the null model

[1]Note that including block as a fixed factor reduces the power in each cell in the analysis – there were, for example, only eight valid and eight invalid trials in each block, and controlling for identity reduces this to only two presentations per block. As such, it is likely that this effect of validity in the final blocks is underpowered to detect a difference rather than participants learning to overcome invalid cues (see Appendix A).

significantly improved the fit (β = 0.02, SE = 0.01, $\chi^2(1)$ = 9.14, p = 0.002), but including validity did not (β = -0.02, SE = 0.01, $\chi^2(1)$ = 1.63, p = 0.202). Comparison of the two-fixed-factor models (block + validity and block * validity) found no evidence of an interaction (β = 0.01, SE = 0.01, $\chi^2(1)$ = 2.40, p = 0.121). Closer examination of the last two blocks found a marginal effect of block (β = -0.03, SE = 0.02, $\chi^2(1)$ = 2.89, p = 0.089), but none of validity (β = -0.01, SE = 0.02, $\chi^2(1)$ = 0.36, p = 0.547), and no interaction (β = 0.05, SE = 0.04, $\chi^2(1)$ = 2.04, p = 0.153).

Trustworthiness ratings

The changes in trustworthiness ratings for the faces in Experiment 3.5 are shown in Figure 3.12. Adding time to the null model did not make it fit the data significantly better (β = 0.28, SE = 1.63, $\chi^2(1)$ = 0.03, p = 0.864), and including validity only marginally improved the fit (β = -2.96, SE = 1.58, $\chi^2(1)$ = 3.49, p = 0.062). However, the interaction model fit the data significantly better than did the full model, where both factors were modelled but without an interaction (β = -11.79, SE = 3.15, $\chi^2(1)$ = 13.96, p <.001).

We ran further analysis of the changes in trustworthiness as a function of time for valid (with time | identity removed) and invalid faces (with no random slope terms) separately, and in this experiment including time as a fixed factor significantly improved the model fit for both valid (β = 6.17, SE = 2.21, $\chi^2(1)$ = 7.57, p = 0.006) and invalid faces (β = -5.62, SE = 2.18, $\chi^2(1)$ = 6.59, p = 0.010).

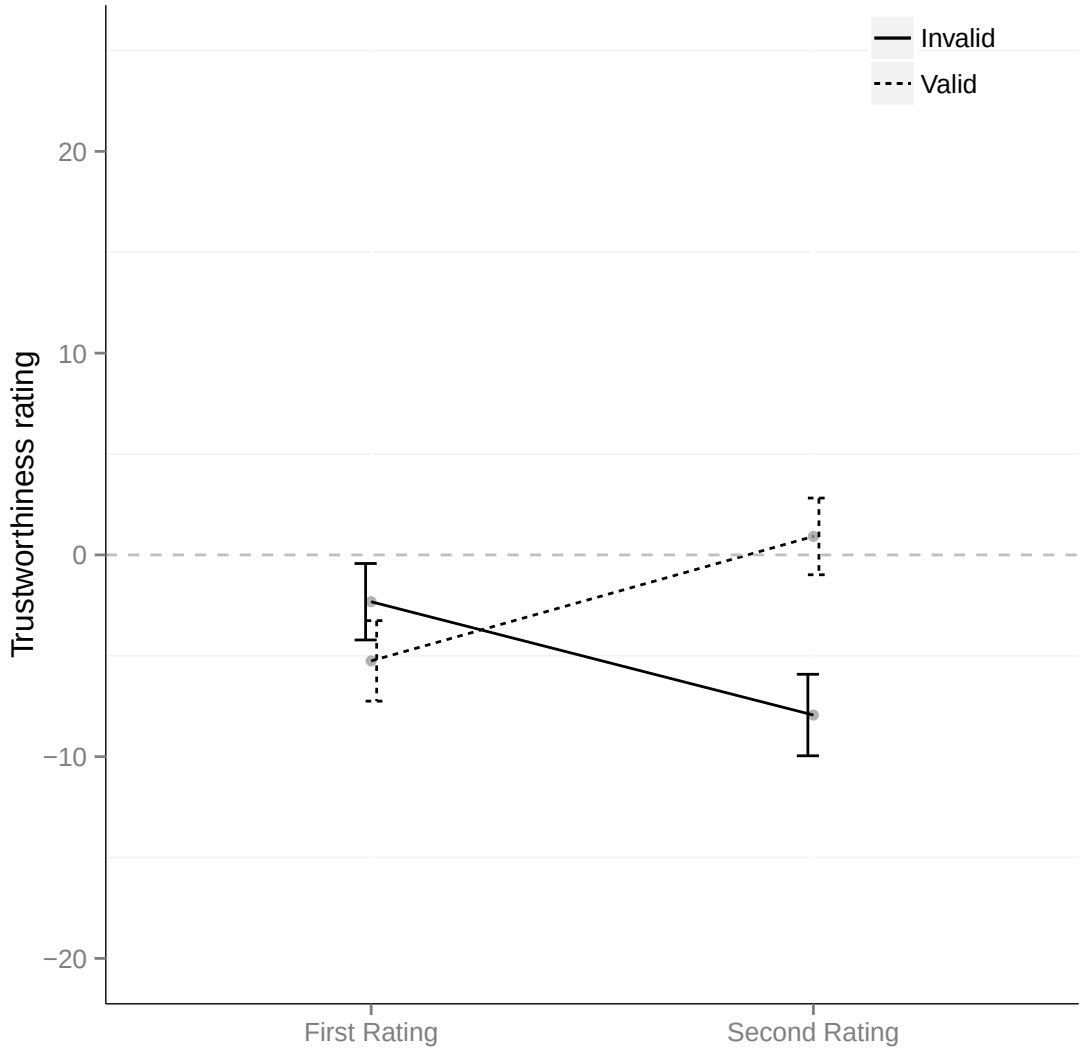

Figure 3.12: Time course of trustworthiness ratings over Experiment 3.5 for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

Cross-experiment analysis

In order to see how an interference task and familiarity level impacted the overall effect, the results of Experiments 3.1, 3.2, 3.3, 3.4 and 3.5 were combined (see Figure 3.13). In this analysis, the null model included the maximum random structure and no features had to be removed to allow for convergence. The outcome variable was now change in trustworthiness over the course of the experiments (as such, time was no longer a factor and was now a property of the measured variable). Fixed factors were validity (valid/invalid) and experiment (1-4). The null model was compared to validity-only and

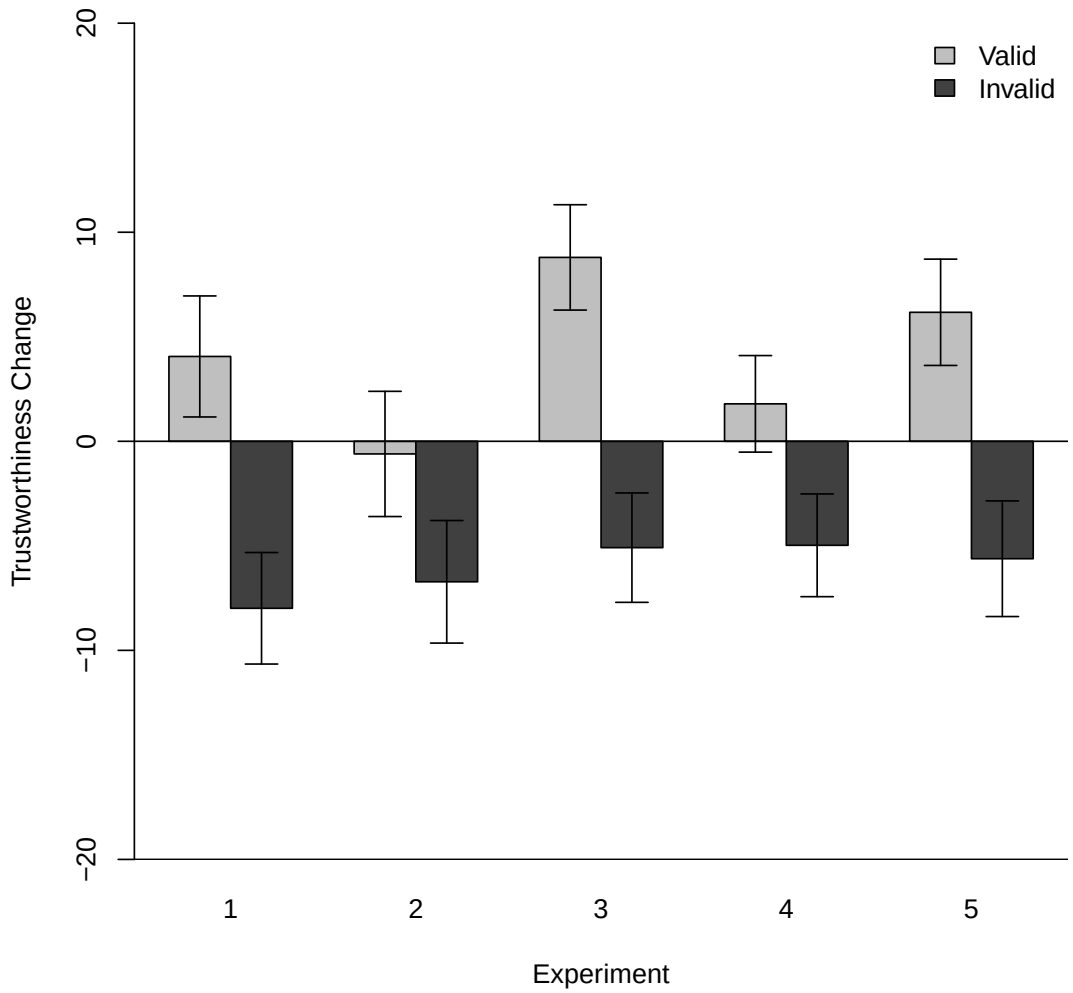

Figure 3.13: Changes in trustworthiness in Experiment 3.1 (where there was no filler and no familiarisation task), Experiment 3.2 (filler task but no familiarisation task), Experiment 3.3 (both a filler and familiarisation task), Experiment 3.4 (a familiarisation task and an hour's gap before second rating), Experiment 3.5 (no familiarisation task, but where the faces changed their cueing behaviour for the last block). The graph shows the change in trustworthiness ratings over the course of the experiment for valid (dotted) and invalid (solid line) faces. Error bars show standard error.

an experiment-only models separately, and then an interaction model (validity x experiment) was compared with a non-interaction model (validity + experiment). No models would converge with the validity | subject term included.

Adding validity to the null model significantly improved the fit ($\beta = 10.12$, $SE = 1.23$, $\chi^2(1) = 66.80$, $p < .001$), but including experiment did not ($\beta = 0.66$, $SE = 0.69$, $\chi^2(1) = 0.90$, $p = 0.343$). The interaction model of validity x experiment did not fit the

data better than did the model where both factors were modelled but without an interaction ($\beta = 0.01$, $SE = 0.87$, $\chi^2(1) = 0.00$, $p = 0.988$), indicating that incidental learning of trust was largely similar across experiments.

3.6 Chapter Discussion

Even when ignoring a face its pattern of eye-gaze behaviour can be learned, subsequently influencing ratings of the faces' trustworthiness: the key question explored here concerned the stability of this incidental associative learning process. After replicating the basic effect in Experiment 3.1, we showed in Experiment 3.2 that with minimal interference the evidence of trust learning was weaker, although we did not completely eradicate the pattern.

We go on to show in Experiment 3.3 that by including a face familiarisation task at the beginning of the experiment, the effect can now convincingly survive the same interference that weakens it in Experiment 3.2, and in Experiment 3.4 we show that traces of this trust learning can persist an hour after gaze cueing has ended. Finally, Experiment 3.5 demonstrates that even when the same faces are presented and the gaze behaviour is reversed, trust can survive. The change scores of initial to final trust ratings from each experiment are shown together in Figure 3.13.

This is the first study that investigates how long this incidental learning can last, and we show that it can be durable and somewhat resilient to interference. While Experiments 3.2, 3.4 and 3.5 tend to show weaker learning effects than we see in Experiments 3.1 and 3.3, the overall profile of results persists, and we are confident that although these interference tasks do appear to weaken the effect, it nonetheless survives. This is supported by the fact that a cross-experiment analysis found that modelling an

interaction of validity and experiment when predicting changes in trustworthiness to faces did not fit the data significantly better than modelling no interaction.

Of course there could be other interfering tasks that are more disruptive of incidentally learned trustworthiness. For example, re-presenting the faces used in the experiment in a task where no gaze cueing was observed is likely to cause extinction of prior learning (c.f. Rogers et al., 2014). However, Experiment 3.5 shows that when there is still gaze cueing, these representations are maintained, even when the stored information is at odds with incoming experience. Experiment 4.3 in Chapter 4 provides further converging evidence for this point. This maintenance may be due to context − if the faces appear in a context outside of gaze cueing we may abandon maintaining the memory of their helpful or deceptive cueing behaviour. However, if they continue to provide cues (even if these cues now change) participants appear to maintain memories for task-relevant information even if it does not necessarily fit with what the face is currently doing. In contrast, it is likely that exposure to faces not presented in the experiment might not disrupt prior learning. For example, during the one-hour interference task where participants left the laboratory (Experiment 3.4) they were exposed to other faces as they moved around the campus, and the effect survived. Clearly more formal studies of the stability of incidentally learned trust will be worthwhile.

The stability of trust learning even after faces demonstrate the opposite cueing behaviour raises the question of how much information is necessary to learn about trustworthiness, as well as to override these stored representations. Manssuer, Roberts and Tipper (2015) investigated this with EEG and found a late positive potential that differentiated between valid and invalid faces that arose around the fifth presentation of the face. As such, it may be that two presentations of each face experienced in the sixth

block of Experiment 3.5 were not enough to override the learned information from the ten exposures in the preceding five blocks.

We initially considered two possibilities: that this incidental trust learning might be short-lived, and reflect in-the-moment monitoring of statistical contingencies, or that these traces might be integrated into a longer and more durable representation. We find evidence supporting the latter interpretation, and taken with previous research using a similar paradigm we begin to see a complex underlying mechanism for social learning emerge. This suggests that at some point during gaze cueing (likely around the fifth presentation of the face identity, c.f. Manssuer, Roberts & Tipper, 2015) information about the cueing behaviour of faces is transferred to a longer term and more durable representation that feeds into a network that focuses on invariant aspects of identity recognition.

Interestingly, in Experiment 3.1 we replicate the result of Experiment 2.1 that the change in trustworthiness over the course of the experiment is primarily driven by a decrease in trust towards invalid faces. This pattern persists into Experiment 3.2, where despite the weaker learning we nonetheless see that the decrease in trust to invalid faces is greater than the increase to valid. However, this does not emerge in Experiment 3.3 (or in Experiment 3.4, although this could be due to the much weaker learning that we see after an hour's interference). In Experiment 3.3, trust learning appears to be driven instead by an increase in trust to valid faces. The internal representation of faces differ when the faces become familiar. Although not reported in previous work, it is possible that the prioritising of deception detection when interacting with unknown individuals shifts to greater sensitivity to encoding trustworthy actions when people are more familiar. This shift to remembering helpful behaviour rather than cataloguing deceivers

could be a point for future research to investigate further.

However, when the results of these five experiments are seen together, as in Figure 3.13, it seems striking that the decrease for invalid faces (while it may change slightly across experiments) is rather more stable than the change for valid faces. This lends support to our hypothesis that there may be differences in how traces of learning survive between valid and invalid faces, as it appears that memory for invalid faces is more stable and appears to survive the interference that reduces the signal in Experiments 3.2 and 3.4. This finding supports previous literature that shows memory advantages for cheaters or deceivers over co-operators (Bell, Buchner, Erdfelder et al., 2012; Buchner et al., 2009; Suzuki & Suga, 2010). As people's default expectation of others is for them to co-operate rather than deceive, invalid gaze cues provide a clear error signal that results in a stronger and lasting memory for the interaction partners involved.

A final aim of this chapter was to explore whether an additional block where the gaze cueing patterns reversed would lead to a cost in RTs or accuracy. Previous research has shown that implicit learning of underlying patterns can be seen in RT costs when the underlying pattern changes (Destrebecqz & Cleeremans, 2001). However, Experiment 3.5 explored this and found no evidence that such costs emerged. A possible explanation for this is that while the underlying pattern for individual faces changed (valid became invalid and vice versa), the global attentional cueing pattern remained the same (half of the trials provided valid cues and half provided invalid cues). Similarly, Driver et al. (1999) showed that gaze cueing is automatic and not easily affected by other factors such as informing participants that gaze direction was counter-predictive. What our finding shows is that although participants are learning about the faces, as evidenced by intact trust learning, this does not impact their attentional cueing performance. It is possible

that the active processes that lead to impression formation are happening later in the trials, following on from gaze cues but not able to impact participants' preparedness or strategies.

This question of whether trust learning could be measured using a technique other than asking directly about the face raises an interesting point. Although measuring it through RTs in this way is not possible, it is potentially advantageous to be able to measure trustworthiness without having to ask about the face. More implicit measures could, for example, allow multiple tests of the same subjects without the fear of demand characteristics. This is the question that the experiments in the next chapter attempt to address.

In conclusion, we have reported the results of five experiments that examine for the first time the question of how durable cueing-induced changes in trustworthiness are, and we show that although the effect tends to deteriorate over time, it is still surprisingly resilient and traces of it do survive, particularly in the form of decreased trust towards invalid faces. It can even survive instances where the faces change their behaviour. With more familiar faces these effects can be seen up to an hour after cueing exposure has ended. Taken together these results point to a mechanism for building robust, lasting representations of the identities of deceptive or unhelpful interaction partners, even when not explicitly attending to them while focused on a different task.

Chapter 4. Exploring alternative measures of incidentally learned trust

In Chapters 2 and 3, incidental learning of trust has been replicated several times using a dual scalar rating system to measure changes in social attitudes. While this paradigm has been shown to be effective at not cueing participants in to the nature of the experiment during gaze cueing (c.f. Experiment 2.4), it nonetheless does ask directly about the faces at the beginning and end of the experiment.

It is likely that the repeated request for ratings of face trustworthiness would alert participants to the nature of the study, and change their strategies. Hence such explicit ratings would prevent multiple measures of the effect − for example, tracking how the effect endures over time within-subjects, or investigating test-retest reliability. The aim of this chapter is to explore possible alternative measures of participants' incidental learning, ones that access trustworthiness impressions in an implicit way without directly mentioning the face's trustworthiness, to make it possible to ask such questions in the future.

The first part of this chapter aims to investigate whether gaze cueing behaviour can systematically affect participants' memory for the physiognomic architecture of the faces. Physical cues to trustworthiness are so ubiquitously and automatically associated with reliable social judgements of trust that it seems plausible that these could be systematically incorporated into memory for faces − a potentially efficient way of remembering that a certain individual is untrustworthy could be to adapt the memory of their face's physical features to express the architectural features typically associated with untrustworthiness (e.g. lowered brow ridge, square jaw, etc.) to facilitate access to this learned trustworthiness information. While top-down influences do not affect *perception* (and those studies that find such effects are vulnerable to certain pitfalls, see

Firestone & Scholl, 2015), it could nonetheless be that *memory* for perceptual features may be more susceptible to top-down interference, as it is not directly beholden to incoming sensory input and relies on integration with other aspects of social memory.

Experiment 4.1 replaces the scalar ratings used in Experiment 2.1 with a post-experiment memory task. In this task participants were required to adjust a morph of the faces to match exactly what they recalled viewing earlier. The morphs ranged from low to high trustworthiness. We predicted that if prior incidental learning of trust from eye-gaze could affect perceptual memory of the physical quality of the faces, previously invalid cueing faces would be morphed to appear less trustworthy.

Experiment 4.2 is a simplified version of Experiment 4.1 using just two images (one morphed to appear trustworthy, one morphed to appear untrustworthy) presented side-by-side. The pretence of the experiment was that the image the participants had seen during the experiment was one of a pair of twins, and they were asked to select the twin they believed they had seen during the experiment. Not only are Experiments 4.1 and 4.2 subtler measures of the effect, as they do not ask about trust at all, but they can also give us insight into how the participants' perceptual representations of the faces may be altered by the faces' prior gaze behaviour.

The second part of this chapter investigates whether the learned trustworthiness associated with faces can generalise to affect judgements of unrelated stimuli. There is some evidence that information about faces can influence judgements of non-face objects: for example, Strick, Holland and van Knippenberg (2008) showed that facial attractiveness could increase desire for associated objects. In Experiment 4.3 we explore whether the same can be said for incidentally learned trustworthiness: we investigate whether the previous cueing validity of an associated face can be detected in the

aesthetic judgements of artistic images (decorative letter Hs, Mandelbrot fractals and Kandinsky-style artwork). The driving hypothesis is that images presented alongside valid faces will be rated more positively than those presented alongside invalid faces, as the attitudes associated with the face bleed over onto associated stimuli.

During the image rating procedure, a face would appear in the centre of the screen maintaining direct gaze, then shift its gaze to provide a valid cue to the location of the subsequent image. Previous research suggests that properties of faces (such as emotion) can transfer over to object, but only when the face is gazing at the object (Bayliss, Frischen, Fenske & Tipper, 2007).

However, this also gives us the opportunity to further investigate previous findings by Manssuer, Roberts and Tipper (2015), who ran a similar gaze-cueing paradigm in EEG and found that neural signals associated with cueing validity emerged around 1,000ms after the image had appeared in the form of a late positive potential. This suggests that, when seeing a face, retrieving the stored representation of its trustworthiness takes approximately a second. A gaze shift, however, can dislodge a person's attention to the face and redirect it elsewhere in space. As such, it may be that a short (less than 1,000ms) initial presentation of direct gaze may be insufficient for trust to be retrieved. To explore this, we included two between-subjects conditions in the image ratings in Experiment 4.3 − one where the stimulus onset asynchrony (SOA) between face onset and gaze shift was too short (500ms) to access this stored memory, and one where it was sufficient (1,000ms) to see if this influenced either the image or trustworthiness ratings. Our hypothesis was that trust learning would be disrupted in the 500ms SOA condition as participants were exposed to the face without the opportunity to consolidate their learned representations of trust, while trust learning

would survive the interference with 1,000ms SOA.

4.1 Experiment 4.1

This experiment attempts to see if the incidentally learned trustworthiness observed in Chapters 2 and 3 can be measured in terms of distortions in the memory for physical features of the faces. Instead of scalar ratings, participants could physically alter the face until it matched their memory of what they had seen in the experiment, to see if these memories had been distorted by gaze cueing behaviour.

4.1.1 Methods

Participants

A mixture of undergraduate and postgraduate students at the University of York volunteered for this study in return for payment or course credit. There were 25 participants in total, but one participant's data were not collected due to a runtime error, so there were 24 available for analysis (22 female, $M_{age} = 21.43$, $s.d. = 4.21$ years). Participants were all Caucasian to control for other-race effects impacting how participants evaluated subtle changes in the stimuli during the morphing procedure. All participants provided written consent and the study was given ethical approval by the Departmental Ethics Committee of the University of York Psychology Department.

Stimuli

The sixteen face stimuli (eight male, eight female) that were used in the gaze-cueing experiment were identical to those used in previous chapters. Stimuli used in the final task were morphed images of the faces. Morphed faces were made using the KDEF face

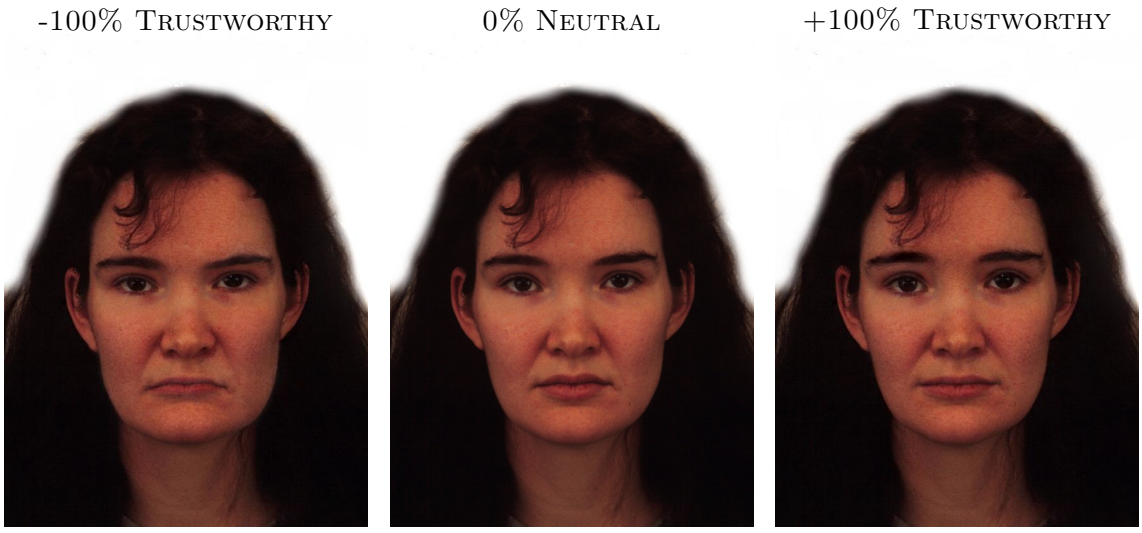

Figure 4.1: Examples of the morphed stimuli, with the original face in the centre and morphed prototypes for untrustworthy (left) and trustworthy (right).

stimuli, which were warped between trustworthy and untrustworthy prototypes (the prototype templates were based on work by Todorov et al., 2008) using JPsychomorph (Tiddeman, Burt & Perrett, 2001) to generate 20 images of each face identity morphing towards 'trustworthy' in 5% increments, and 20 morphing towards 'untrustworthy'. An example of the extremes compared with the original image are shown in Figure 4.1.

The presentation of the morphed stimuli was coded in Matlab R2012a and presented as a compiled application using the same hardware used in Experiment 2.1. The gaze-cueing section was coded in E-Studio 2.0 and presented using E-Run 2.0 on a white background.

Design and Procedure

The procedure of the gaze cueing part of the experiment used the same parameters as Experiment 2.1, the only difference being the lack of explicit trustworthiness ratings at the beginning and end. Importantly, the participants were told at the beginning of the experiment to ignore the faces as distractors and focus only on the objects that appeared

on the screen when they classified them. While this was the same in Experiment 2.1, it is highlighted here because it illustrates that participants were not informed that they would be tested on the faces later, or that they should even attend to any aspect of the face identity, meaning any effects in the subsequent procedure would be entirely incidental.

For the morphing procedure, participants were told after they had completed the gaze-cueing that they would be tested on how well they remembered the faces in the trials. They were told that morphed versions of the faces had been generated along a continuum, and their task was to use button presses on the keyboard (the comma and full stop keys, or < and >) to try to find the image that had been used in the experiment. They were told that as the changes were quite subtle, there was no time limit and they could make as many adjustments to the face as they felt necessary.

In every trial, the original image would appear on the screen (that is, the first image they saw was the 'correct' answer, although participants were not informed of this). When they were happy that the image was the same as the one they had been exposed to in the gaze-cueing section of the experiment, participants pressed the space bar to advance to the next trial. After each trial, a screen appeared asking them to rate their confidence in their decision on a scale from 1 (not very confident) to 9 (very confident). These confidence ratings were taken because previous research has indicated that people demonstrate memory advantages for cheaters or deceivers (Buchner et al., 2009; Bell, Buchner, Erdfelder et al., 2012), and Bayliss and Tipper (2006) found that participants were more likely to judge invalid cueing faces as having appeared more frequently in the experiment than valid. Although this effect only trended towards significance, we were interested in seeing whether a related but separate measure − self-reported confidence in

their responses to the supposed memory test − would yield clearer results. As such it was expected that confidence ratings would be greater in response to faces from invalid trials than from valid trials.

There were four identities used as practice trials to familiarise participants with the procedure. Participants had not seen these faces before, and so they were shown the two extremes of the scale (see Figure 4.1) and asked to morph the face to the midpoint between the two (once again, the initial face served as the correct answer).

Data analysis

RT filters were applied in the same way as in Experiment 2.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials. RT and accuracy models would not converge with the validity | subject error term included.

In the morphing procedure, participants' final image for each identity was recorded, as was their confidence in their decision on a scale of 1 (not at all confident) to 9 (extremely confident). Participants' responses were coded as the degree of separation between the chosen image and the original image (negative numbers indicate untrustworthiness), that were then multiplied by 5 to give a percentage to which participants morphed the face (since each image represented a 5% step along the continuum). A linear mixed effects model approach comparing a null model with a validity-only model was used to investigate whether the degree to which faces were morphed differed for valid or invalid faces, and all models converged with the maximum random structure.

For confidence ratings, participants' ratings of how confident they felt were compared in a similar way (comparing a null model with a validity-only model to see if

participants were more confident to one condition over the other), and again all models converged.

As with previous chapters, see Appendix A for more conventional ANOVAs and RT and accuracy rates broken down by experimental block.

4.1.2 Results and Discussion

Gaze-cueing

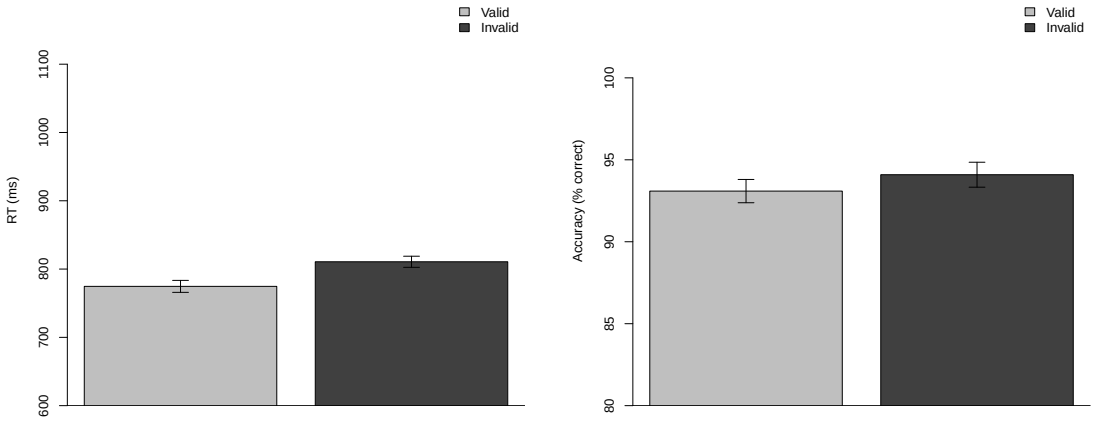

Figure 4.2: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 4.1 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

The RT and accuracy results of Experiment 4.1 are shown in Figure 4.2. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -36.40, SE = 8.51, $\chi^2(1)$ = 18.22, p <.001). This improvement was not seen for accuracy scores (β = -0.02, SE = 0.01, $\chi^2(1)$ = 1.82, p = 0.177).

Morphing results

On the whole, participants were more likely to morph images that had been invalid to appear less trustworthy (percentage distortion: M = -4.63, $s.d.$ = 39.34), and the same

Figure 4.3: Examples of the average final image chosen as the original face during cueing, when the face was valid (left) and invalid (right).

was true for valid faces (percentage distortion: M = -1.76, *s.d.* = 43.67). Fitting validity to the null model did not help to explain significantly more of these data (β = -2.88, *SE* = 4.16, $\chi^2(1) = 0.48$, $p = 0.488$). It is notable that the standard deviations of the responses were extremely high, and so this lack of difference cannot be explained by participants being accurate in their responses and answering that the initial image (or one close to it) was the correct answer − rather it appears that participants demonstrated little bias or consistency in how they morphed the faces.

The two face images in Figure 4.3 show the images with maximum probability of being the final image for invalid and valid images. As can be seen, the most probable image is further towards the untrustworthy or negative end of the scale for invalid faces (-1 images from neutral) than for valid (0 images from neutral). There was no evidence that validity biased participants' judgements in either direction from neutral.

Confidence ratings

Data were also collected that measured participants' confidence in their decisions that the image they had chosen was the original, as we felt this might give some insight into the participants' decisions. For example, even if participants were inaccurate or inconsistent in the image they chose as the original, if they differed in their confidence ratings for valid and invalid faces that could indicate that at least they felt they had a stronger representation of the features of some faces than others.

However, mean confidence ratings were largely similar for valid (M = 4.88, *s.d.* = 1.99) and invalid faces (M = 4.73, *s.d.* = 2.07), and fitting validity to the null model did not significantly improve the fit (β = -0.14, SE = 0.15, $\chi^2(1) = 0.85$, $p = 0.357$), indicating that participants reported feeling similarly confident about decisions for both kinds of faces.

The results of Experiment 4.1 showed no evidence of any effects of gaze-cueing behaviour on participants' memory of the physical features of the faces in any of the measures examined. That is, the incidental learning of trust from gaze does not change the memory for the previously viewed face, and invalid cueing faces do not appear physically less trustworthy.

However, it is possible that this may simply have been too complicated a measure for participants to report their impressions of the faces. Due to the nature of the task, it may have simply been that this null result is due to participants not being able to remember the physical details of the faces. Certainly, our memory for faces is not always perfect − particularly for unfamiliar faces during a cognitively demanding task (Jenkins, Lavie & Driver, 2005) − and asking them to recall such fine-detailed features of faces as this task nominally did (although of course the hope was to capture a broader valenced

impression of the faces, this was not what the participants were asked to do) may have swamped any effects if there were any.

This might also explain why no apparent difference was found between confidence ratings for valid and invalid faces. When interviewed afterwards most participants admitted that the final morphing task was particularly difficult, and one or two admitted that they had mostly felt like they were guessing which face it should be, which may have meant that there were no trials where participants felt particularly more confident than others. It might be that a simpler paradigm would avoid swamping any differences in participants' confidence judgements with the overall level of difficulty.

In order to address this and see if a simpler paradigm could yield a better signal-to-noise ratio, we repeated this experiment but replaced the morphing procedure at the end with a 2AFC task. We took the 50% trustworthy and untrustworthy morphs from Experiment 4.1 and told participants that the images were those of identical twins and they were to select the twin that they felt they had seen in the experiment. Broadly speaking, this is the same task as Experiment 4.1, but with more structure to the participants' response.

4.2 Experiment 4.2

This experiment simplifies the morphing response of Experiment 4.1 by getting participants to make a 2AFC of a pair of faces (one morphed to look trustworthy, one morphed to look untrustworthy) that they believe they had seen during the experiment. By simplifying the response that the participants made and limiting the decision to a categorical distinction (one face or the other), it was hoped that this would limit confounds identified in Experiment 4.1.

4.2.1 Methods

Participants

A mixture of undergraduate and postgraduate students at the University of York volunteered for this study in return for payment or course credit. There were 24 participants in total, but one had to be removed after RT filters were applied and so the total number available for analysis was 23 (18 female, $M_{age} = 21.52$, $s.d. = 4.26$ years). Participants were all Caucasian to control for other-race effects impacting how participants evaluated subtle changes in the stimuli during the morphing procedure.

Stimuli, Design and Procedure

Figure 4.4: Examples of the 'twin' stimuli. Images such as these were presented side by side and participants were asked to judge which they had seen during the gaze-cueing experiment.

Stimuli were the same as in Experiment 4.1 for the gaze-cueing procedure. For the subsequent judgements, participants were shown a screen immediately after the cueing

procedure and told that the images they had seen in the experiment had been taken from a database of identical twins, and that we were interested in seeing how well they could implicitly learn the details of a face. The twin images were actually the 50% morphed images from Experiment 4.1 − these were chosen to avoid any distortions present in the 100% prototypes while still ensuring that they were an equal distance from neutral and were recognisably different[1]. See Figure 4.4 for examples.

Each 'twin' pair was presented side-by-side for a maximum of 3,000ms − this was done to give participants enough exposure to the face to make their decision, but not enough to examine the faces too closely such that they still had to rely on a gut decision. After 3,000ms the faces would disappear from the screen, but there was no upper time limit on how long the participants took to make their response.

Once their response was registered, participants were asked to rate their confidence in their decision on a 9-point Likert scale (1-not very, 9-very confident) and had to press the space bar to advance onto the next trial. The response was classed as which side of the screen they believed the correct image was on, with the response keys Z for the left hand side and M for the right. Each face pair appeared twice, and the order was fixed but counterbalanced across participants such that each face was always the same number of trials away from its repeat as every other one (i.e. the first pair would be the first to repeat, the second would be second, and therefore each pair would be separated from its repeat by 15 other pairs of faces). The current analysis looks only at the judgements made on the first presentation of the faces.

One consideration about Experiment 4.1 that may have contributed to a null result

[1] A pilot study with five participants was run to ensure that the images were dissociably different, and that the trustworthy or positive images did actually appear more trustworthy. Independent judges went through each face pair twice and selected either the trustworthier or the less trustworthy of each pair. No judge scored more than one error on the 32 trials.

was that we removed the initial trustworthiness ratings that were present in Experiment 2.1. Although in the initial experiment these were used to calculate trustworthiness changes over the course of the experiment, they also serve the unintended purpose of making participants study the faces and evaluate them in terms of their physical characteristics, a deeper level of encoding than one gets from simply following gaze direction alone, and one that could lead to more stable memory for the identities (see Experiment 3.3). As such, in this experiment we included an initial trustworthiness rating to allow for deeper initial encoding of the faces.

Data analysis

The same RT filters were applied to these data as in Experiment 4.1, and all but one participant retained enough trials and achieved high enough accuracy to be included in the final analysis. No RT or accuracy models would converge until the validity | subject term was removed.

When participant chose the image they believed they had seen in the experiment, they were coded either as congruent (valid face, trustworthy morph; invalid face, untrustworthy morph) or incongruent (valid face, untrustworthy morph; invalid face, trustworthy morph), and these counts were analysed to see if more congruent responses were made than incongruent. As this is a binary response, we analysed the number of congruent responses using a binomial test.

For confidence ratings, we compared these using linear mixed effects models and fit a validity-only model to compare with a null model to see if validity explained any of the variance in participants' confidence ratings. Both models converged with the maximum random structure. We also generated a model with congruency (whether the selected

image matched to the validity in line with our predictions) as a fixed factor, which we compared with the null model, to see if participants were more confident when the image they chose matched with the face's previous behaviour, and with the validity-only model, to see whether confidence was better predicted by face validity or choice congruency. The congruency model would not converge with any random slope terms defined, and so these were removed from all models to allow for direct comparison.

Initial trustworthiness ratings were not analysed as they did not inform our hypotheses.

4.2.2 Results and Discussion

Gaze-cueing

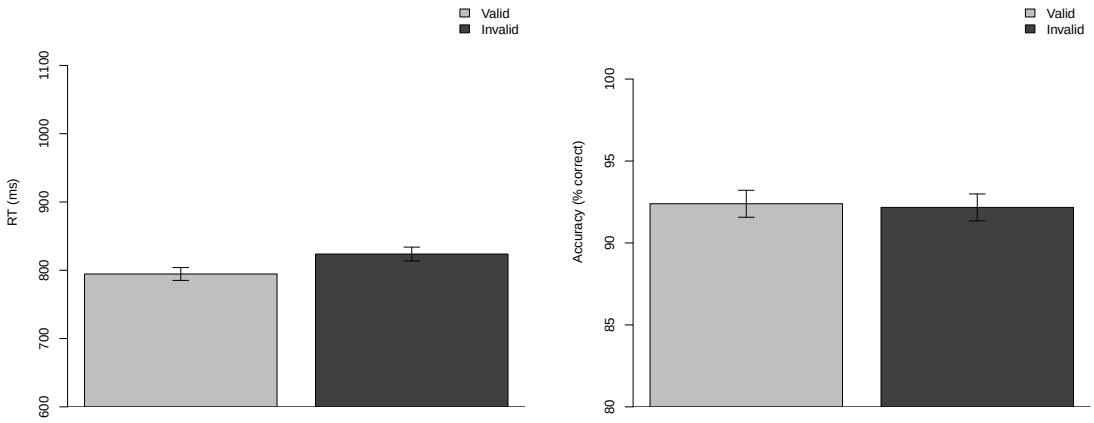

Figure 4.5: Timecourse of reaction times in milliseconds (left plot) and accuracy rates (percent correct; right plot) across all five blocks in Experiment 4.2 in response to valid (light grey) and invalid (dark grey) trials. Error bars show standard error.

The RT and accuracy results of Experiment 4.2 are shown in Figure 4.5. Fitting validity to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -28.23, SE = 9.84, $\chi^2(1) = 8.22$, $p = 0.004$). This improvement was not seen for accuracy scores ($\beta = 1.00$, $SE = 0.74$, $\chi^2(1) = 1.82$, $p = 0.177$).

Image choices

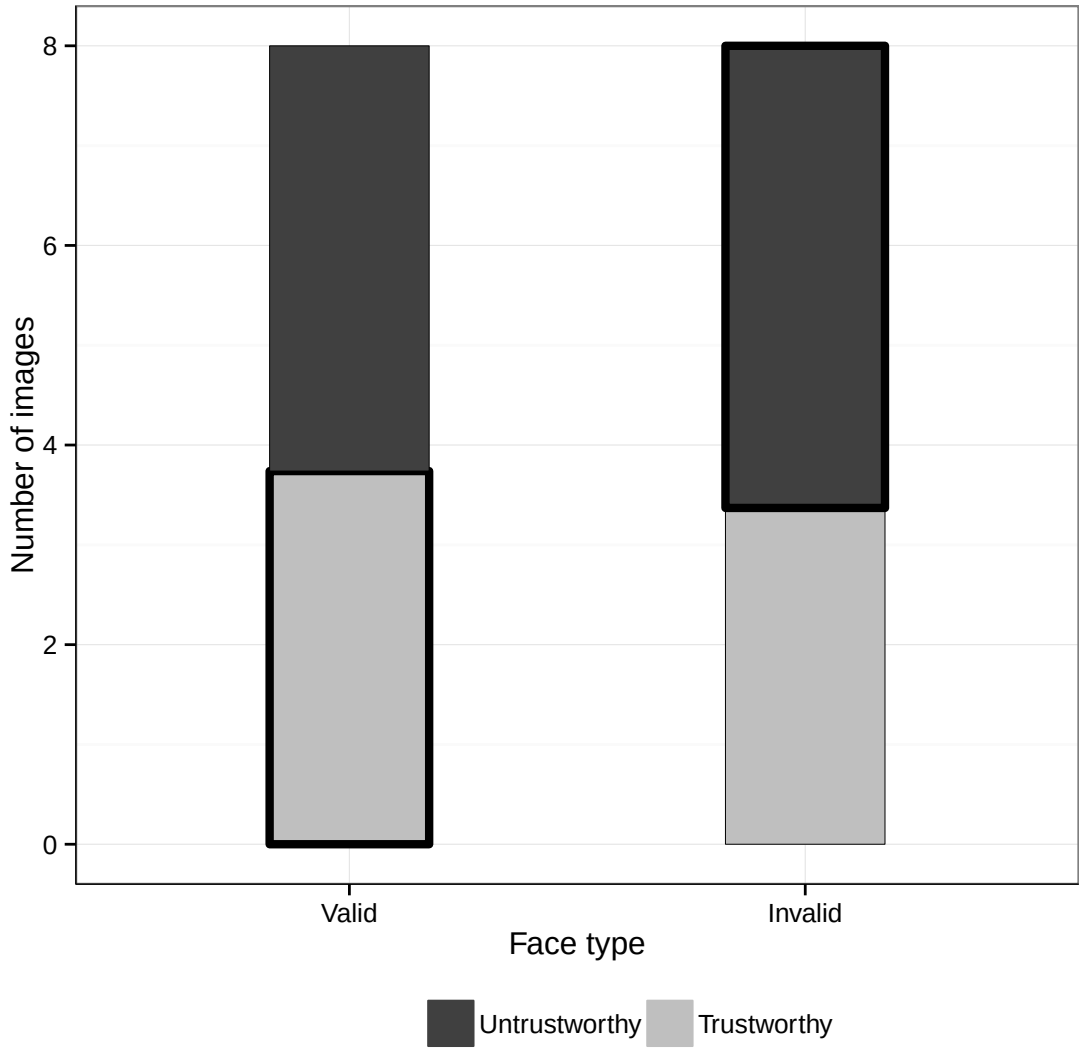

Figure 4.6: Graph to show the proportion of the four different choice outcomes – selecting the more trustworthy (bottom row, light grey) or untrustworthy image (top row, dark grey) of a valid-cueing (left) or invalid-cueing (right) face. The congruent choices are denoted by thick black borders. Entirely congruent choices would show the left bar as entirely blue (all 8 valid faces chosen as trustworthy image) and the right as entirely red (all 8 invalid as untrustworthy image).

The results of Experiment 4.2 are shown in Figure 4.6. The graph shows the proportion of the 8 valid and 8 invalid images that were selected as positive or negative. For each participant, the number of faces that were chosen in line with our hypotheses (Congruent; positive valid faces, negative invalid faces, 8.37) and the number that were not (Incongruent; negative valid faces, positive invalid, 7.63) were calculated. The results of the binomial test found that participants did not select the congruent image

significantly more often than could be explained by chance ($p = 0.224$).

Confidence ratings

Although they did not reach significance, there were differences in confidence ratings that participants made for the different faces. Participants reported feeling more confident in their decisions about invalid faces ($M = 5.44$, $s.d. = 2.10$) than valid ($M = 5.09$, $s.d. = 2.02$). Fitting validity to the null model significantly improved the fit ($\beta =$ -0.35, $SE = 0.12$, $\chi^2(1) = 8.24$, $p = 0.004$).

Confidence ratings for incongruent trials (positive image, invalid face or negative image, valid face: $M = 5.16$, $s.d. = 1.99$) were lower than those for congruent trials (positive image, valid face or negative image, invalid face: $M = 5.37$, $s.d. = 2.13$), but fitting congruency to the model did not significantly improve the fit ($\beta =$ -0.11, $SE = 0.13$, $\chi^2(1) = 0.76$, $p = 0.385$). In fact, comparing the validity and congruency models showed that the validity model fit the data significantly better than did a congruency model ($\chi^2(1) = 7.48$, $p <.001$), which indicates that validity was a better predictor of confidence ratings (in that participants felt more confident about decisions regarding invalid faces than valid) than was congruency (the decision that would reflect the true cueing validity of the face).

Given that the task participants were asked to perform was to choose the image they remembered from the experiment, a possible explanation of this effect is that participants placed more confidence in their stored representations of invalid faces (that is, they felt they remembered them better) than valid faces, but this did not necessarily bias them towards selecting the untrustworthy exemplar over the trustworthy one. This result is reminiscent of that observed by Bayliss and Tipper (2006), where participants

reported that they had viewed invalid faces more often.

4.3 Experiment 4.3

This experiment aimed to investigate whether incidentally learned trustworthiness representations of face identities could be accessed without asking anything directly about the faces in question. Such a measure would allow for the investigation of aspects of the effect such as its test-retest reliability that have not been possible to address before. In this experiment, then, participants rated artistic images while the face images remained irrelevant background distractors that participants were never asked about, to see if the validity of the face would have an effect on an unrelated liking judgement.

4.3.1 Methods

Participants

In the 500ms SOA condition, there were 22 participants. Two of these participants' data were not collected due to runtime errors, leaving 20 (all female, $M_{age} = 19.60$, $s.d. = 1.53$). In the 1,000ms SOA condition there were 20 participants in total (all female, $M_{age} = 20.25$, $s.d. = 2.39$). As such, across the two conditions there were a total of 40 participants.

Stimuli, Design and Procedure

This experiment was identical to Experiment 3.3 in that it included a face-matching familiarisation task at the beginning of the experiment, followed by an initial trustworthiness rating and five blocks of gaze-cueing. Where Experiment 3.3 continued with a video filler task, however, this experiment replaced this with an alternative

measure of the attitudes to the faces, described below.

In this task, we intended to see whether attitudes about the faces could generalise to simultaneously-presented objects. In order to explore this there were two between-subjects conditions, which varied based on stimulus onset asynchrony (SOA). Previous research with ERPs by Manssuer, Roberts and Tipper (2015) found a late positive potential emerging around 1,000ms that appears to differentiate between valid and invalid faces, suggesting that it takes this long to access the stored representation of trustworthiness associated with that face. However, to date such a delay has not been shown behaviourally. As such, we included two levels of SOA between the face appearing and the gaze shifting to look at where the target would appear: the gaze shift would occur either 500ms (short SOA, too early to access the stored representation) or 1,000ms (long SOA, long enough to access properties of trust).

Three types of image were used; non-directional arrows, Mandelbrot fractals and Kandinsky- inspired abstract images (see Appendix D for examples). There were 16 examples of each type of image, and each face from the experiment appeared alongside one example of each image. Image type was blocked and counterbalanced across participants, such that where one participant might see arrows, fractals and then Kandinsky images, another would see Kandinsky, arrows, and then fractals, etc. An example trial is shown in Figure 4.7. Image position (left/right side of screen) alternated predictably every trial, although participants were not explicitly told this.

We were interested in the rating that participants made of the image alongside the face – whether those images associated with previously valid faces would be rated more positively, even when all faces now gazed at the images.

At the end of the experiment, participants once again completed the normal

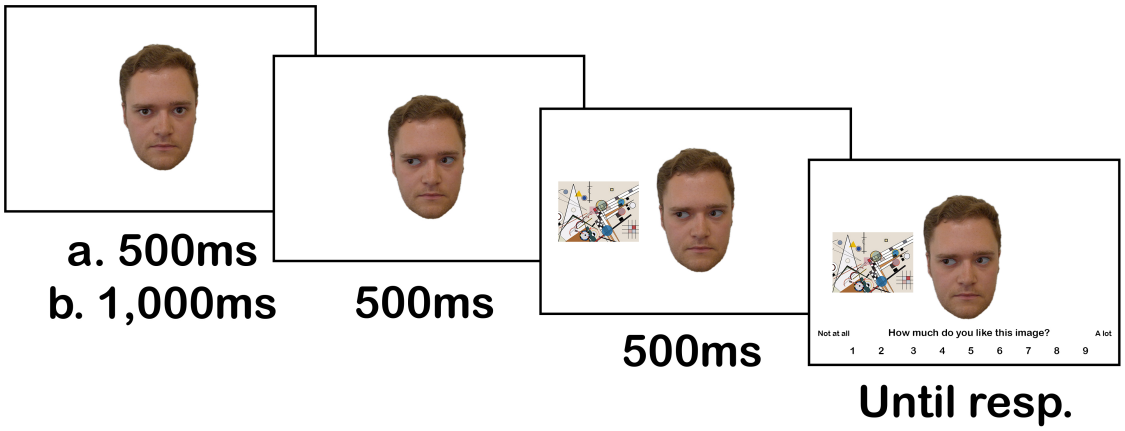

Figure 4.7: Example of an image rating trial from Experiment 4.3, with timings for a. the short and b. the long SOA conditions. In a trial, a face would appear in the centre of the screen and shift its gaze either left or right after a. 500 (short) or b. 1,000ms (long). The image (either a Mandelbrot fractal, a non-directional arrow, or a Kandinsky-inspired abstract image (pictured in trial sequence); see Appendix D) would then appear for 500ms before the question and rating scale appeared on the bottom of the screen. Participants reported how much they liked the image on a scale of 1 to 9.

trustworthiness ratings of the faces.

Data analysis

RT filters were applied in the same way as in Experiment 4.1. Data were analysed in the same way as described in Chapter 3, but with the addition of SOA (500ms/1,000ms) as a fixed factor, and models that examined the interaction of validity and SOA. This was done for peace of mind to ensure that the between-subjects groups of participants were processing the faces' gaze behaviour in a similar way during gaze cueing (e.g. if we did not find trust learning in the 500ms SOA condition as we expect, that this was not due to those participants showing abnormal insensitivity to gaze cues). For RTs, the maximum random structure would not converge until the validity | subject term was removed, and so this was removed from this model and the validity-only and SOA-only models to allow for direct comparison. The two-factor models converged with the maximum random structure.

For accuracy rates, the validity-only, SOA-only, and interaction models would not

converge with the maximum random structure until the SOA | identity term was removed, so this was removed from all models to allow for direct comparison.

When modelling trustworthiness ratings no models would converge until the time | subject term was removed. For image ratings, all models converged with the maximum random structure − which in this instance included the image identity as a random factor as well as face identity.

4.3.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 4.3 are shown in Figures 4.8 and 4.9, respectively. When analysing RTs, including validity as a fixed factor significantly improved the fit when applied to RTs, indicating a cueing effect (β = -52.10, SE = 7.76, $\chi^2(1) = 44.77$, p <.001), and including SOA had a marginal effect (β = 59.07, SE = 32.76, $\chi^2(1) = 3.28$, $p = 0.070$). This appears to be driven by faster RTs overall in the long SOA condition than the other. However, given that the manipulation of SOA did not affect the experiment until after gaze cueing had ended, this is likely noise driven by slightly different populations of participants. There is no cause for concern that differences in gaze cueing may drive any differences in learning, as there was no evidence of an interaction of validity and SOA (β = 2.80, SE = 15.87, $\chi^2(1) = 0.03$, $p = 0.860$).

Models of accuracy rates found no evidence to support an effect of validity (β = 0.01, SE = 0.01, $\chi^2(1) = 0.46$, $p = 0.499$), or SOA (β = -0.01, SE = 0.03, $\chi^2(1) = 0.04$, $p = 0.839$), nor any interaction of the two (β = -0.01, SE = 0.03, $\chi^2(1) = 0.07$, p = 0.791).

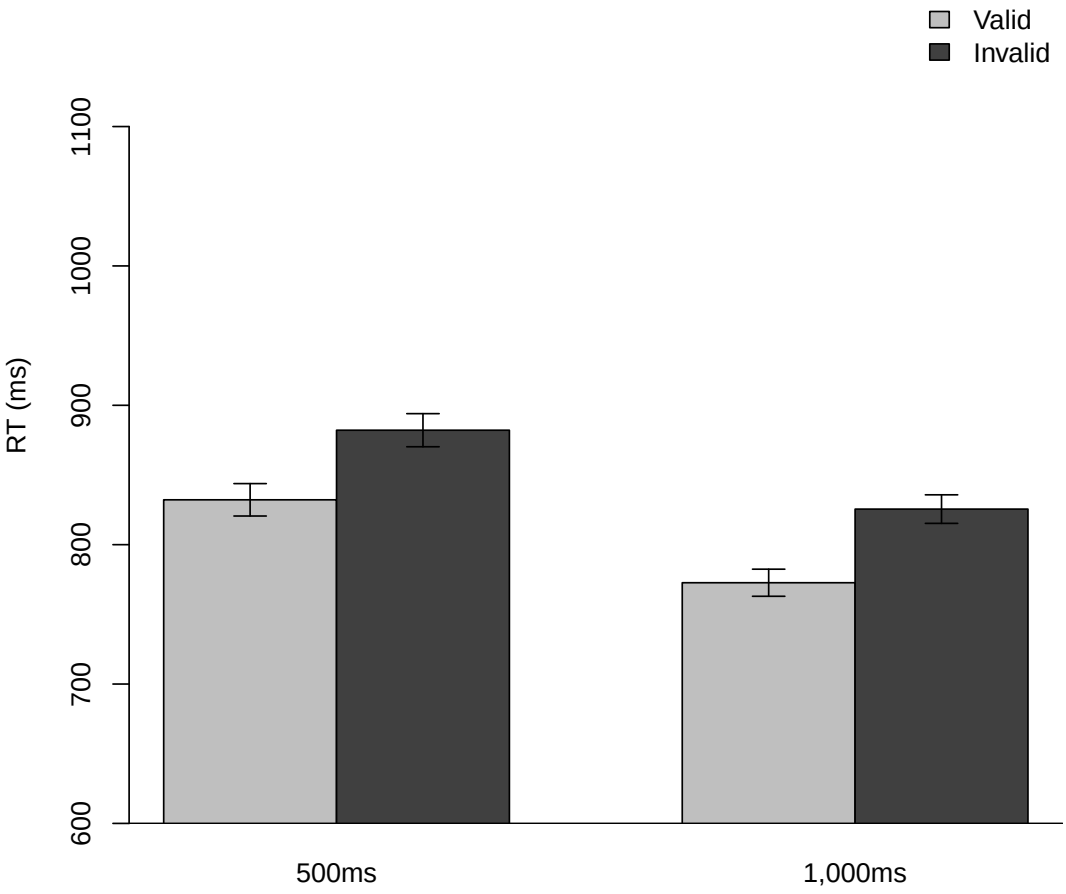

Figure 4.8: Averaged reaction times (milliseconds) in Experiment 4.3 in response to valid (light grey) and invalid (dark grey) trials in the short (500ms; left plot) and long (1,000ms; right plot) conditions. Despite different conditions, all experimental details up to the end of gaze-cueing were identical across different conditions. Error bars show standard error.

Trustworthiness ratings

The changes in trustworthiness ratings for the faces in Experiment 4.3 are shown in Figure 4.10. Adding time to the null model did not significantly improve the fit (β = -0.21, SE = 1.65, $\chi^2(1) = 0.02$, $p = 0.899$), nor did including SOA ($\beta = 0.79$, $SE = 4.34$, $\chi^2(1) = 0.03$, $p = 0.864$). However, including validity did improve the fit (β = -5.01, SE = 1.52, $\chi^2(1) = 10.59$, $p = 0.001$). The comparison of the two fixed-factor models (time +/* validity) revealed a significant two-way interaction of time and validity (β = -7.56,

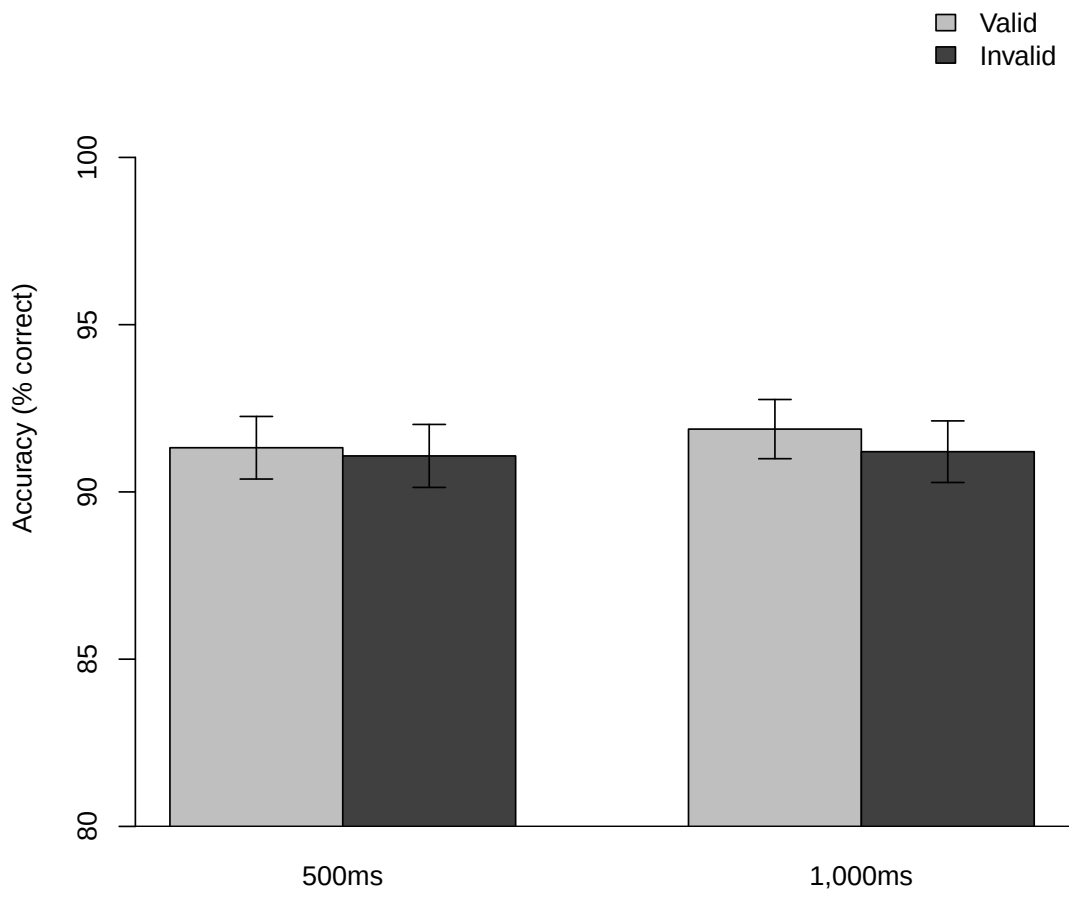

Figure 4.9: Averaged accuracy rates (percent correct) in Experiment 4.3 in response to valid (light grey) and invalid (dark grey) trials in the short (500ms; left plot) and long (1,000ms; right plot) conditions. Error bars show standard error.

SE = 2.91, $\chi^2(1) = 6.75$, $p = 0.009$), and including SOA to make a three-way interaction marginally improved the fit (β = -3.19, SE = 5.80, $\chi^2(3) = 7.12$, $p = 0.068$).

To explore this three-way interaction further, analyses were performed on the short and long SOA conditions separately. For the 1,000ms SOA condition, where the face showed direct gaze for a full second during image ratings before shifting its gaze, we examined a two-way interaction model of time and validity. The model converged after the time | subject slope term was removed, and comparison of the two models found a significant interaction ($\beta = 10.31$, $SE = 6.59$, $\chi^2(1) = 4.84$, $p = 0.028$). For the 500ms

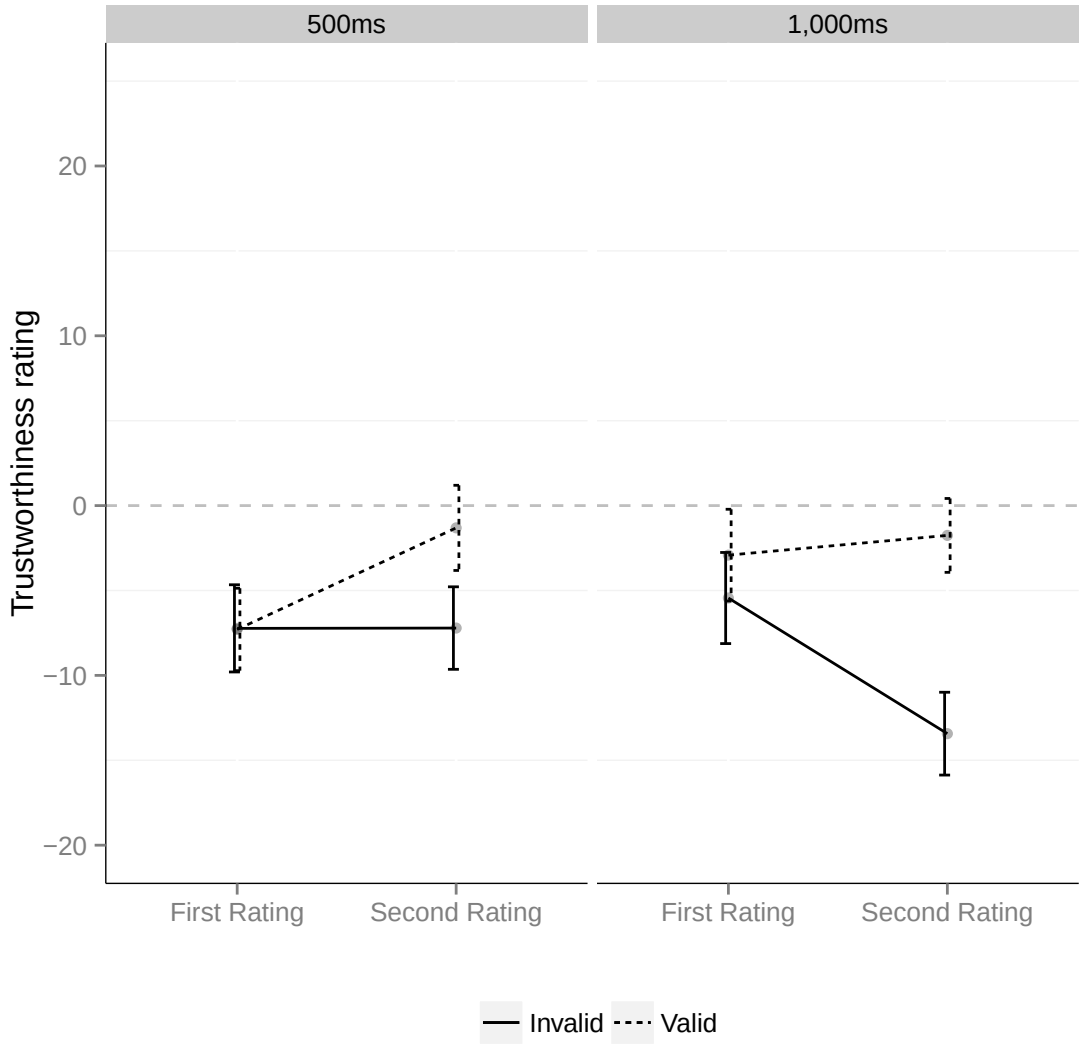

Figure 4.10: Time course of trustworthiness ratings over the experiment for valid (dotted) and invalid (solid line) faces in short (500ms; left plot) and long (1,000ms; right plot) conditions in Experiment 4.3. Error bars show standard error.

SOA condition, where the face showed direct gaze for only half a second during image ratings before shifting its gaze, the models would not converge with the maximum random structure and so both the time | identity and time | subject slope terms had to be removed. This analysis found no significant interaction ($\beta = 11.94$, $SE = 6.45$, $\chi^2(1) = 2.14$, $p = 0.143$) indicating that participants' learning of gaze cue contingencies had been selectively disrupted in this condition. This was in line with our a priori hypotheses that trust learning would be disrupted when the face shifted attention away before participants had enough time to access their stored representation of its gaze behaviour.

We also examined the change in trustworthiness for each condition of the faces (short SOA vs. long SOA; valid vs. invalid). No models converged with any random terms except for the long SOA valid models, which converged with time | identity included only. These analyses found that including time in the model, which reflects a change over the course of the experiment, significantly improved the fit for valid faces in the short SOA (β = 5.98, SE = 2.83, $\chi^2(1)$ = 4.46, p = 0.035), but not in the long SOA condition (β = 1.16, SE = 3.21, $\chi^2(1)$ = 0.13, p = 0.715). However, for invalid faces there was a significant improvement of fit in the long SOA (β = -7.99, SE = 2.67, $\chi^2(1)$ = 8.84, p = 0.003), but not in the short (β = 0.02, SE = 2.85, $\chi^2(1)$ = 0.00, p = 0.995).

Image ratings

The results of the image rating tasks are shown in Figure 4.11 as standardised rating values around the centre of the 1-9 Likert scale. Participants were generally conservative with their image ratings, as all averages tended towards the centre of the scale. Adding validity to the null model did not significantly improve the fit, suggesting that participants did not base their image rating decisions on the cueing validity of the paired face (β = -0.06, SE = 0.08, $\chi^2(1)$ = 0.55, p = 0.457). Including SOA similarly did not improve the fit (β = -0.38, SE = 0.31, $\chi^2(1)$ = 1.59, p = 0.208), but a two-way interaction of validity and SOA approached significance (β = 0.33, SE = 0.17, $\chi^2(1)$ = 3.70, p = 0.054).

However, closer examination of the two SOA conditions separately found that the effect of validity approached significance only in the short SOA condition (β = -0.21, SE = 0.12, $\chi^2(1)$ = 3.24, p = 0.072), but not in the long SOA condition (β = 0.11, SE = 0.12, $\chi^2(1)$ = 0.80, p = 0.371). This did not fit with our hypothesis that participants

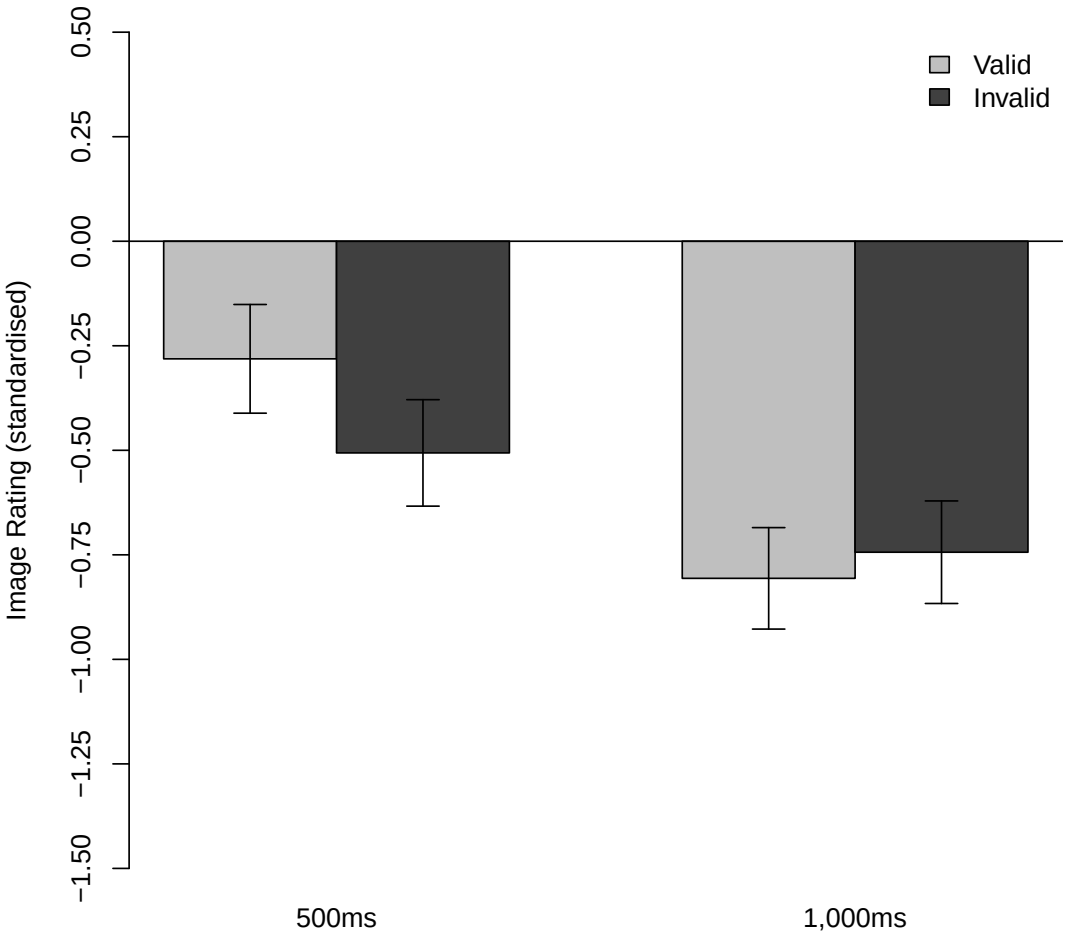

Figure 4.11: Averaged image ratings for images paired with valid faces (dotted) and images paired with invalid faces (solid line) for short (500ms; left) and long SOAs (1,000ms; right). Values are standardised as deviations from the midpoint of the response scale (5), such that negative values denote ratings below 5 and positive values denote ratings above 5. Error bars show standard error.

would only be able to access the stored trustworthiness representations in the long SOA condition, and not in the short.

4.4 Chapter Discussion

Experiment 4.1 asked whether the standard incidental trust learning could be accessed through measuring distortions in participants' memory for the physical features of faces. To do this, it used a morphing procedure in place of trustworthiness ratings; participants

could morph the face to look more or less trustworthy to match what they remembered from the gaze-cueing, with the expectation that if participants had updated their memory of the faces to appear more or less trustworthy in line with their cueing behaviour, this would cause them to morph the faces to physically match these stored representations. No systematic bias towards one direction or another was found for either valid or invalid faces. However, the paradigm used was quite complicated and required participants to attempt to match a new image to a stored representation of a face that they had been instructed to ignore.

This sort of morphing technique could prove to be a valid one, particularly in paradigms that focus on memory for faces as this can be a difficult process to examine; using a participant-controlled morphing application allows participants to feel more control to get the image to match their memory or representation that is not available when using recognition or 2AFC measures. Nonetheless, in this instance it appears that the measure was somewhat complicated and challenging for participants, and as such lacked sensitivity.

Experiment 4.2 attempted to simplify the morphing procedure of Experiment 4.1 and limited participants' choices to either a trustworthy or untrustworthy morphed 'twin'. It also included the pre-rating from Experiment 2.1, as one difference between Experiments 2.1 and 4.1 was that there was no initial consideration of the trustworthiness of the face in Experiment 4.1, which may have led to more shallow encoding of the faces (cf. Chapter 3). However, this experiment also found no evidence of gaze behaviour impacting the low-level physical properties of participants' memories of faces. This reinforces the conclusion that changes in trustworthiness elicited by gaze cues do not result in systematic differences in memory for the perceptual features of the

face. These results show that as well as not affecting perception (Firestone & Scholl, 2015), top-down learning of trustworthiness also does not seem to affect memory for perceptual features (although we must be careful to avoid over-interpreting null effects).

Both Experiments 4.1 and 4.2 also asked participants to rate how confident they felt in their answers. The morphing task in Experiment 4.1 did not yield any systematic differences in confidence, which is not surprising when the difficulty of the task is taken into account. The small changes in facial features and the difficulty recalling fine details of faces to match to the images mean that it was unlikely that participants felt particularly confident about any of their decisions. In contrast, the simpler 2AFC task used in Experiment 4.2 found that participants reported greater confidence in decisions made about invalid faces, which suggests that they may feel they remembered these faces better. Although this did not result in a response bias towards the congruent image (valid-trustworthy and invalid-untrustworthy), this does fit with other findings throughout the literature and this thesis that individuals associated with deception are privileged in memory (Bell, Buchner, Erdfelder et al., 2012; Buchner et al., 2009; Bayliss & Tipper, 2006, and the fact that decreases for invalid faces are more stable than changes for valid faces, particularly in Chapter 3).

Experiment 4.3 explored whether the trustworthiness of the faces could be accessed without asking directly about the faces at all. To this end, faces were presented alongside artistic images that participants were asked to rate, to see if the face's previous validity had an effect on image ratings. No significant effect of face validity on ratings of the associated images was found in either the short or the long SOA condition. While previous research has shown that facial attractiveness can manipulate desirability of associated objects (Strick et al., 2008), these data suggest that the same is not true for

incidentally learned trustworthiness.[2] As Strick et al. used faces that varied in their physical levels of attractiveness, it might be that such physical cues in the face are necessary to affect non-social judgements, and as shown in Experiments 4.1 and 4.2, this incidental learning does not affect memory or perception of physical features.

A second aim of Experiment 4.3 was to investigate how the timing of the intervening task might impact trust learning. Between learning (gaze cueing) and recall (trustworthiness ratings), the representations of faces' cueing behaviour must be maintained and consolidated. Experiment 3.4 has shown that this can occur during naturalistic interference in the real world, and Experiment 3.5 showed that this learning can survive reversed gaze cueing in a final block. Experiment 4.3 aimed to explore this further by making all faces now look towards a subsequent target − essentially removing the gaze cueing component while retaining the visual experience of a gaze shift. The key question was whether disrupting processing of the face by triggering an attentional shift away from it (using a gaze shift) would disrupt the memory for individuals' trustworthiness, and whether this might selectively impact valid or invalid faces.

In general, in support of the data obtained in Chapter 3 where memory for trust could survive intervening tasks, Figure 4.10 shows similar patterns of trust learning. It is noteworthy that the trust learning appears to be more robust when faces were viewed for 1000ms prior to gaze shifts in the previous art rating task. We might tentatively propose at this time, that this longer viewing time enabled retrieval of prior trust/deception associated with the faces (e.g., Manssuer, Roberts & Tipper, 2015), and this supported the consolidation of trust during this intervening period. Certainly the current trust

[2]Note that Strick et al. (2008) also found that direct gaze was necessary to show this transfer. Although they are not reported here, we also ran versions of Experiment 4.3 where the face maintained direct gaze throughout the trial, but found no effects of face validity on image ratings in either short or long SOAs. This is more in line with the findings of Bayliss et al. (2007), who found that emotional information conveyed by expression (which is more closely related to judgements of trustworthiness than are judgements of attractiveness) requires that gaze be directed towards the object.

rating data supports Experiment 3.5. That is, memory for incidentally learned trust can even survive when the same faces are exposed again in a different task context, even though there is no longer a speeded response component to the task, and no invalid cues from any faces. Interestingly, it seems that this representation is maintained even when faces demonstrate a non-informative gaze shift.

Although none of the approaches described here would serve as alternative measures of incidentally learned trust, there are other options that may serve as a useful basis for future research. For example, the economic trust games used by Rogers et al. (2014) can be used to investigate different facets of this learning, and using different types of game from the one-shot games used in that study (for example, seeing how participants then adapt to particular investment behaviours from valid and invalid faces) could yield important insights. Alternatively there are other measures that are starting to emerge from technological advances with motion capture and virtual reality: Proxemic imaging is a technique that has been used to measure approach and avoidance behaviours in social interaction by measuring how participants position themselves in relation to their partner (McCall & Singer, 2015). This has the advantage of providing a nuanced and rich source of data without ever having to ask directly about participants' attitudes towards the identities.

The main aim of this chapter was to explore alternative measures of incidental trust learning, which it did in three ways over three separate experiments. However, none of the avenues explored proved viable alternatives to directly asking about trustworthiness. They did, however, reinforce some important points. Experiment 4.2 shows that participants appear to feel more confident about their memories for invalid face than valid faces, while Experiment 4.3 showed that an intervening task where the faces were

re-exposed does not completely disrupt retrieval of prior incidental learning of trust, similarly to results with non-face interference tasks reported in Chapter 3.

Chapter 5. The effect of minimal group membership on incidental trust learning

The results of Chapters 2, 3, and 4 have brought to light some of the complexities and intricacies of incidental learning of trust from gaze cues. However, one outstanding question that remains to be addressed is the issue of how the identity of the cueing faces might affect how we learn about them. In the real world, one hardly ever experiences a face entirely in isolation − people also carry with them a wealth of social information that we use to form predictions and evaluations of the person and their behaviour.

Humans are social creatures, and we rely on social groups in order to survive. These social groups can range from a smaller personal scale (e.g. a close circle of friends) to a much larger societal scale (our national identity, gender, race, etc.; Lickel, Hamilton & Sherman, 2001). People prefer individuals who are members of their in-group over their out-group, even when this group distinction is a new category that has been learned in the laboratory (Allen & Wilder, 1975; Tajfel, Billig, Bundy & Flament, 1971).

Given that people tend to trust in-group members over out-group members, there is an expectation for in-group members to behave cooperatively and out-group members to behave deceptively. According to a model of learning by Rescorla and Wagner (1972), violations of such expectations should lead to increased learning as the cognitive system responsible has to reconcile the error between what is expected and what occurs. In this case, the expectation would be that in-group faces should co-operate and that out-group faces should deceive. As such, in-group invalid faces that deceive will be judged more punitively than out-group invalid faces due to violating their expectations. This relates to social psychological research on the black sheep effect whereby in-group members who act negatively are judged more harshly than out-group members who act negatively (Marques, Yzerbyt & Leyens, 1988). As for out-group valid faces that facilitate

performance, they could lead to a greater increase in trust than their in-group counterparts due to greater learning of expectancy-violating events.

There is an alternative hypothesis to this, however. Most previous research into trust and group membership has focussed primarily on broad group-level differences. However, the current paradigm involves in-the-moment experiential learning of individual behaviours, which are conceptually separate from group-level prejudices. As in this paradigm an equal number of in-group members as out-group members deceive the participant, it is possible that learning trustworthiness from gaze cues may override the group representation − therefore, exposure to valuable information about an individual's behaviour may be a way of reducing this classic in-group favouritism, as people judge others on the basis of their behaviour rather than who they are.

Alternatively, and perhaps more pessimistically, group membership may supersede individual learning, and so the presence of an out-group (particularly one that is generated as a result of explicit experimental instructions) could drive participants to default to a group-level representation without including information about that person's individual history of behaviour. That is, once people are aware they are interacting with two groups of individuals, both in- and out-group, learning from unique patterns of individual behaviour no longer takes place as broader categorisation at the group level dominates.

Over the course of two experiments, this chapter aims to explore this issue of how knowledge of group membership may influence incidental trust learning from gaze cues. In Experiment 5.1, we assign participants to minimal groups and find that this does extinguish any changes in trust on the basis of individual behaviour. Experiment 5.2 serves as a control experiment and removes explicit references to these groups, to match

previous versions of the paradigm and replicate the original incidental learning of trust effect.

5.1 Experiment 5.1

Experiment 5.1 aims to explore whether incidental trust learning is affected by social knowledge about others by using a minimal groups paradigm. Participants were assigned to one of two minimal groups to see if identifying with half of the faces affected how they learned about those and other faces' trustworthiness.

5.1.1 Methods

Participants

A mixture of undergraduate and postgraduate students at the University of York volunteered for this study in return for payment or course credit. There were 34 participants in total, but due to runtime errors data had to be removed from three of these, and one further participant was removed after RT filters left less than 70% of the original trials. As such, there were 30 participants included in the final analysis (18 female, $M_{age} = 20.17$, s.d. $= 1.57$).

Stimuli

All stimuli were the same for the gaze-cueing portion of the experiment as the smiling faces used in Experiment 2.2, with the exception of the colour of the face images' t-shirts. Shirt colours were edited in PhotoShop and 30% chromatic filters were applied to the default grey. Examples of the stimuli used are shown in Figure 5.1. At the beginning of the experiment participants were assigned to either the Overestimator group or the Underestimator group, and were instructed at the beginning of the

experiment that the faces had completed a similar classification task. For half of the participants, blue shirts signalled an in-group member, yellow an out-group, while for the other half this was reversed. Smiling faces were used because this has been shown to lead to more robust learning and a bidirectional change in ratings of trustworthiness (cf. Experiment 2.2), which would allow us to better track any effect of group membership.

Original Blue Yellow

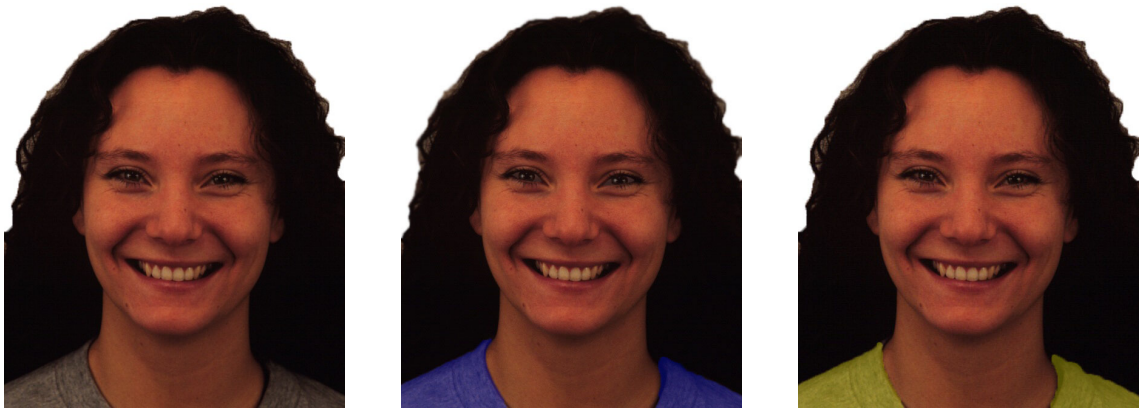

Figure 5.1: Examples of the stimuli used in the group-membership experiment, along with the original grey-shirt image (left). Participants were instructed that either the blue or yellow shirt signified their in-group, and that the other signified their out-group.

The classification task undertaken at the beginning of the experiment involved participants watching 12 arrays of moving dots. Participants were asked to estimate the number of dots in each array. The number of dots was either 25, 50 or 70 white dots moving incoherently at 1/500 units per frame on a grey background for 5 seconds. Participants had a sliding scale to report the number of dots. After the total number of trials, participants were then shown a screen that read, 'Processing...' for one second to give the impression of the machine making calculations, and then were shown a screen that read as follows:

> Your results show that you are an [overestimator/underestimator].
>
> Research has shown that people who score like you on these tests tend to perform well at tasks involving visual attention and memory, and also work

well in team exercises.

The additional blurb did not vary depending on how participants were classified. This blurb was added both to increase belief in the fact that this group classification would be somehow related to the gaze-cueing portion of the experiment, and to make it seem like they were expected to perform particularly well, which would lead to more negative consequences for misleading gaze cues. This supposed calibration procedure was coded using PsychoPy2 Experiment Builder (Peirce, 2007) and run as a standalone Python file.

Design and Procedure

The gaze-cueing portion of the experiment was essentially similar to that in Experiment 2.2, where smiling faces appeared on the screen and cued either the correct or incorrect location. The only difference was the addition of shirt colour, which participants were told signalled whether the person was 'like them' (either an over- or underestimator) or not like them.

In order to ensure that participants were familiar with which colour corresponded to their group, participants consent forms were 'filed' in a clear plastic wallet with either a blue or yellow sticker, corresponding to the colour associated with that group that was left on the desk next to the participant, to serve as a reminder of which colour was their own group. At the beginning of the experiment participants also performed a categorisation task with the faces − a face would appear in the centre of the screen wearing a blue or yellow t-shirt and participants (having been instructed that, for example, blue t-shirts denoted that the person was an overestimator like them) judged whether the person was 'Like me' or 'Not like me' using the keyboard keys Z and M.

This encouraged participants to focus on the t-shirt colour early on and associate it with either an in-group or out-group. Participants also saw refresher screens during every break to make sure that they did not forget which faces were which.

There were 16 faces in total, 8 of which were in-group members and 8 out-group. Of these, 4 provided valid gaze cues and 4 were invalid, meaning that each participant saw four groups of faces, each composed of 4 identities, two male, two female: In-Group Valid, In-Group Invalid, Out-Group Valid and Out-Group Invalid.

Data analysis

RT filters were applied in the same way as in Experiment 2.1, and in this experiment one participant had to be removed for retaining less than 70% of their original trials.

All data were analysed using linear mixed effects models as in previous chapters. The inclusion of face identity as a factor meant that for RTs and accuracy rates we also generated models that included group as a fixed factor, and models that explored a validity x group interaction to see if gaze cueing effects changed as a result of whether the face was an in-group or out-group member. For RTs, no models would converge until the validity | subject term was removed. No accuracy models would converge until the validity | subject term was removed.

For trustworthiness ratings, the process was largely similar. We again generated a maximum random structure then generated models for each fixed factor individually (time-only, validity-only and group-only models). To investigate interactions, we first compared a 2-factor interaction model (time x validity) with a model that included both factors without an interaction (time + validity). We then explored whether a three-way interaction model fit the data better than a two-way interaction. To do this, we modelled

the two-way interaction and included group as an additional fixed factor (time x validity + group), which we compared with the three-way interaction (time x validity x group). This was done to be sure that any improvement of model fit was not simply due to the inclusion of a third factor. All models included a group | subject error term as well.

All single-factor models except for the group-only model failed to converge until the time | identity, validity | subject and group | subject error terms were removed. The group-only model would not converge with any random slope terms defined, and so this model alone was compared with a simplified null model that included the random intercepts but no defined error slope terms. The two-factor models of time and validity (time +/* validity) would not converge until the validity | subject and group | subject terms were removed. The three-factor models (time +/* validity +/* group) would not converge until the time | identity, time | subject, and validity | subject terms were all removed.

However, we include an important caveat here. As in previous experiments, we included a set of 16 individual faces for participants to learn. In experiments with only validity as a single fixed factor this allows for eight identities in each condition, but the inclusion of minimal group as an orthogonal factor to validity reduces the number of faces in each cell to four. As such, controlling for both stimulus and subject-level variance could lead us vulnerable to a Type II error due to limited power. Indeed, this may be why linear mixed models were reluctant to converge for trustworthiness ratings. A way around this could be to use more stimuli, but the issue of set capacity in this incidental learning has never been explored (see Chapter 8: General Discussion). Instead, while we still present the results of linear mixed effects models, we also report the results of factorial ANOVAs for trustworthiness ratings in the main text of this

chapter, to allow converging statistical techniques to strengthen our interpretation.

5.1.2 Results and Discussion

Gaze-cueing

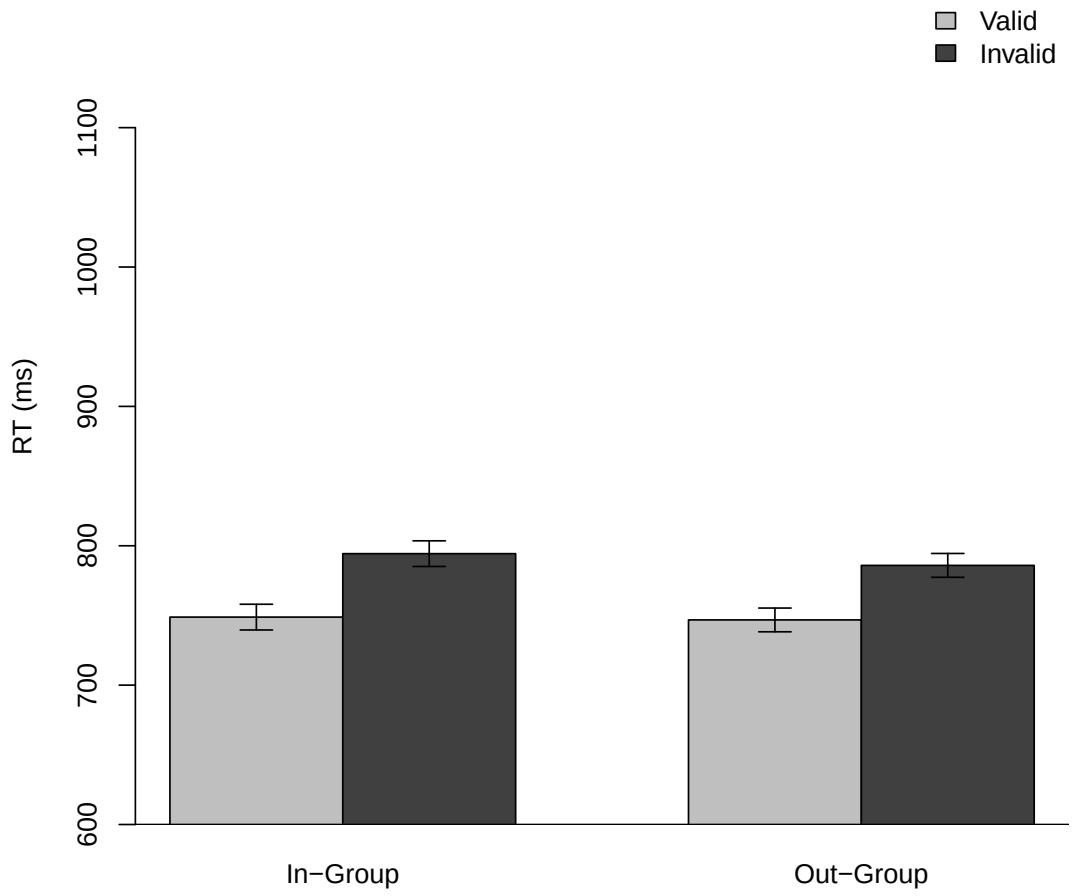

Figure 5.2: Averaged reaction times (milliseconds) in Experiment 5.1 in response to valid (light grey) and invalid (dark grey) trials and in- (left) and out-group (right) faces. Error bars show standard error.

The RT and accuracy results of Experiment 5.1 are shown in Figures 5.2 and 5.3, respectively. Including validity as a fixed factor significantly improved the fit when applied to RTs, indicating a cueing effect (β = -41.78, SE = 7.21, $\chi^2(1)$ = 33.34, p <.001), but including group did not, which suggests that participants were not faster to

respond on trials with faces of either their in- or out-group (β = -5.47, SE = 7.26, $\chi^2(1)$ = 0.57, p = 0.451). As well as this, an interaction of these two factors (validity x race) did not fit the data significantly better than did a model with both factors included without an interaction (β = 6.22, SE = 14.45, $\chi^2(1)$ = 0.19, p = 0.667).

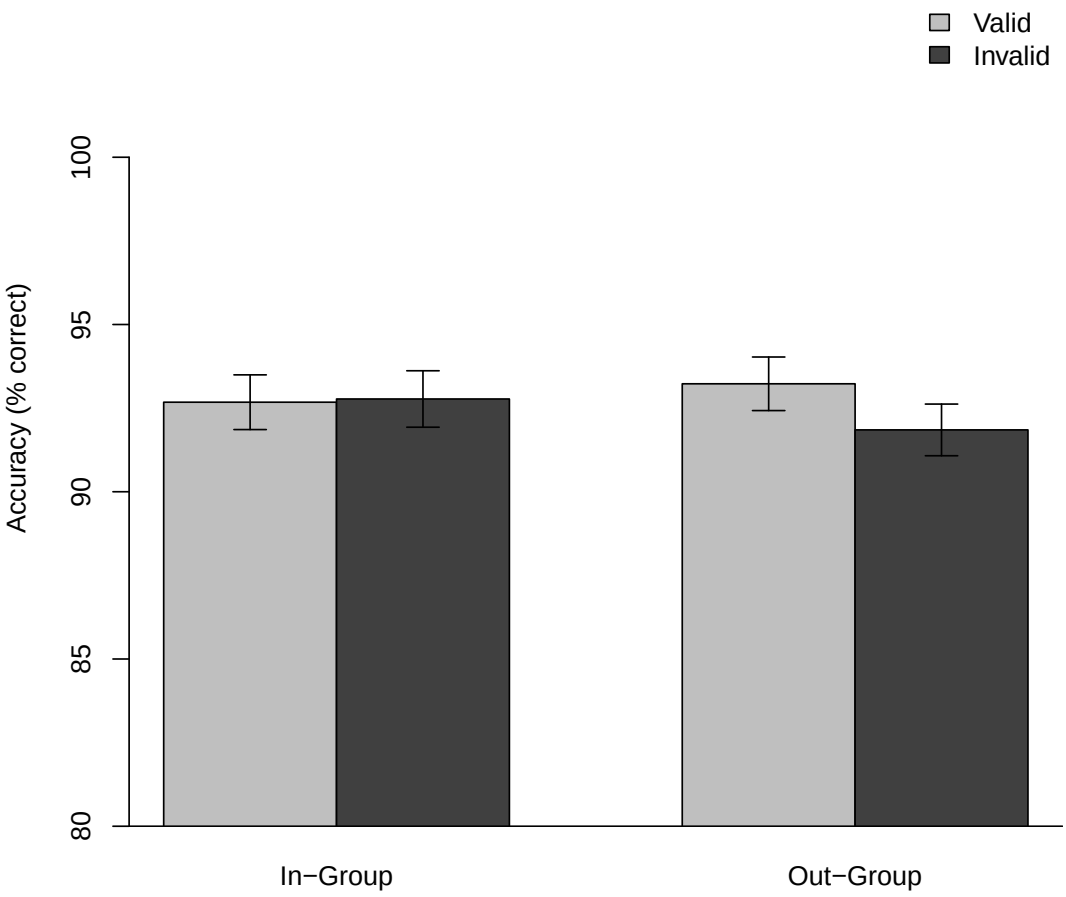

Figure 5.3: Accuracy rates (percent correct) in Experiment 5.1 in response to valid (light grey) and invalid (dark grey) trials and in- (left) and out-group (right) faces. Error bars show standard error.

When modelling accuracy scores, including validity as a fixed factor did not significantly improve the model fit (β = 0.66, SE = 0.70, $\chi^2(1)$ = 0.88, p = 0.347), nor did including group membership (β = -0.17, SE = 0.79, $\chi^2(1)$ = 0.05, p = 0.829), and there was no evidence to support an interaction of the two (β = 1.52, SE = 1.40, $\chi^2(1)$

= 1.18, $p = 0.278$).

Trustworthiness ratings

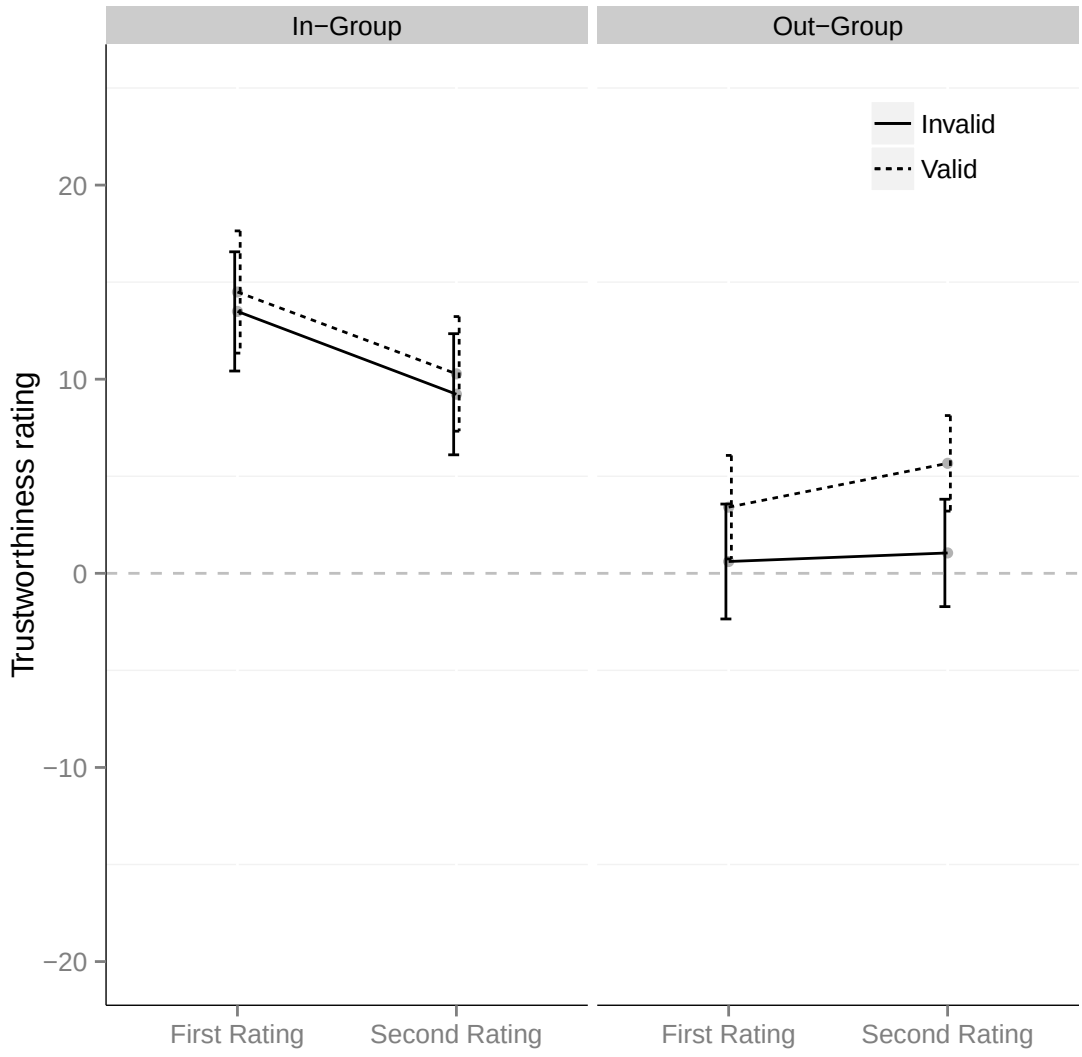

Figure 5.4: Time course of trustworthiness ratings over the course of Experiment 5.1 for valid (dotted) and invalid (solid line) faces for both in-group (left) and out-group (right) members. Error bars show standard error.

Linear mixed effects models. The results of Experiment 5.1 are shown in Figure 5.4. We report the results of linear mixed effects models in the interests of continuity with other experiments, but our confidence in these statistics is weakened by the small number of identities in each condition, as described above. Adding time to the null model did not significantly improve the fit (β = -1.45, SE = 1.91, $\chi^2(1) = 0.58$, p =

0.447), nor did including validity (β = -2.37, SE = 1.82, $\chi^2(1) = 1.69$, $p = 0.194$). However, including group as a fixed factor did significantly improve the fit (β = -9.38, SE = 1.80, $\chi^2(1) = 26.67$, $p < .001$)[1]. The comparison of the two fixed-factor models (time +/* validity) did not reveal a significant two-way interaction (β = -0.98, SE = 2.69, $\chi^2(1) = 0.07$, $p = 0.798$), and neither did the comparison of a three-way interaction fit better than a two-way interaction with group included (time * validity + group vs. time * validity * group), indicating no three-way interaction (β = -4.22, SE = 3.26, $\chi^2(3) = 3.72$, $p = 0.293$).

We ran further analysis of the changes in trustworthiness as a function of time for each condition (In-Group Valid; In-Group Invalid; Out-Group Valid; and Out-Group Invalid) separately. These models found that time did not significantly improve the model fit for In-Group invalid (β = -4.26, SE = 3.45, $\chi^2(1) = 1.53$, $p = 0.216$), Out-Group invalid (β = 0.49, SE = 3.24, $\chi^2(1) = 0.02$, $p = 0.879$), In-Group valid (β = -4.22, SE = 3.38, $\chi^2(1) = 1.56$, $p = 0.211$), or Out-Group valid faces (β = 2.29, SE = 3.22, $\chi^2(1) = 0.51$, $p = 0.474$).

ANOVAs. As well as mixed effects models we also report the results of a factorial ANOVA, given that controlling for both subject and identity in linear mixed effects models may unacceptably raise the risk of a Type II error. A 2x2x2 ANOVA looking at time, validity and group membership found no main effect of time ($F(1,29) = 0.57$, $p = 0.455$, $\eta_P^2 = 0.02$), or of validity ($F(1,29) = 1.36$, $p = 0.253$, $\eta_P^2 = 0.04$), but a marginal effect of group membership ($F(1,29) = 3.95$, $p = 0.057$, $\eta_P^2 = 0.12$), which was driven by the fact that in-group members were rated as more trustworthy than out-group members. No significant interactions emerged (all Fs < 1.4).

[1]Note that this model was compared with a simplified null model with no random terms, to allow for direct comparison

Follow-up paired-samples t-tests between the beginning and end ratings for each condition found that there was no significant change in trustworthiness over the course of the experiment for valid in-group ($t(29) = 1.02$, [95% CI -4.26 to 12.70], $p = 0.318$, $d = 0.19$), valid out-group ($t(29) = -0.68$, [95% CI -9.09 to 4.57], $p = 0.504$, $d = 0.10$), invalid in-group ($t(29) = 1.01$, [95% CI -4.41 to 12.95], $p = 0.323$, $d = 0.18$), or invalid out-group faces ($t(29) = -0.15$, [95% CI -6.30 to 5.42], $p = 0.879$, $d = 0.02$).

The results of Experiment 5.1 suggest that participants identified with their in-group over their out-group, despite the fact that group membership was minimal at best, and the cover story described an artificial group distinction that does not emerge often in the real world. What is particularly striking is that participants experience the same gaze cueing paradigm as in previous studies and that gaze shifts in the in- and out-group faces evoke the same orienting of attention, as the cueing effects are strikingly similar. Nevertheless, when participants are asked to make trustworthiness judgements in this experiment they do not appear to incorporate this information into their judgements, and instead default to judging trustworthiness according to which group individuals belong.

Of course, an alternative explanation is that the presence of blue and yellow t-shirts during trustworthiness ratings was distracting and so participants simply forgot the cueing behaviour of the faces. This explanation seems unlikely, but in order to rule it out we replicated this experiment but removed all references to there being two groups. Experiment 5.2 removed the instruction screen categorising participants as over/underestimators in the calibration task, and removed the face categorisation task at the beginning of the experiment. As such, this was more closely matched to Experiment 2.1, with the addition of a visual estimation task and the inclusion of blue and yellow

shirts, which were not mentioned or highlighted at any point.

5.2 Experiment 5.2

This experiment replicated Experiment 5.1 but removed any reference to dichotomous groups, to see if this addition had prompted participants to represent the trustworthiness of faces on a group level as opposed to an identity level.

5.2.1 Methods

Participants

30 participants (22 female, $M_{age} = 19.79$, $s.d. = 2.20$) volunteered for this study in return for a mixture of course credit and payment. No participants were removed after RT filters were applied.

Stimuli, Design and Procedure

This experiment was identical to Experiment 5.1 in every way except that participants were not allocated to any groups − during the 'calibration' task where participants were asked to estimate the number of moving dots, they were not given any feedback as to whether they had been classed as an overestimator or underestimator, and faces were similarly not explicitly identified as belonging to either of these groups. Participants also no longer completed any 2AFC task at the beginning of the experiment where they identified the group membership of the faces on the basis of shirt colour. As such, in this experiment shirt colour was an incidental background feature that was never explicitly mentioned, as opposed to Experiment 5.1, where it was a salient cue to group membership.

Data analysis

RT filters were applied in the same way as in Experiment 5.1, and in this experiment no participants had to be removed for retaining less than 70% of their original trials. In this experiment, there was no in-group/out-group distinction and so these data were instead analysed using shirt colour as an additional fixed factor alongside validity for RT and accuracy models, and using shirt colour, validity and time as fixed factors in trustworthiness models. We did not expect any effects of shirt colour, but include it in the interests of consistency with Experiment 5.1.

All RT models converged with the maximum random structure defined. The colour-only model of accuracy rates would not converge until the colour | subject term was removed, so this was removed from this, the null, and the validity-only models. The two-factor models of accuracy rates would not converge until the validity | subject term was removed.

For trustworthiness models, all single-factor models except for the group-only model failed to converge until the time | identity, validity | subject and shirt colour | subject error terms were removed. The shirt colour-only model would not converge with any random slope terms defined, and so this model alone was compared with a simplified null model that included the random intercepts but no defined error slope terms. The interaction model of time and validity (time * validity) would not converge until the time | subject, time | identity, and shirt colour | subject terms were removed, so these were removed from both two-factor models. The three-factor models (time +/* validity +/* shirt colour) would not converge until the time | identity, time | subject, and validity | subject terms were all removed.

Trustworthiness ratings were also analysed using factorial ANOVAs, which are

reported in the main text as with Experiment 5.1.

5.2.2 Results and Discussion

Gaze-cueing

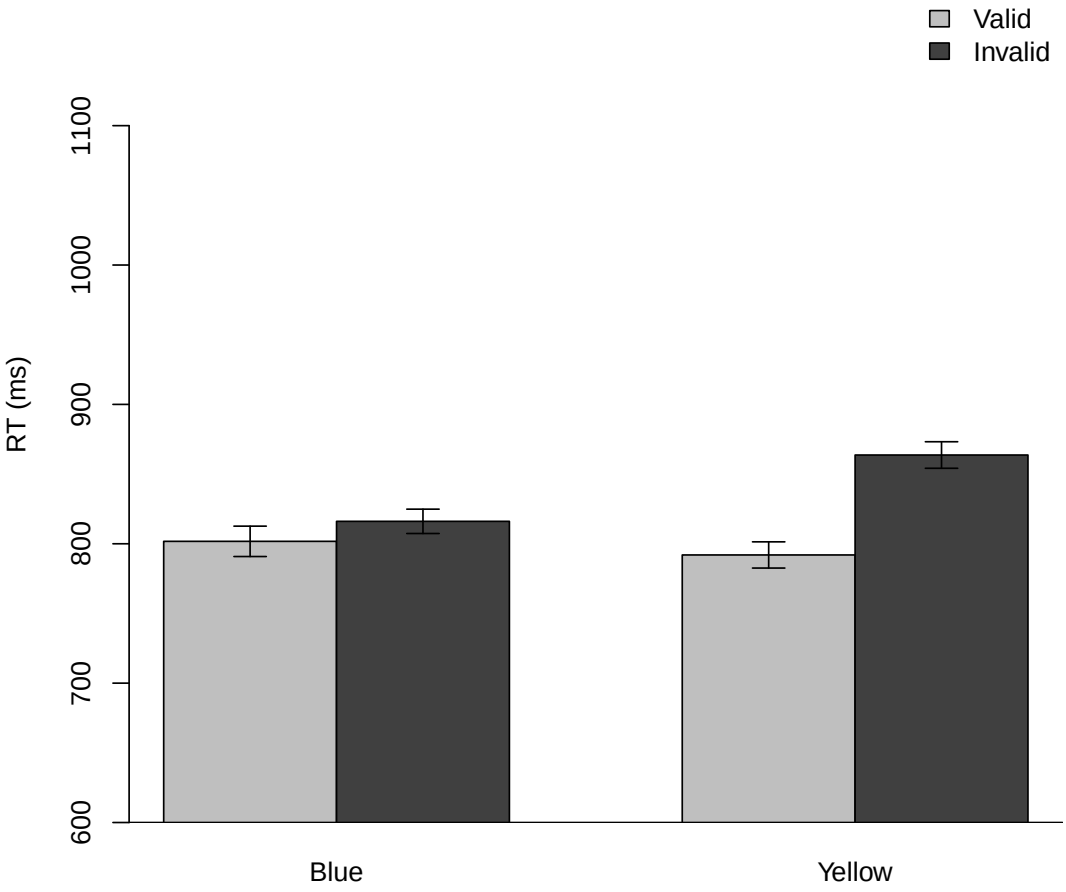

Figure 5.5: Averaged reaction times (milliseconds) in Experiment 5.2 in response to valid (light grey) and invalid (dark grey) trials and faces wearing blue (left) and yellow (right) shirts. Error bars show standard error.

The RT and accuracy results of Experiment 5.2 are shown in Figures 5.5 and 5.6, respectively. Including validity as a fixed factor significantly improved the fit when applied to RTs, indicating a cueing effect (β = -42.85, SE = 9.02, $\chi^2(1) = 17.38$, p <.001), as did including shirt colour ($\beta = 58.58$, $SE = 18.12$, $\chi^2(1) = 4.28$, $p = 0.039$).

Fitting a two-way interaction model fit the data significantly better than a two-factor model without an interaction (β = -57.15, SE = 15.70, $\chi^2(1)$ = 13.21, p <.001). This appears to be driven by the fact that cueing was slightly stronger for faces in yellow shirts than blue, but it is not clear why this happened and it is difficult to say what impact, if any, this result may have on our interpretation.

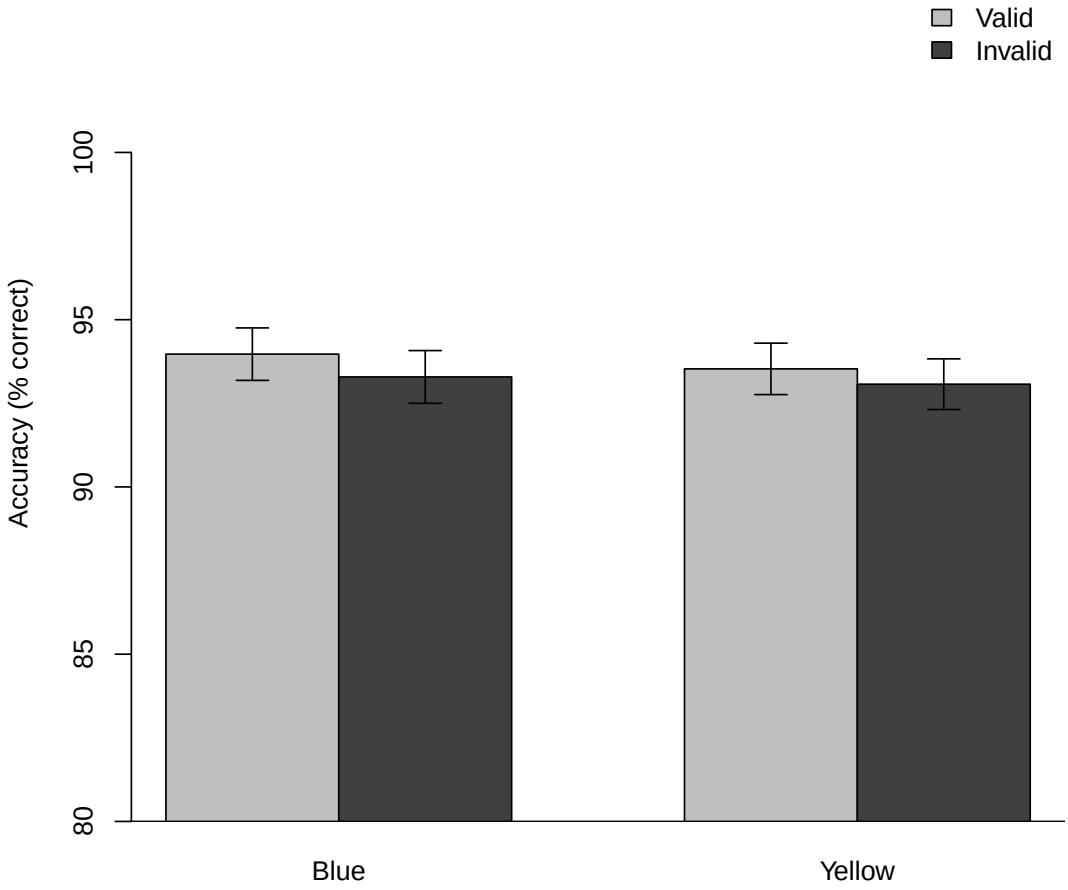

Figure 5.6: Accuracy rates (percent correct) in Experiment 5.2 in response to valid (light grey) and invalid (dark grey) trials and faces wearing blue (left) and yellow (right) shirts. Error bars show standard error.

When modelling accuracy scores, including validity as a fixed factor did not significantly improve the model fit (β = 0.55, SE = 0.66, $\chi^2(1)$ = 0.70, p = 0.404), nor did including shirt colour (β = -0.32, SE = 0.66, $\chi^2(1)$ = 0.23, p = 0.631), and there

was no evidence to support an interaction of the two (β = -0.26, SE = 1.34, $\chi^2(1)$ = 0.04, p = 0.846).

Trustworthiness ratings

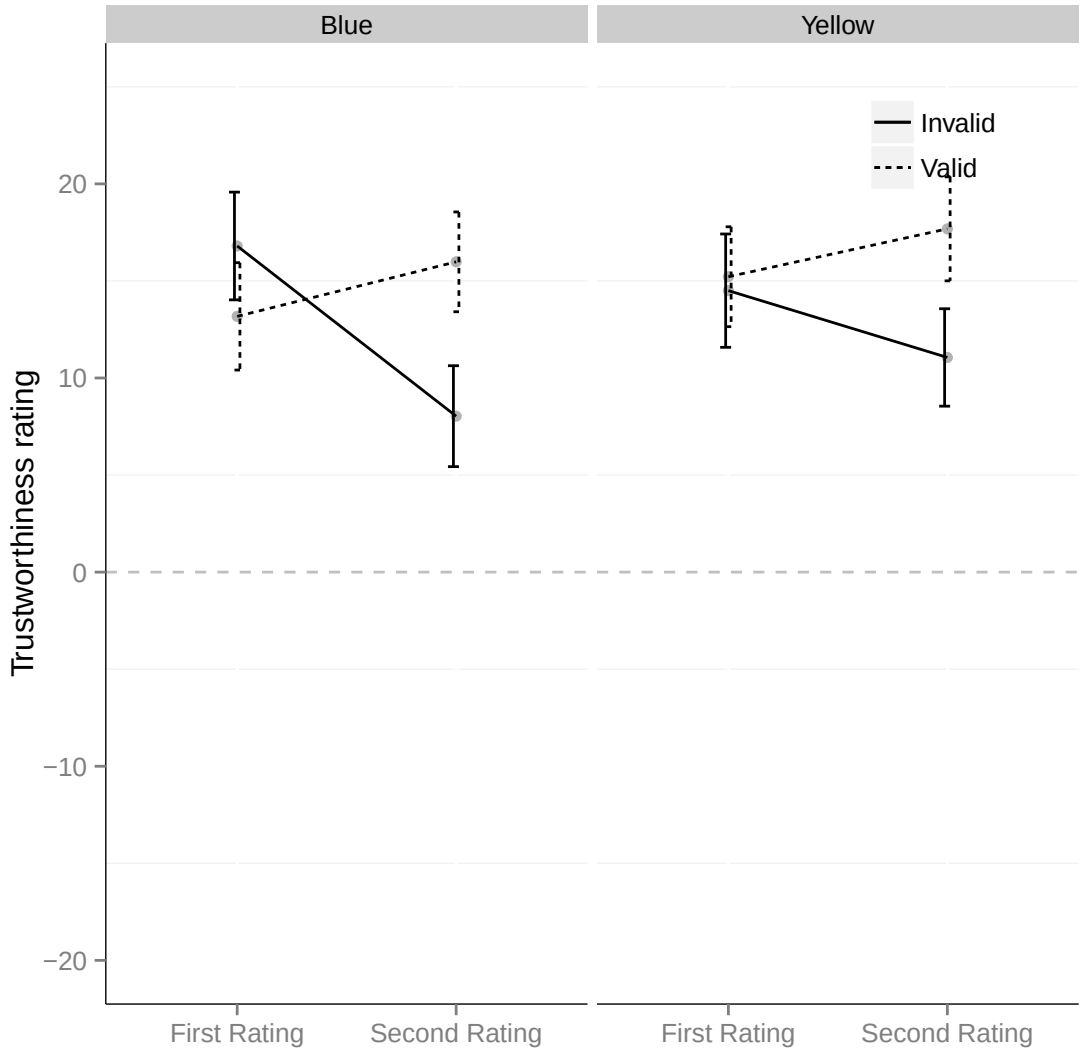

Figure 5.7: Time course of trustworthiness ratings over the course of Experiment 5.1 for valid (dotted) and invalid (solid line) faces for and faces wearing blue (left) and yellow (right) shirts. Error bars show standard error.

Linear mixed effects models. The results of Experiment 5.2 are shown in Figure 5.7. We report the results of linear mixed effects models in the interests of continuity with other experiments, but our confidence in these statistics is weakened by the small number of identities in each condition, as described above. Adding time to the null model did

not significantly improve the fit (β = -1.73, SE = 1.66, $\chi^2(1) = 1.11$, $p = 0.291$), nor did including shirt colour (β = 1.12, SE = 1.61, $\chi^2(1) = 0.48$, $p = 0.488$). Including validity did marginally improve the model fit (β = -2.92, SE = 1.61, $\chi^2(1) = 3.30$, $p = 0.069$).

The comparison of the two fixed-factor models (time +/* validity) revealed a significant two-way interaction (β = -8.74, SE = 3.18, $\chi^2(1) = 7.55$, $p = 0.006$), but including a three-way interaction did not significantly improve the fit over a two-way interaction with shirt colour as an additional factor (time * validity + shirt colour vs. time * validity * shirt colour), indicating that trust learning was not modulated by shirt colour (β = 5.67, SE = 6.34, $\chi^2(3) = 1.44$, $p = 0.697$).

We ran further analysis of the changes in trustworthiness as a function of time for each condition (Blue Valid; Blue Invalid; Yellow Valid; Yellow Invalid) separately.[2] These models found that time did not significantly improve the model fit for Blue Valid (β = 2.81, SE = 3.19, $\chi^2(1) = 0.78$, $p = 0.378$), or for Yellow Valid (β = 2.47, SE = 3.07, $\chi^2(1) = 0.65$, $p = 0.420$), or Yellow Invalid faces (β = -3.44, SE = 3.19, $\chi^2(1) = 1.17$, $p = 0.278$), but it did improve the fit for Blue Invalid (β = -8.77, SE = 3.36, $\chi^2(1) = 6.74$, $p = 0.009$).

ANOVAs. As well as mixed effects models we also report the results of a factorial ANOVA, given that controlling for both subject and identity in linear mixed effects models may unacceptably raise the risk of a Type II error. A 2x2x2 ANOVA looking at time, validity and shirt colour found no main effect of time ($F(1,29) = 1.10$, $p = 0.304$, $\eta^2_P = 0.04$), or of validity ($F(1,29) = 1.05$, $p = 0.314$, $\eta^2_P = 0.04$), or shirt colour ($F(1,29) = 0.28$, $p = 0.599$, $\eta^2_P = 0.01$). The interaction of time and validity approached significance ($F(1,29) = 3.83$, $p = 0.060$, $\eta^2_P = 0.12$), but no other interactions did (all Fs

[2] No models for blue faces would converge with any random terms defined, while models for yellow faces converged after the time | subject term was removed.

< 1.23).

Follow-up paired-samples t-tests between the beginning and end ratings for each condition found that there was no significant change in trustworthiness over the course of the experiment for Blue Valid faces ($t(29)$ = -0.82, [95% CI -9.78 to 4.17], p = 0.417, d = 0.12), Yellow Valid faces ($t(29)$ = -1.14, [95% CI -6.89 to 1.96], p = 0.264, d = 0.11), Yellow Invalid faces ($t(29)$ = 1.00, [95% CI -3.60 to 10.49], p = 0.326, d = 0.14), but there was a significant change for Blue Invalid faces ($t(29)$ = 2.26, [95% CI 0.83 to 16.70], p = 0.031, d = 0.35).

In all, it seems that Experiment 5.2 has managed to overcome some of the extinction of trust learning seen in Experiment 5.1, although it seems weaker than the learning seen in previous experiments. While there were some differences between learning for blue and yellow faces, these were not so strong as to indicate a three-way interaction. This is likely due to the fact that the presence of different shirt colours still serves as a cue to discrete face categories, and although participants are not explicitly instructed to think of these faces as different groups, enough of them may spontaneously default to this representation that it weakens the classic profile of trust learning, which leaves them less likely to use individuals' behaviour to inform trustworthiness judgements.

5.3 Chapter Discussion

This chapter investigated how the identity of the cueing faces may affect trust learning, by assigning face identities to one of two minimal groups. Although we expected that this manipulation may have some selective effects (i.e. enhancing or disrupting trust learning for a particular type of face, depending on condition), Experiment 5.1 instead found no learning of trust from gaze cues for any type of face. Instead, participants seem to use

the group membership of the identities as a heuristic in their social decision making, as group category appears to drive their decisions more than individual gaze behaviour.

Learning trust from gaze cues is demanding, and results of previous studies suggest that it occurs automatically and outside of conscious awareness (c.f. Experiment 2.4). However, the results of the current chapter suggest that in the presence of an alternative source of information − explicitly instructed group membership − participants may default to representing faces in terms of more easily accessible information that they can use to inform trustworthiness judgements. It may be that learning on the basis of experience with individuals is more cognitively demanding, and so the current paradigm is insufficient to override the salience of group membership.

Further research may look to explore the reasons for this. For example, it is not clear from the current study whether participants still learn about trustworthiness from gaze cues but do not use this information later in their social judgements, or whether it is learning itself that is inhibited during gaze cueing. Similar cueing effects for both in- and out-group members seem to indicate the former interpretation, but follow-up studies could use alternative measures to interrogate this. Manssuer, Roberts and Tipper (2015) found neural correlates of incidental trust learning during gaze cueing (that is, while participants are learning about the cueing validity of the faces) using EEG, and so this may be a way of exploring whether it is learning or retrieval that is affected by minimal group membership. If these neural signals are preserved when groups are present, then this suggests that learning may occur but group membership overrides it at retrieval. On the other hand, if these neural correlates are disrupted by group membership then this suggests that it is specifically learning that is disrupted.

If the explanation behind these findings is that trust learning is absent because

either forming or accessing these memories for individual behaviour is more cognitively demanding, it could be interesting to see what happens when the typical gaze-cueing paradigm used throughout this thesis is replicated under high cognitive load. If this learning is more demanding (therefore higher in intrinsic cognitive load) then increasing the demand on the cognitive system should extinguish trust learning in a non-social context. However, if learning persists, this would suggest that this interference is due to higher-order social interference from the minimal groups that is specific to this paradigm.

The results of Experiment 5.2 suggest that some participants at least are still resorting to group-level representations. Clothing choices are, after all, a salient indicator as to an individual's beliefs and opinions in the real world that reflect an explicit and conscious decision (sports team colours, for example). On the other hand, there are some group distinctions that are independent of personal choice or preference, such as race, and as these occur naturally in the real world it may be that these are subtle enough manipulations that they do not trigger participants to represent them as explicit groups, but rather treat them more implicitly. Chapter 6 explores this question in further detail.

In conclusion, this chapter reports the results of three experiments that show that when participants are explicitly instructed that faces during gaze cueing belong to different, albeit minimal, groups, they default to representing these identities on the basis of their group-level characteristics rather than their individual history of gaze behaviour. This suggests that although incidental trust learning from gaze cues is not consciously driven, it may be superseded by concurrent information such as heuristics about group membership.

Chapter 6. The effect of real-world group membership on incidental trust learning

When using an experimental paradigm to explicitly change participants' expectations of a face (e.g. through a minimal groups paradigm, where each identity is represented as a trusted in-group member or a distrusted out-group member, see Chapter 5), this group-level representation appears to override learning of individual identities, as trustworthiness ratings reflect group membership but not learned information about that face's gaze behaviour.

This paradigm, however, relies on an explicit manipulation − that is, that faces are labelled as part of a superficial in-group or out-group based on nothing more than shirt colour. Although these group categories do exist in the real world (sports teams being an example), there are also other ways in which our expectations and ideas about others may vary in a more naturalistic way, such as the physical features of a face that reflect identity. Looking at a group membership category that occurs naturally and is driven by face physical features would therefore allow us to investigate this without explicitly telling participants about the manipulation.

One of the most studied and historically important social group categories, and one that is driven primarily by facial features, is that of race. Aside from showing preferential biases towards own-race over other-race faces (Dasgupta, Mcghee, Greenwald & Banaji, 2000), people are also better at recognising the posed emotions of own-race than other-race faces (Elfenbein & Ambady, 2002b, 2002a), and remember own-race faces better than other-race faces (Meissner & Brigham, 2001). There is also evidence that these own-race biases are linked to decisions about trustworthiness in both explicit ratings and economic games (Stanley, Sokol-Hessner, Banaji & Phelps, 2011).

In this chapter we include faces of different races in order to create more naturalistic

group categories. There are competing predictions. One possibility is that learning about other-race faces will be impaired. This would fit with previous research suggesting that we remember individuals of our own race better than those of other races (Meissner & Brigham, 2001; Ng & Lindsay, 1994; Slone, Brigham & Meissner, 2000), and that emotional learning about other-race faces is less flexible than learning about own-race faces (Dunsmoor, Kubota, Li, Coelho & Phelps, 2016). The trust learning task requires that gaze patterns are associated with a specific face identity (see Chapter 3), and hence if other-race identities are less well represented, learning will be impaired.

However, basing our behavioural decisions on physiognomy and race can be harmful, leading to rigid responses that reinforce discrimination. In addition, it can render us incapable of responding to the dynamics of behaviour and adjusting our behaviour towards others in light of incoming information. Therefore we might expect learning not to be less strong for other-race faces, but learning may instead be influenced by those events that offer opportunities for updating incorrect models or expectations.

Some studies suggest that people generally hold an expectation for own-race faces to cooperate more than other-race faces (Stanley et al., 2011). As such, an alternative hypothesis is that trust learning will follow a Rescorla and Wagner (1972) pattern of learning, where learning is instead tailored to surprising events − in this case, invalid same-race faces and valid other-race faces.

In Experiment 6.1 we explore whether trust learning is the same for individuals who differ in terms of racial group membership in a behavioural paradigm. To this end, we use images of White (in-group) and East Asian (out-group) faces in the cueing paradigm and ask participants to rate them in terms of trust at the beginning and end of the experiment. A critical contrast to the minimal group procedure of the previous chapter

is that the two categories of faces are never explicitly mentioned in the current study.

6.1 Experiment 6.1

6.1.1 Methods

Participants

In Experiment 6.1a there were 30 participants in total. In Experiment 6.1b there were a further 33 participants. One participant was excluded because Experiment 6.1b involved EMG recordings of which participants needed to be naïve to the purpose, and this participant indicated at the end of the experiment that they had completed EMG studies before. Another participant was excluded after applying filters (i.e. for having pre-ratings that were exceptionally skewed towards the ends of the rating scale, which would have created a ceiling or floor effect in trustworthiness ratings − more details are given below). Finally, one participant was removed on the basis that their EMG recordings were too noisy to extract meaningful data, so the final number for analysis in Experiment 6.1b was 30.

These two experiments were then collapsed together to give a total of 60 participants for analysis (all female, all Caucasian, $M_{age} = 20.14$, $s.d. = 1.59$). All participants provided written consent and the study was given ethical approval by the Departmental Ethics Committee of the University of York Psychology Department.

Stimuli

Stimuli were taken from the MR2 face database, a multi-racial high-resolution database of facial stimuli (Strohminger et al., 2015). This database comes with a set of ratings for each face on a range of attributes, including trustworthiness on a scale of 1

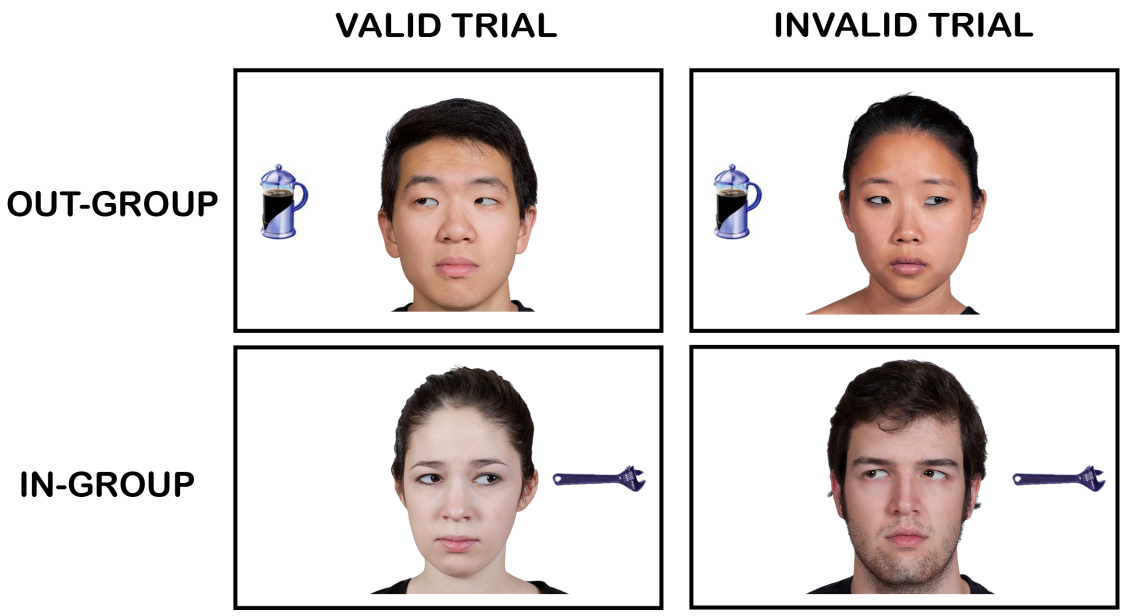

Figure 6.1: Examples of the four different conditions in which faces were presented in Experiment 6.1: out-group valid, out-group invalid, in-group valid and in-group invalid.

(untrustworthy) to 7 (trustworthy), and these are publicly available on the Open Science Framework (OSF; https://osf.io/uwk4v/). 8 East Asian faces and 8 White faces were selected on the basis of these ratings to be similar in terms of apparent trustworthiness (East Asian faces: $M = 4.13$, $s.d. = 0.14$; White faces: $M = 4.12$, $s.d. = 0.21$). Examples of each of the four conditions in which faces could appear are shown in Figure 6.1.

These images were then edited in Adobe Photoshop CS6 to remove the grey background and edit the direction of eye gaze to create three versions of each face: straight, left, and right gaze. These stimuli were used in the gaze-cueing procedure, while faces with eyes unedited were used in the trustworthiness ratings and one-back procedure.

Design and Procedure

The object categorisation task that participants were asked to complete was the same as in previous experiments and chapters; a face appeared in the centre of the screen

maintaining direct gaze for 1,500ms and would look either left or right for 500ms, followed by the target object that participants had to identify as either a kitchen or garage item, which would stay for 2,500ms, followed by a blank screen (500ms in Experiment 6.1a; 2,000ms in Experiment 6.1b). The face would then return to direct gaze for 1,000ms. As in previous experiments, participants received feedback in the form of an error tone for incorrect trials. At the beginning and the end of the experiment, participants completed trustworthiness ratings of each of the faces.

Experiment 3.3 has shown that trust learning is stronger when participants complete a familiarisation task at the beginning of the experiment. In previous experiments, this has involved recognising identity across changes in viewpoint and expression. The MR2 database does not provide any such variation, and so in this experiment we included a one-back recognition task, where faces were presented in sequence and participants had to respond with the SPACE bar if they saw the same face repeated twice in a row. This encourages participants to encode details of the faces and store them in working memory, at least until the next face is shown, and with repeated exposures this should allow participants to become familiar with the face identities.

Experiment 6.1a was an entirely behavioural experiment with the same timings as previous chapters. Experiment 6.1b was identical to Experiment 6.1a with the exception that electromyographic (EMG) recordings were taken from the corrugator supercilii and zygomaticus major and the blank screen between trustworthiness rating trials was shown for 2,000ms rather than 500ms, to allow the EMG signal to recover. These recordings did not yield any significant effects of race or validity during gaze cueing or trustworthiness ratings, and so are not reported here (EMG parameters and results are available in Appendix E). However, given that the two versions of the experiment were

otherwise identical (and preliminary analysis found no significant difference between the two versions), results were collapsed across the two versions of Experiment 6.1, yielding a total of 60 participants.

Data analysis

As in previous experiments, before data were analysed participants' responses were filtered to remove all error trials (where participants reported the incorrect answer) and RT outliers − RTs below 250ms (too short to process the stimuli) and above 2,500ms (indicating that participants had not given a response in the allotted time). The number of remaining trials was then compared with the original number of trials to check that all participants retained at least 70% of their total trials and had not scored below 70% total correct on any one condition.

As well as RT filters, we also examined participants' pre-ratings. Participants' ratings to in-group and out-group faces at the beginning of the experiment were averaged and examined to ensure that the average for neither group exceeded 70 on the 100-point scale in either direction. This was done because an average to one-group that exceeded 70 suggested that participants gave ratings to multiple faces that used the far ends of the scale before any trustworthiness induction was performed, and in a paradigm where there were two 8-member groups of faces at the beginning of the experiment (as opposed to one 16-member group of faces as in previous experiments), we felt that this may unduly affect our results if these participants were left in. One participant was removed on this basis in Experiment 6.1b.

All data were analysed using linear mixed effects models as in previous chapters, with the exception that experiment (EMG/no EMG) was included as a random factor.

As in Experiment 5.1, the inclusion of face identity as a factor meant that for RTs and accuracy rates we also generated models that included race as a fixed factor, and models that explored a validity x race interaction to see if gaze cueing effects changed as a result of the cueing face's race. All RT models converged with the maximum random structure except for the race-only model, which would not converge until the experiment | identity term was removed, and so this was removed from all single-factor models to allow for direct comparison.

For accuracy models, the validity-only would not converge with experiment | identity as a term, so this was removed from all models for the purposes of comparison. When this was removed, the null model would not converge until the validity | subject term was also removed from this model only. The two-way interaction model would not converge until both the experiment | identity and validity | subject terms were removed, so these were removed from both two-factor models to allow for direct comparison.

Regarding analysis of trustworthiness ratings, we report the results of linear mixed effects models in light of consistency with other chapters. We generate a maximum random structure model that we compare with time-only, validity-only and race-only models separately. As regards the interaction of these factors, we first compared a 2-factor interaction model (time x validity) with a model that included both factors without an interaction (time + validity). We then explored whether a three-way interaction model fit the data better than a two-way interaction. To do this, we modelled the two-way interaction and included race as an additional fixed factor (time x validity + race), which we compared with the three-way interaction (time x validity x race). This was done to be sure that any improvement of model fit was not simply due to the inclusion of a third factor.

However, as in Experiment 5.1, we have a caveat about interpreting the results of these models. In experiments with only validity as a single fixed factor this allows for eight identities in each condition, but the inclusion of race as an orthogonal factor to validity reduces the number of faces in each cell to four. As such, controlling for both stimulus and subject-level variance leaves us vulnerable to a Type II error. This is somewhat evident by the fact that no model of trustworthiness rating would converge until most of the maximum random structure slope terms were removed, leaving only time | identity as a random term. As such, in this chapter we also report the results of repeated measures ANOVAs for trustworthiness ratings.

6.1.2 Results and Discussion

Gaze-cueing

The RT and accuracy results of Experiment 6.1 are shown in Figures 6.2 and 6.3, respectively. Including validity as a fixed factor significantly improved the fit when applied to RTs, indicating a cueing effect (β = -29.84, SE = 5.93, $\chi^2(1) = 24.58$, p <.001), but including race did not, which suggests that participants were not faster to respond on trials with faces of a particular race (β = -6.07, SE = 5.90, $\chi^2(1) = 1.06$, p = 0.304). As well as this, an interaction of these two factors (validity x race) did not fit the data significantly better than did a model with both factors included without an interaction (β = 3.46, SE = 11.78, $\chi^2(1) = 0.09$, $p = 0.769$).

When modelling accuracy scores, including validity as a fixed factor did not significantly improve the model fit (β = -0.72, SE = 0.53, $\chi^2(4) = 3.73$, $p = 0.444$), nor did including race (β = 0.03, SE = 0.67, $\chi^2(4) = 1.84$, $p = 0.766$), and there was no evidence to support an interaction of the two (β = -1.13, SE = 0.97, $\chi^2(1) = 1.36$, p =

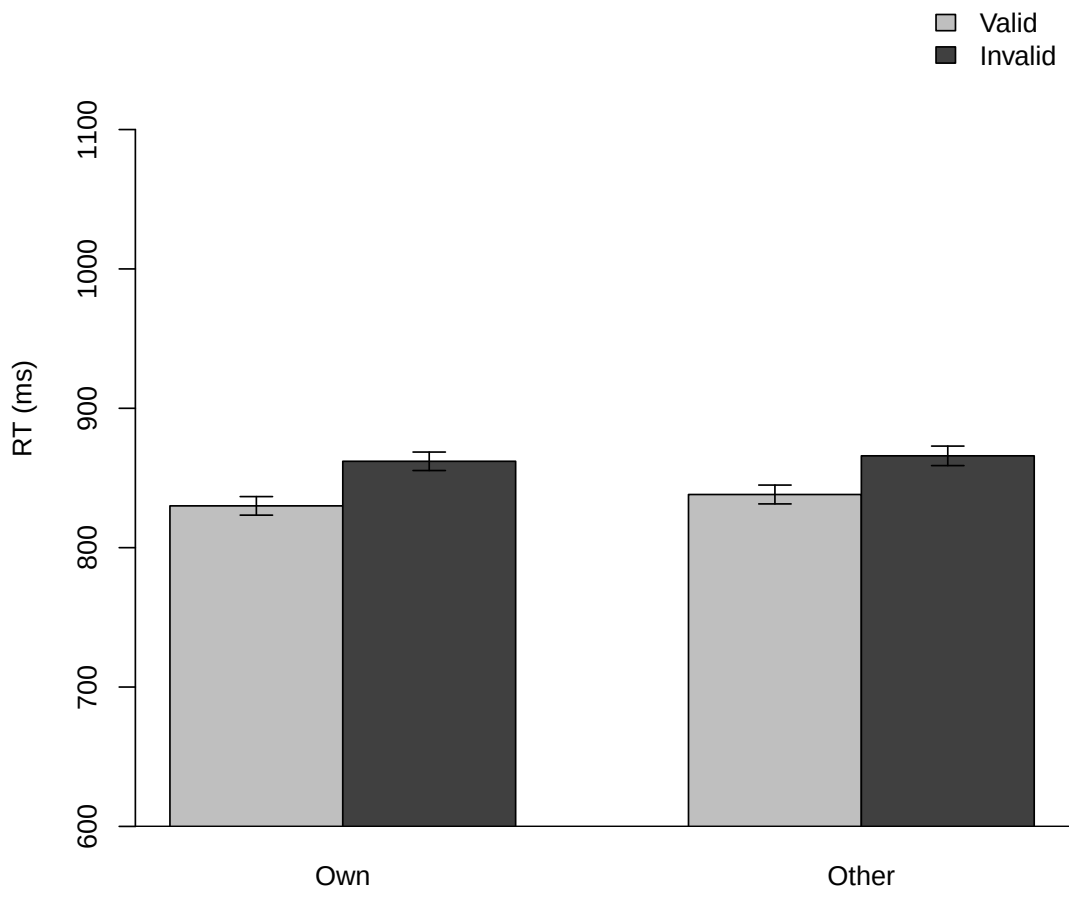

Figure 6.2: Averaged reaction times (milliseconds) in Experiment 6.1 in response to valid (light grey) and invalid (dark grey) trials and own (left) and other (right) faces. Error bars show standard error.

0.244).

Trustworthiness ratings

Linear mixed effects models. We report the results of linear mixed effects models in the interests of continuity with other experiments, but as in Chapter 5 our confidence in these statistics is weakened by the small number of identities in each condition. The trustworthiness ratings in Experiment 6.1 are shown in Figure 6.4. Adding validity to the model significantly improved the fit ($\beta = 9.72$, $SE = 1.48$, $\chi^2(1) = 42.82$, $p < .001$),

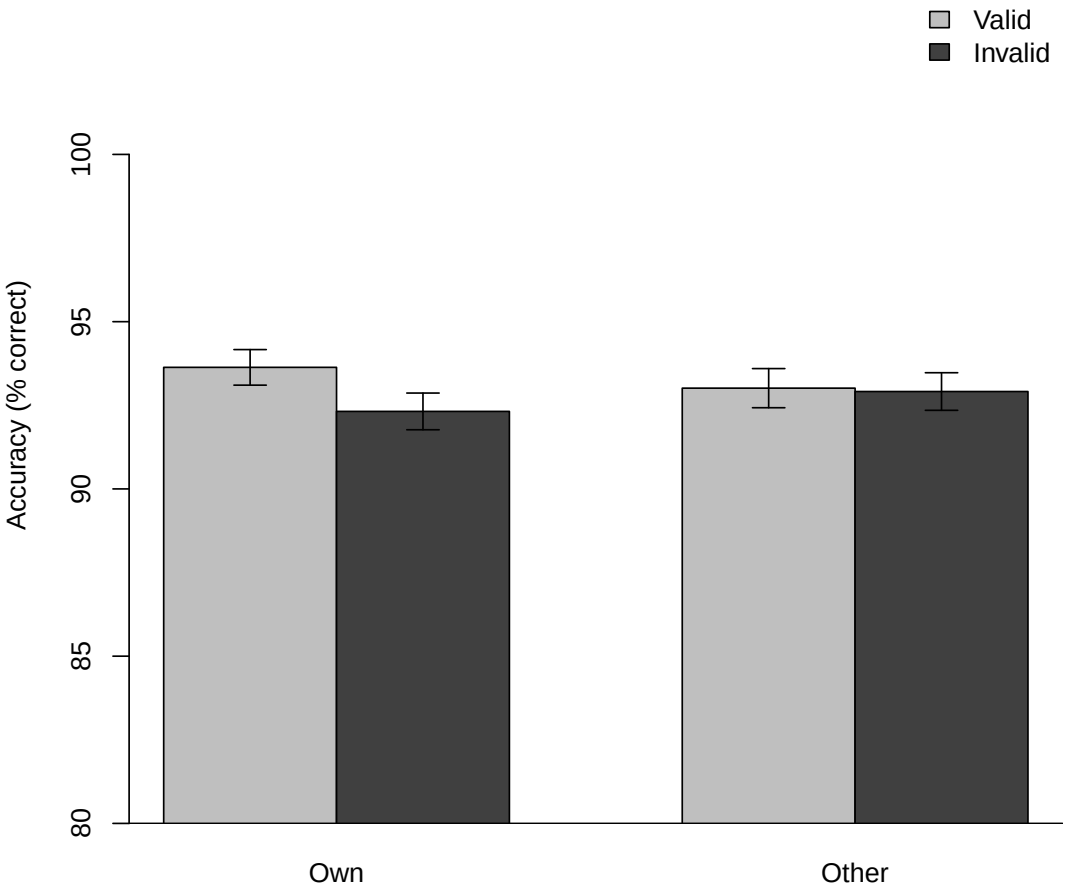

Figure 6.3: Accuracy rates (percent correct) in Experiment 6.1 in response to valid (light grey) and invalid (dark grey) trials and own (left) and other (right) faces. Error bars show standard error.

as did adding time (β = -4.47, SE = 1.77, $\chi^2(1) = 5.85$, $p = 0.016$). The interaction model of these factors (time x validity) fit the data significantly better than when both factors were modelled but without an interaction ($\beta = 19.95$, $SE = 2.92$, $\chi^2(1) = 46.31$, $p <.001$).

Including race in the null model did not significantly improve the fit ($\beta = 5.75$, $SE = 4.55$, $\chi^2(1) = 1.68$, $p = 0.195$), and modelling a three-way interaction of time x validity x race did not fit significantly better than modelling time x validity with race as a separate factor ($\beta = 9.54$, $SE = 5.83$, $\chi^2(3) = 4.30$, $p = 0.231$).

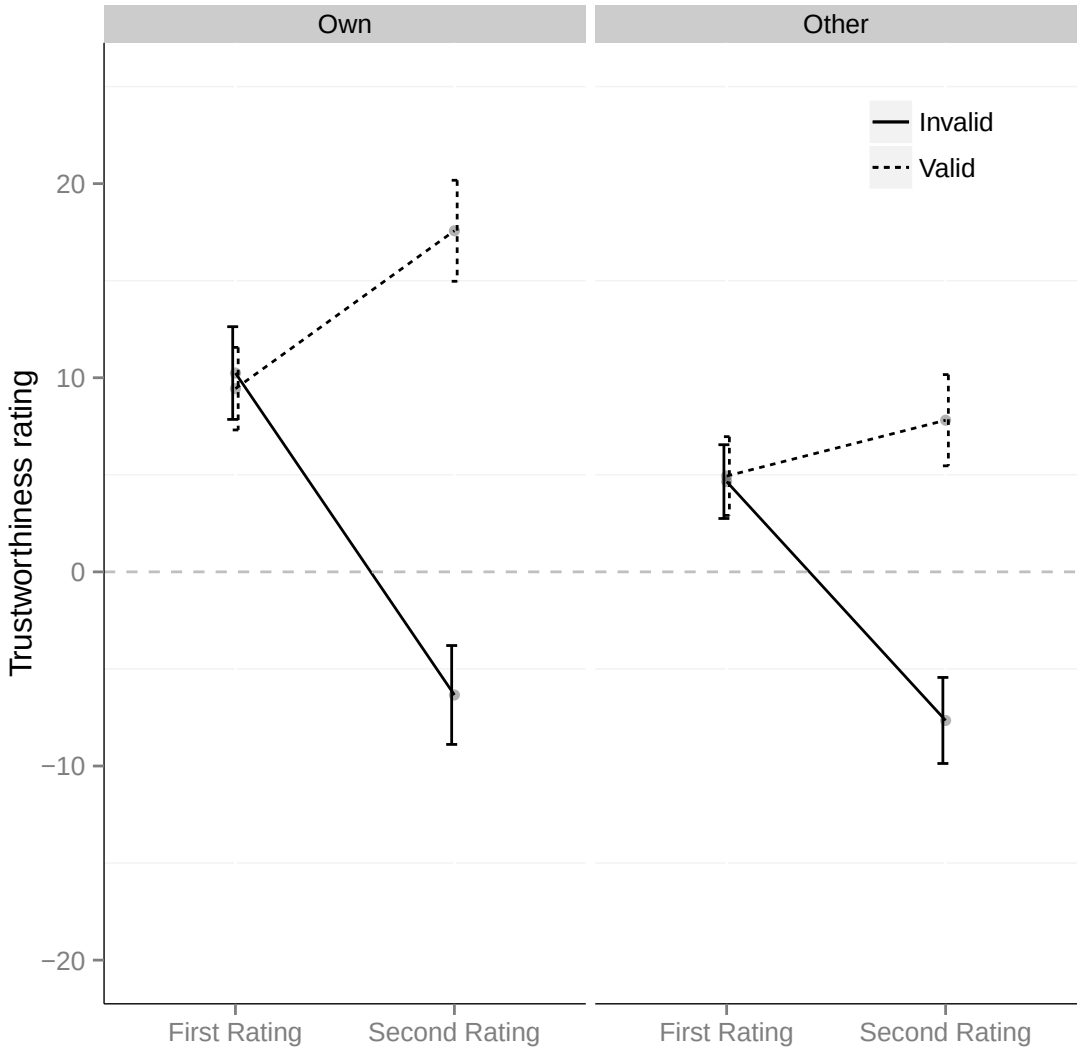

Figure 6.4: Time course of trustworthiness ratings over the course of Experiment 6.1 for valid (dotted) and invalid (solid line) faces for both in-group (left) and out-group (right) members. Error bars show standard error.

We ran further analysis of the changes in trustworthiness as a function of time for each condition (Own Race Valid; Own Race Invalid; Other Race Valid; and Other Race Invalid) separately. These models found that time significantly improved the model fit for Own Race invalid (β = -16.58, SE = 3.14, $\chi^2(1) = 13.76$, p <.001), Other Race invalid faces (β = -12.30, SE = 2.77, $\chi^2(1) = 19.32$, p <.001)[1], and for Own Race valid faces (β = 8.13, SE = 3.78, $\chi^2(1) = 4.04$, $p = 0.044$)[2], but the same was not seen for

[1]The null model of this comparison would not converge with any random terms included, so these were removed from both.

[2]Models would only converge when time | identity was the only error term.

Other Race valid ($\beta = 2.87$, $SE = 2.90$, $\chi^2(1) = 1.00$, $p = 0.316$).

ANOVAs. As well as mixed effects models we also report the results of a factorial ANOVA, given that controlling for both subject and identity in linear mixed effects models may unacceptably raise the risk of a Type II error. Initial analysis included experiment (Experiments 1a and 1b) as a between-subjects factor. Although there was an interaction between experiment, time and race ($F(1,58) = 9.83$, $p = 0.003$, $\eta_P^2 = 0.14$) where the difference between own- and other-race trust judgments was larger at the second rating in Experiment 6.1b than 6.1a, no interactions of experiment with the critical variable of validity were detected.

Collapsing across experiment, a 2x2x2 ANOVA looking at time, validity and race found a main effect of time ($F(1,59) = 8.03$, $p = 0.006$, $\eta_P^2 = 0.12$), one of validity ($F(1,59) = 12.02$, $p = 0.010$, $\eta_P^2 = 0.17$), and a main effect of race ($F(1,59) = 5.46$, $p = 0.023$, $\eta_P^2 = 0.08$). A significant interaction of time and validity was found ($F(1,59) = 17.43$, $p < .001$, $\eta_P^2 = 0.23$), indicating that there was significant learning of trust over time as a function of gaze cueing behaviour. Other two-way interactions were not significant (Fs < 1.1). Importantly, this learning did interact with race, as a three-way interaction was significant ($F(1,59) = 5.77$, $p = 0.019$, $\eta_P^2 = 0.09$).

To explore this interaction further we broke this down into separate analyses for own- and other-race faces and looked primarily at the interaction of time and validity. Analysis of time and validity in own-race faces found a significant interaction of time and validity ($F(1,59) = 22.29$, $p < .001$, $\eta_P^2 = 0.27$). The same analysis in other-race faces also found a significant interaction, but this was weaker than in own-race faces, as evidenced by the smaller F value and partial eta squared effect size statistic ($F(1,59) = 8.81$, $p = 0.004$, $\eta_P^2 = 0.13$). Follow-up paired-samples t-tests between the beginning and

end ratings for each condition found that invalid faces showed a similar decline in both own- ($t(59)$ = -4.21, [95% CI -24.47 to -8.70], p <.001, d = 0.66) and other-race faces ($t(59)$ = -3.47, [95% CI -19.39 to -5.21], p = 0.002, d = 0.57), but while own-race valid faces showed a significant increase in trustworthiness ratings ($t(59)$ = 3.28, [95% CI 3.17 to 13.10], p = 0.002, d = 0.35), this was not the case for other-race valid faces ($t(59)$ = 1.07, [95% CI -2.48 to 8.23], p = 0.287, d = 0.13).

6.2 Chapter Discussion

There are two key findings from this chapter. First, gaze cueing where participants follow the gaze direction of another person is unaffected by whether the viewed face is a racial in-group or out-group member, which suggests that gaze following is not sensitive to the identity of the face (see also Frischen & Tipper, 2004). Second, and in sharp contrast to attention cueing effects, incidental learning of trust from the predictive gaze patterns of ignored faces was influenced by race. That is, trust learning was larger and more robust for own-race faces.

As noted above, there is a wealth of previous literature that suggests we might see a difference in incidental learning processes between faces of different races (Dasgupta et al., 2000; Elfenbein & Ambady, 2002b, 2002a; Meissner & Brigham, 2001), but the underlying mechanisms remain unclear. We initially offered two potential mechanisms by which differences in learning between racial in-groups and out-groups may occur. The first point relates to expectations about cooperation − as we may expect out-group members to deceive more than in-group members (Stanley et al., 2011), differences in trust could related to a Rescorla and Wagner (1972) model of learning where unexpected information (i.e. in-group deceivers and out-group cooperators) is learned better than

stereotypical information. However, we found no evidence of such asymmetries in trust learning in this study − the magnitude of change in trust for other-race faces was smaller than own-race faces in both valid and invalid conditions. This is not to say that expectancy violation does not play a role in this learning, but we cannot draw this conclusion from our current data.

An alternative explanation that fits the current data better is that of learning efficiency. There is extensive prior research demonstrating that other-race faces are identified and remembered less efficiently than own-race faces (see Meissner & Brigham, 2001, for a review). Hence, we predicted that learning of trust from gaze would be less efficient in other-race faces. This was based on the idea that during the task where faces are irrelevant and to-be-ignored, an association has to be learned between a specific face identity and the pattern of eye-gaze it produces. It follows that the association between identity and gaze behaviour will be more easily learned if there is a strong/specific representation of the face identity. Experiments 3.2 and 3.3 confirmed this by manipulating the strength of face identity representations, demonstrating that stronger representations (greater familiarity with the faces) result in greater learning of trust from gaze behaviour. These data appear to support this latter interpretation, but it is important to note that this study aimed only to explore whether there was a *difference* between trust learning of own- and other-race faces, and was not designed to explore what this difference was. Follow-up studies would be needed to clarify the mechanisms underlying these differences.

There is also the possibility that this learning may be at least partly affected by stereotype content − that is, that incidental trust learning is affected by participants' preconceptions about the identities involved. For example, if one has a stereotype of an

out-group as being particularly violent and aggressive, then it may be that invalid faces would be particularly associated with negative trustworthiness ratings, while if one has a stereotype of an out-group as inscrutable or difficult to read then it may accentuate feelings of out-group homogeneity. With only one out-group in this experiment (East Asian faces) it is difficult to tease these apart. Future research may wish to use additional out-group categories (e.g. Black faces) or to explore the nature of participants' preconceived and stereotypical views as a way of exploring how stereotype content may affect learning.

The results of this chapter are particularly striking when compared with the results of Chapter 5, where faces were also members of discrete social categories. In a minimal groups paradigm, trust learning was extinguished and group-level representations were used to inform trustworthiness judgements. On the other hand in the present experiment, where race was the social category across which faces varied, trust learning was shown for both in-group and out-group members, but was stronger for in-group members.

This intriguing contrast could be because of the nature of the group manipulation – trust learning was extinguished when participants were explicitly assigned a personal trait of over- or under-estimator and this personal property was then explicitly linked to one group of people via shirt colour. In sharp contrast, in the current study no personal property was highlighted and explicitly linked to a group of people. Rather, participants simply passively viewed a range of faces while identifying peripheral targets in the gaze cueing procedure.

Therefore it seems that explicitly priming participants to represent people in terms of their group membership is able to supersede individual-level representations, but if participants respond to real-world groups without explicit instruction, then these

individual representations are maintained and influenced by the wider social context. One would therefore expect that individuals with stronger implicit beliefs or expectations about different racial groups might show greater differences in trust learning, so an approach for future research may be to measure individual differences in implicit attitude biases (Dasgupta et al., 2000) to correlate with learning of trust.

In conclusion, this chapter reports the result that shows that incidental learning of trust from gaze cues is affected by the race of the faces involved. While we see clear patterns of trust learning across both in-group and out-group members, this learning is stronger for in-group members, which could point to more efficient processing and more stable representations of in-group identities than out-group. Contrasts between this experiment and those in Chapter 5 indicate that this group distinction must be implicitly driven or naturally occurring in order not to override the representations of individual identities.

Chapter 7. The contribution of visuomotor fluency to incidental trust learning

A key feature of the previous research looking at incidental trust learning from gaze cues is that trust was influenced by eye-gaze behaviour of another person. Clearly looking towards or away from relevant objects is a means of deceiving another person and initiates joint attention, which recruits reward-related neurocircuitry (Gordon, Eilbott, Feldman, Pelphrey & Vander Wyk, 2013; Schilbach et al., 2010). However, if this rewarding sense of joint attention were the driving force behind this effect, then one would expect learning to be driven by increases in trustworthiness to valid faces, rather than the characteristic decrease in trust to invalid faces that has emerged throughout this thesis. Invalid faces are associated with a lack of rewarding joint attention, but beyond that they are also associated with disruptions to visuomotor fluency during responses (that is, there is a cost to RTs that is reliable when invalid faces appear that is absent for valid faces).

It is possible that this decrease in trustworthiness for invalid faces occurs due to them *withholding* the rewarding sense of joint attention, but it remains to be seen whether this effect is wholly dependent on this joint attention feature or if similar effects can be induced purely through selective disruptions of visuomotor fluency in the absence of any physical changes to the face. Previous research has shown that perceptual fluency (e.g., Reber & Schwarz, 1999) and motor fluency (e.g., Hayes, Paul, Beuger & Tipper, 2008) can influence emotional assessments of stimuli. Can impaired processing of a face with no physical changes, such as eye-gaze shifts, also influence trust judgements?

Therefore we investigate the incidental learning of trust in a task where enhanced visuomotor fluency is associated with some faces and impaired visuomotor fluency is associated with other faces in a similar way to gaze cueing studies. However, there are

no face behaviours, such as gaze shifts, that might be associated with deception. To this end we develop two new task-switching procedures in Experiments 7.1, 7.2 and 7.3. A task-switching paradigm involves participants performing two judgements of a stimulus on different trials. For example, two trials might require reporting the colour of a stimulus, while the next two trials might require reporting the identity of a stimulus. These paired trials and predictable switches between tasks continue throughout the experiment. When the task changes, a visuomotor cost (slower RTs, greater probability of errors) is associated with responses on that switch trial (e.g., Monsell, 2003; Wylie & Allport, 2000; Yeung, 2006), even when the change sequence is predictable and therefore switches can be anticipated (Kiesel et al., 2010; Rogers & Monsell, 1995).

If changing visuomotor fluency is sufficient to evoke affective reactions (see Constable, Bayliss, Tipper & Kritikos, 2013; Hayes et al., 2008) then creating disfluency while processing a particular face identity will reduce trust ratings. That is, throughout the experiment particular face identities are always presented on switch trials where RTs are slowed and errors are more likely, while other face identities are always presented on repeat trials where RTs are fast and accurate. In Experiment 7.1 we introduce a procedure that closely matches the gaze cueing procedure used in previous chapters; faces appear in the centre of the screen while objects appear on the left or right, and the judgement that participants make about these objects changes every other trial.

In Experiments 7.2 and 7.3 we introduce a task-switching procedure where the faces are now the targets of participants' decisions. In gaze cueing experiments, the sense of disfluency is contingent on the faces, which means there is some dynamic attention-grabbing feature that is difficult to inhibit. However, in Experiment 7.1 the identity of these faces may be easier to inhibit due to the lack of any dynamic changes

and the fact that they have nothing to do with the participant's disfluency (as opposed to an invalid gaze cue, which is entirely driven by the face). As such, these experiments change the task to focus on the faces, identifying either the colour (green or yellow) or identity (male or female) of the faces. This means that the faces are now the targets of the participants' judgements, which could facilitate trust learning as participants' attention is wholly on the faces, or could disrupt it, as literature on distractor devaluation suggests that to-be-ignored information often shows a devaluation that to-be-attended information does not (see Raymond, 2009). It could be that we would not see any learning of trust (particularly the characteristic decrease for invalid faces, those associated with low fluency, evident in gaze cueing experiments) in Experiments 7.2 and 7.3 because it is more difficult to devalue targets than distractors.

This two-trial task-switching procedure selectively disrupts visuomotor fluency on half of the trials that participants are exposed to. However, given the predictable nature of the task-switching procedure, repeat trials − which result in high fluency − are also subject to anticipation of an imminent task-switch, which may confuse the learned associations that are integral to this procedure. To get around this, Experiment 7.3 replicates Experiment 7.2 but includes a third trial in each sequence (a switch-repeat-prepare paradigm as opposed to a switch-repeat paradigm) to disentangle the effects of visuomotor fluency and imminent switch anticipation.

The primary hypothesis of this chapter is that if visuomotor fluency is the key driver for the learning of trust, then effects should be detected in one or more of these new tasks. On the other hand, if cues to deception such as eye-gaze are necessary, then no learning of trust should be detected.

7.1 Experiment 7.1

This experiment investigates whether the usual trust effect is contingent on gaze behaviour or if a similar result can be elicited just by manipulating the visuomotor fluency associated with the face in the absence of any physical changes to the facial features. As such, this experiment replaces the gaze-cueing procedure with a task-switching paradigm.

7.1.1 Methods

Participants

28 participants volunteered for this study in return for a mixture of course credit and payment. Four participants had to be removed after RT filters were applied, and so the final number available for analysis was 24 (17 female; $M_{age} = 18.95$, $s.d. = 1.38$).

Stimuli, Design and Procedure

This experiment closely matched the gaze-cueing experiment used elsewhere, particularly Experiment 4.1, but the faces no longer shifted their gaze. Instead, participants were told that they would be making one of two possible judgements on a given trial; the first was object TYPE, where they would categorise the object as either a kitchen or garage item (as in other experiments), while the second was object COLOUR, where they would judge whether the object was blue or yellow.

Changes to the stimuli from previous experiments were the introduction of yellow-coloured objects (the same objects as used in previous experiments but digitally manipulated to appear yellow instead of blue) and the fact that when faces appeared in the centre of the screen we used unaltered, original, neutral images rather than those

digitally manipulated to shift their gaze. We also introduced a task cue before each trial, to remind participants of whether they were supposed to judge the object's TYPE (kitchen/garage) or COLOUR (blue/yellow).

The task that participants completed altered every other trial in a switch/repeat task-switching procedure. As such, participants might start judging colour for the first two trials, then switch to type for the next two, and so on (order counterbalanced). Rather than providing valid or invalid cues, in this experiment faces would either always appear on a switch trial (the trial immediately following a change in task) or a repeat trial (the trial immediately preceding the change). As such, each face appeared at a point in the sequence that would associate it either with low (switch) or high (repeat) levels of visuomotor fluency (see Figure 7.1).

Data analysis

Before data were analysed, participants' responses were filtered to remove all error trials (where participants reported the incorrect answer) and RT outliers, as in previous gaze-cueing experiments − RTs below 250ms (too short to process the stimuli) and above 2,500ms (indicating that participants had not given a response in the allotted time). The number of remaining trials was then compared with the original number of trials to check that all participants retained at least 70% of their total trials and had not scored below 70% total correct on any one condition. Four participants were removed on this basis.

Data were analysed using linear mixed models as in previous chapters. When analysing RTs, the null model would not converge with any random terms included and so these were removed from both the null and trial-only models. For accuracy rates, all models converged with the maximum random structure. For trustworthiness ratings, the

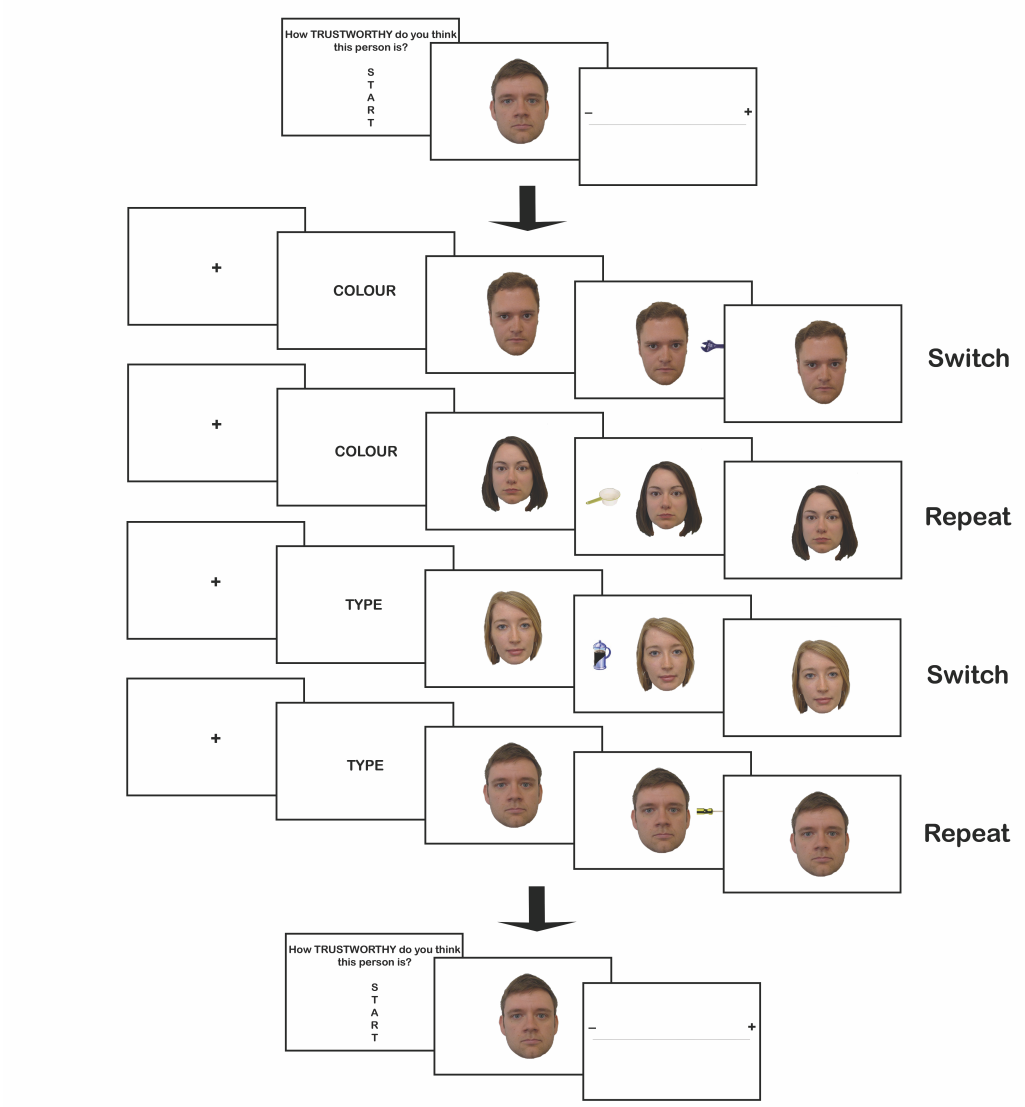

Figure 7.1: Example trials used in the task-switching paradigm in Experiment 7.1. As in previous gaze-cueing experiments, participants complete trustworthiness ratings at the beginning and the end. During task-switching, participants responded with the prompted information that alternated on a switch/repeat basis.

null and time-only models would not converge until the time | subject and trial | subject error terms were removed, and so these were removed from these models and the trial-only model for the sake of direct comparison. The two-factor models would not converge until only the time | subject error term was removed.

As with Chapters 2, 3, and 4, see Appendix A for more conventional ANOVAs and RT and accuracy rates broken down by experimental block.

7.1.2 Results and Discussion

Task-switching

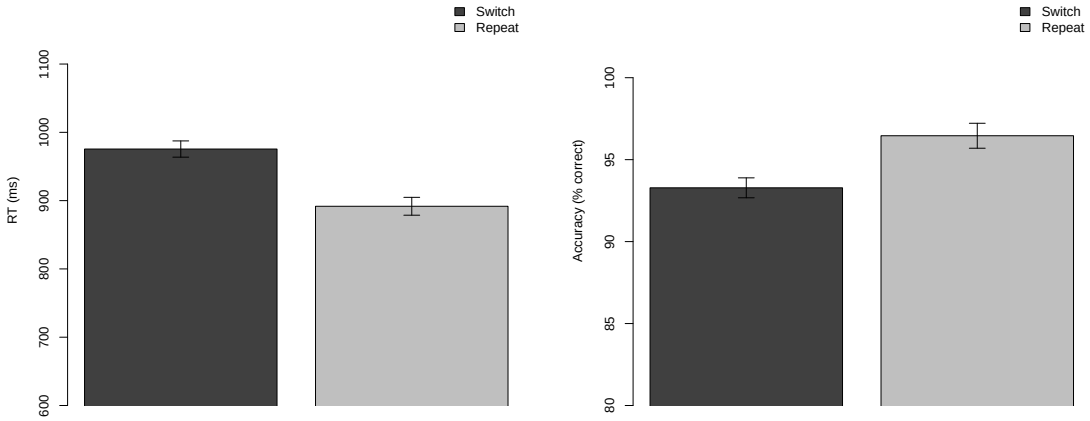

Figure 7.2: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 7.1 in response to switch (dark grey) and repeat (light grey) trials. Error bars show standard error.

The RT and accuracy results of Experiment 7.1 are shown in Figure 7.2. Fitting trial type (switch/repeat) to the null linear mixed effects model significantly improved the fit when explaining RTs (β = 93.09, SE = 14.24, $\chi^2(1)$ = 42.21, p <.001), as responses were faster on repeat trials than switch. A similar effect was seen in accuracy scores (β = -0.02, SE = 0.01, $\chi^2(1)$ = 8.65, p = 0.003), where accuracy rates were higher on repeat trials than switch.

Trustworthiness Ratings

The changes in trustworthiness ratings for the faces in Experiment 7.1 are shown in Figure 7.3. Adding time to the null model did not improve the fit (β = 1.79, SE = 1.64, $\chi^2(1)$ = 1.19, p = 0.275), but including trial type did (β = -4.14, SE = 1.64, $\chi^2(1)$ = 6.36, p = 0.012). The interaction model (time x trial type) did not significantly improve the model fit beyond the full model (time + trial type), where both factors were

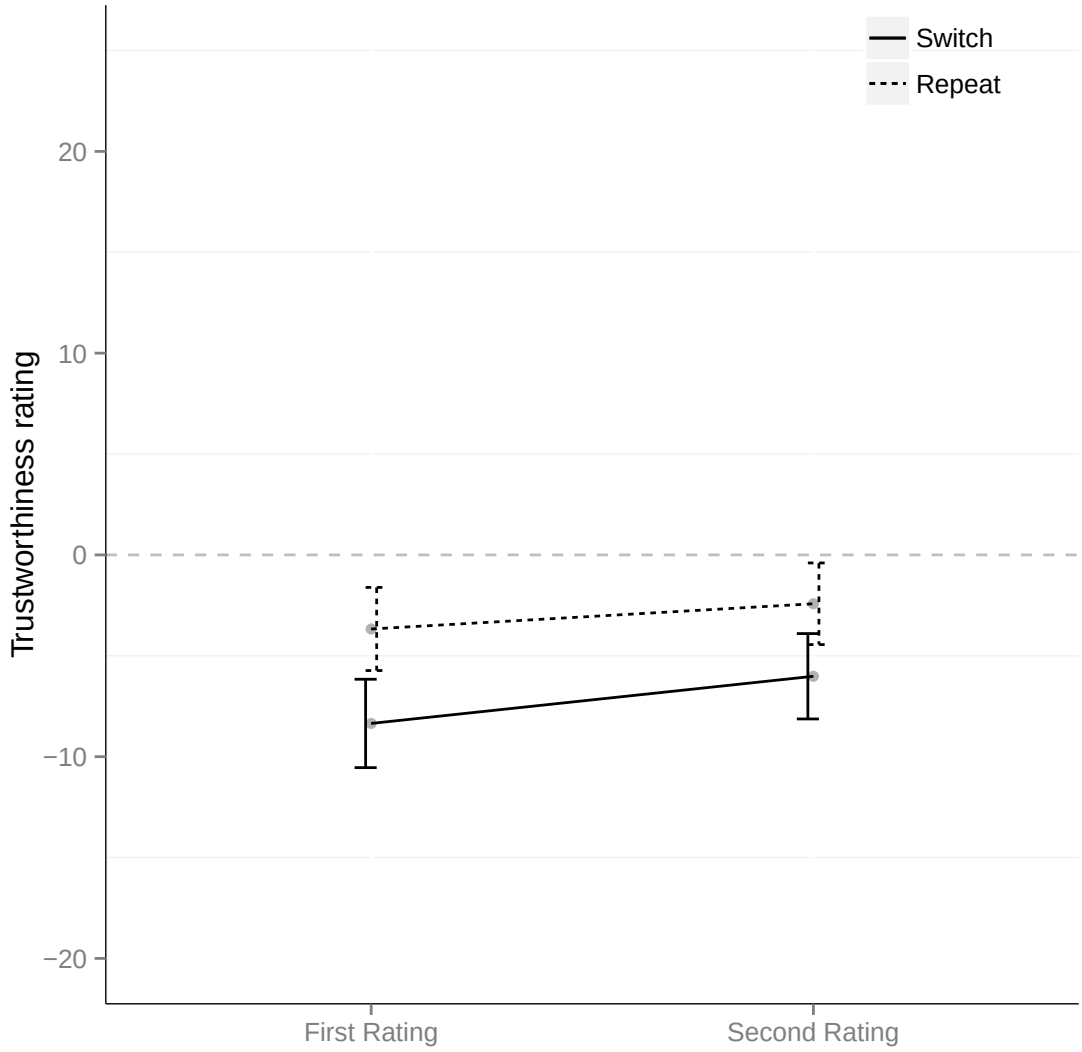

Figure 7.3: Time course of trustworthiness ratings over Experiment 7.1 for switch (solid) and repeat (dotted line) faces. Error bars show standard error.

modelled but without an interaction ($\beta = 1.09$, $SE = 3.28$, $\chi^2(1) = 0.11$, $p = 0.739$).

We ran further analysis of the changes in trustworthiness as a function of time for switch and repeat faces separately. These models found that time did not significantly improve the model fit for either switch faces ($\beta = 2.34$, $SE = 2.36$, $\chi^2(1) = 0.98$, $p = 0.322$) or for repeat faces ($\beta = 1.25$, $SE = 2.27$, $\chi^2(1) = 0.30$, $p = 0.581$).

Although the results of Experiment 7.1 do show a significant effect of face position, this appears to be due to chance differences in the pre-ratings − there is no logical reason to suppose that visuomotor fluency could have an effect before participants

encounter it, and so these differences must be due to random chance. The fact that they do not change over the course of the experiment, as evidenced by the lack of interaction and the remarkably flat profile of changes, is evidence that this incidental learning cannot be explained in terms of visuomotor fluency.

7.2 Experiment 7.2

This experiment replicates the task-switching procedure used in Experiment 7.1 but alters it so that the face images are now targets rather than distractors.

7.2.1 Methods

Participants

32 participants volunteered for this study in return for a mixture of course credit and payment. Eight participants had to be removed after RT filters were applied, and so the final number available for analysis was 24 (21 female; $M_{age} = 21.17$, $s.d. = 2.12$).

Stimuli, Design and Procedure

Stimuli were generated from the same KDEF faces used in Experiment 7.1. Participants completed the same trustworthiness ratings as in Experiment 7.1 before and after the experiment, using full colour unaltered images. During the main portion of the experiment, however, the paradigm was changed from gaze-cueing to task-switching, and for this all face images were superimposed with a chromatic hue in Adobe Photoshop CS6 to appear either green or yellow (see Figure 7.4 for examples).

Participants were told that they would be asked to make one of two judgements about a face image that appeared on the screen; they would either be asked to judge the

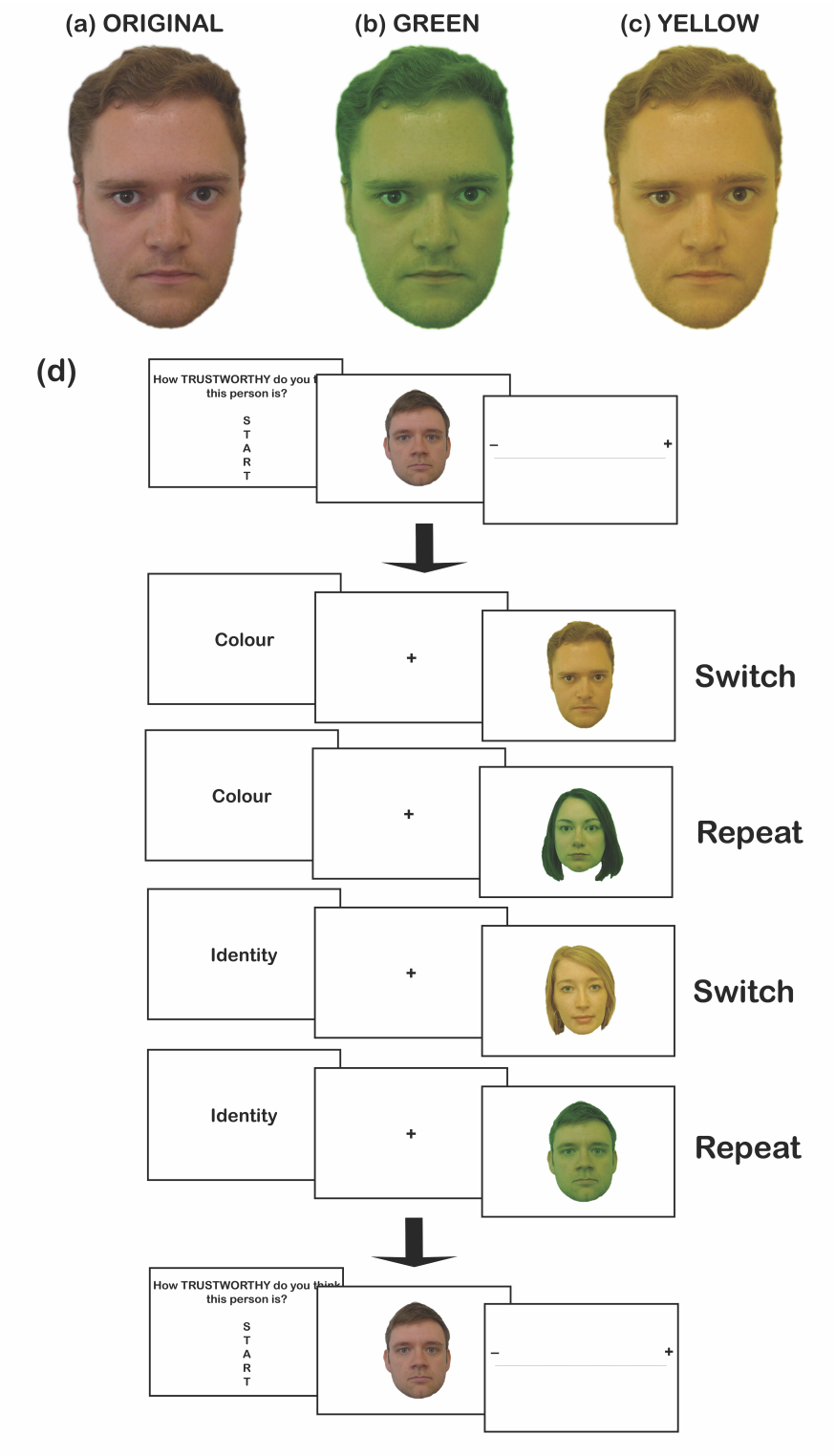

Figure 7.4: Examples of the coloured stimuli used in the task-switching experiment. (a) The original uncoloured images were used during trustworthiness ratings, while the (b) green and (c) yellow images were used in the task-switching portion. (d) Trial sequence. Participants reported whether the face was coloured in green or yellow or if the face was male or female, depending on a prompt before each trial.

colour of the image (Colour condition: green or yellow) or to judge the sex of the image

(Identity condition: male or female). Participants were told that the task they were to

perform would be shown to them as a reminder before each trial, but that the task would change every other trial such that they would perform two Colour trials, then two Identity, and so on. As in Experiment 7.1, half of the identities only appeared immediately after a task-switch, in the first position of the sequence (switch trial) and half appeared immediately before the switch in the second position (repeat trial). Identity and trial position were counterbalanced across participants.

During the course of a trial, a condition cue (either 'Colour' or 'Identity' alternating every two trials) would appear on the screen for 1,000ms to make participants aware of the task they were performing, followed by a 500ms fixation cross. The target image would then appear on the screen for 500ms, followed by a blank screen for 1,000ms. Participants could respond at any point in this 1,500ms window but anything after that was classed as incorrect. Participant responses were the keyboard buttons Z and M, each of which corresponded to a different answer in the two tasks (i.e. Z, male and green; M, female and yellow − counterbalanced across participants).

Data analysis

The same RT filters were applied to the data as in Experiment 7.1, with the exception that the upper time limit was reduced to 1,500ms in line with the new timings. Incorrect responses and responses faster than 250ms were removed from the data and the participants' accuracy and number of trials were considered to see if they retained more than 70% of their original number of trials. In this experiment, eight participants in total committed too many errors to be suitable for inclusion.

When analysing RTs and accuracy rates, no models would converge with any random terms, so these were removed. For trustworthiness ratings, the null and

time-only models would not converge until the time | identity and trial | subject random slope terms were removed. The trial-only model, however, would not converge until all random terms were removed. This does not allow for direct comparison with the null model, and so we generated a simplified null that included no random slope terms and compared the trial-only model with this, while the time-only model was compared with the typical maximum random structure that would converge.

The two-way interaction model would not converge until the time | subject and trial | subject terms were removed, and so these were removed from the two-factor and interaction models.

7.2.2 Results and Discussion

Task-switching

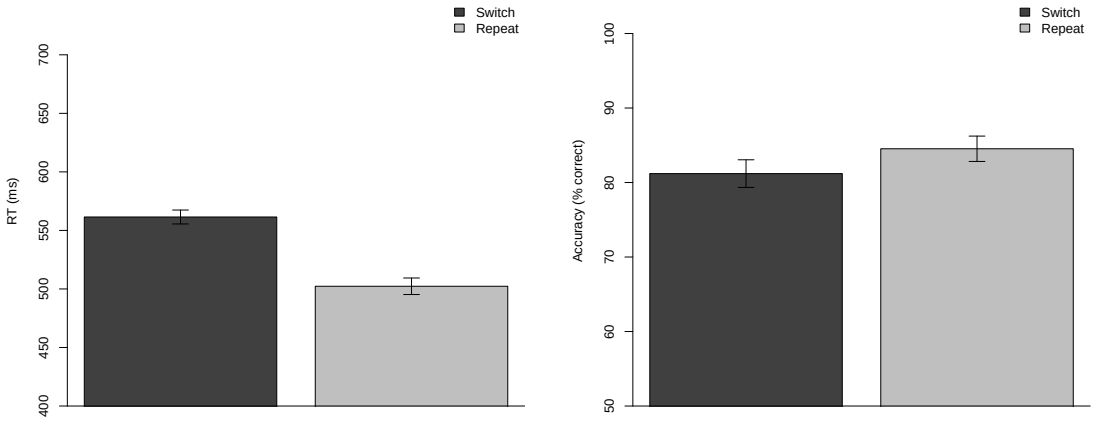

Figure 7.5: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 7.2 in response to switch (dark grey) and repeat (light grey) trials. Error bars show standard error.

The RT and accuracy results of Experiment 7.2 are shown in Figure 7.5. Fitting trial type to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -62.02, SE = 6.43, $\chi^2(1) = 90.79$, $p < .001$) as responses were faster

on repeat trials than switch. A similar effect was seen in accuracy scores ($\beta = 3.50$, $SE = 1.03$, $\chi^2(1) = 11.55$, $p < .001$), where responses were more accurate on repeat trials than switch.

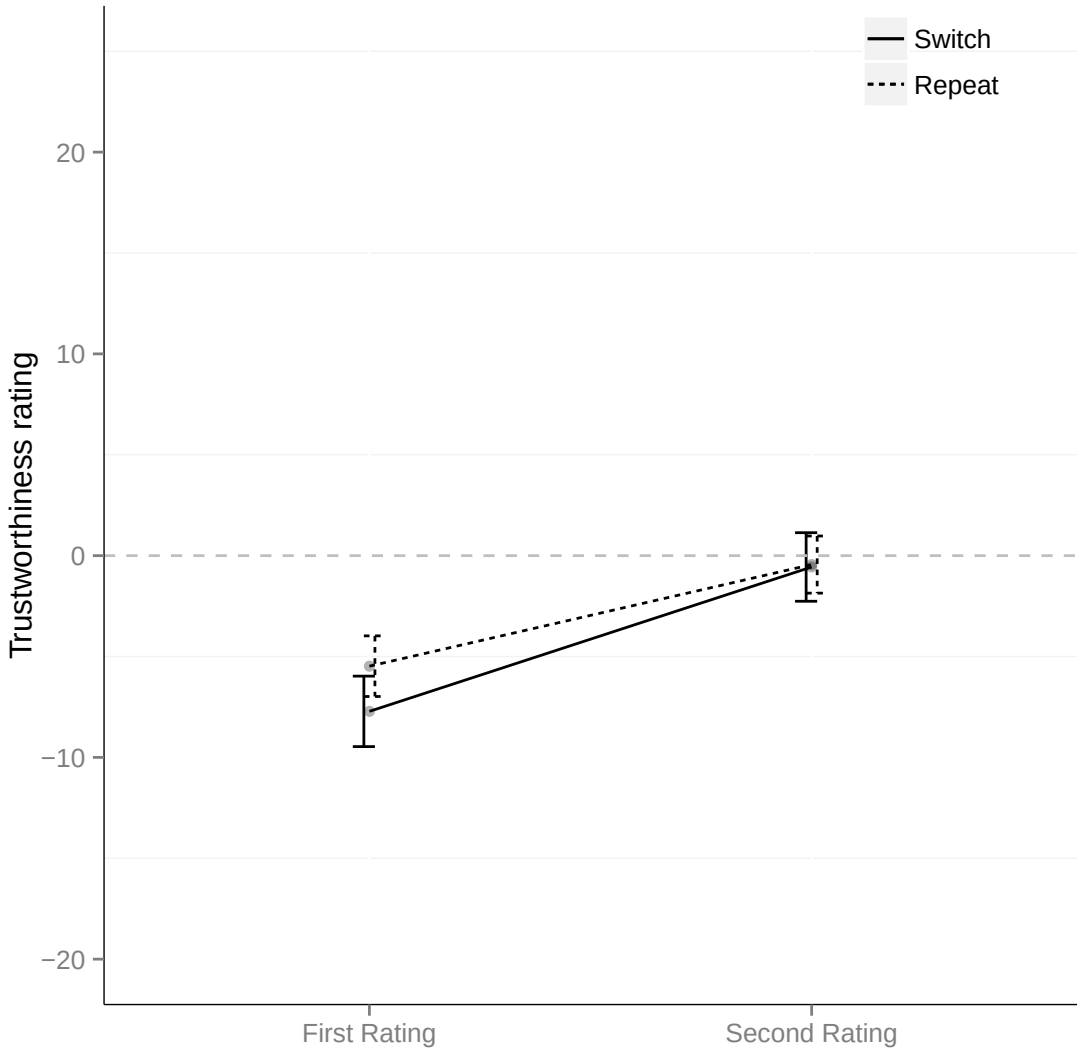

Figure 7.6: Time course of trustworthiness ratings over Experiment 7.2 for switch and repeat faces. Error bars show standard error.

Trustworthiness Ratings

The changes in trustworthiness ratings for the faces in Experiment 7.2 are shown in Figure 7.6. Adding time to the null model significantly improved the fit ($\beta = 6.10$, $SE = 2.86$, $\chi^2(1) = 4.33$, $p = 0.037$), but including trial type did not ($\beta = 1.18$, $SE = 1.66$,

$\chi^2(1) = 0.79$, $p = 0.375$), and the interaction model (time x trial type) did not significantly improve the model fit beyond the full model (time + trial type) (β = -2.12, $SE = 2.62$, $\chi^2(1) = 0.66$, $p = 0.418$). Further analysis of the changes in trustworthiness as a function of time for switch and repeat faces separately found that time significantly improved the model fit for both switch faces ($\beta = 5.04$, $SE = 1.81$, $\chi^2(1) = 7.69$, $p = 0.006$) and repeat faces ($\beta = 7.16$, $SE = 1.83$, $\chi^2(1) = 15.02$, $p <.001$).

The finding of an effect of time in this experiment may reflect the faces' status as targets rather than distractors − as targets, they may be more susceptible to an effect of mere exposure, where repeated exposure to stimuli results in more positive valuations (Burgess & Sales, 1971), whereas such an effect was not observed in Experiment 7.1 because they are irrelevant distractors.

7.3 Experiment 7.3

Experiment 7.3 replicates Experiment 7.2 but adds a third trial to the task sequence to dissociate effects of repetition from imminent switch preparation. The trial sequence now progresses switch, repeat, prepare. We also included blocks of Pure trials (no switching) with different faces to acclimatise participants to the responses and attempt to reduce the number of excluded participants.

7.3.1 Methods

Participants

A further 27 undergraduate and postgraduate students at the University of York volunteered for this study in return for a mixture of course credit and payment. Three participants were removed following the application of RT filters, and so the final

number available for analysis was 24 (21 female, age data not collected).

Stimuli, Design and Procedure

The methods were identical to Experiment 7.2 with some notable changes, the primary change being that rather than changing every other trial, the task now changed every three trials.To accommodate the increased complexity of the learning (three levels of trial type rather than two) the number of stimuli was reduced to 12 to allow for four (two male/ two female) in each group (switch/repeat/prepare). Faces were presented eight times each over four experimental blocks, meaning that each face was presented 32 times over the course of the experiment. This gave participants much more opportunity to learn the fluency associations with the faces than in Experiment 7.2 and we felt this might give more of an opportunity to learn associations of identity and fluency if such an effect were true.

As well as this, we also included a familiarisation procedure at the beginning of the task switching, where participants completed blocks of 'pure' trials (that is, a block where they judged colour and a block where they judged identity, with no switching). This procedure used six separate identities that did not appear in the main task-switching procedure. These additional faces were included as filler in the trustworthiness ratings at the beginning and the end of the experiment but are not included in our analyses.

Data analysis

The same RT filters were applied to the data as in Experiment 7.2. In this experiment, three participants had to be removed on the basis of RT filters. When analysing RTs and accuracy rates, all models converged with the maximum random structure. For

trustworthiness ratings, the null and single factor models would not converge with any random slope terms defined. The two-factor and interaction models converged with only the time | subject term removed.

7.3.2 Results and Discussion

Task-switching

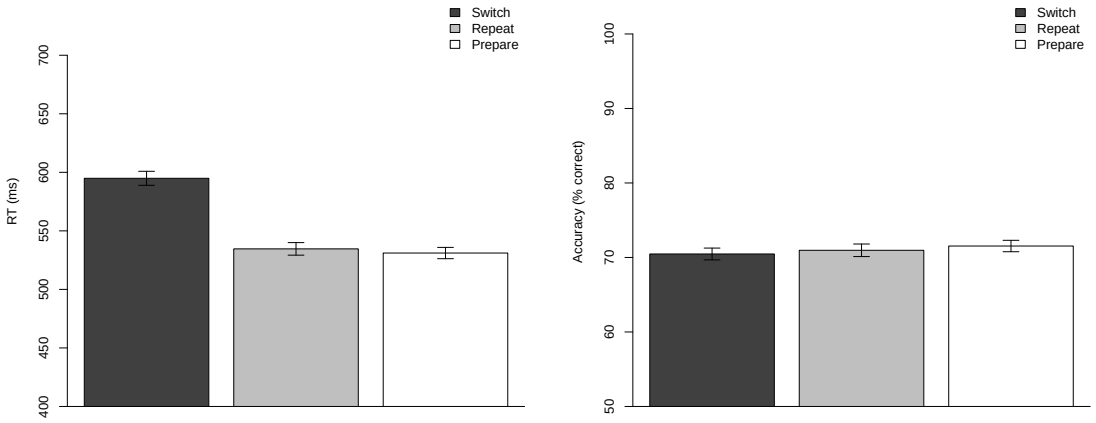

Figure 7.7: Averaged reaction times (milliseconds; left plot) and accuracy rates (percent correct; right plot) in Experiment 7.3 in response to switch (dark grey), repeat (light grey) and prepare (white) trials. Error bars show standard error.

The RT and accuracy results of Experiment 7.3 are shown in Figure 7.7. Fitting trial type to the null linear mixed effects model significantly improved the fit when explaining RTs (β = -60.29, SE = 10.31, $\chi^2(2)$ = 22.92, p <.001), as participants were slower to switch trials than either repeat or prepare trials. This effect was not seen in accuracy scores (β = 0.49, SE = 0.94, $\chi^2(2)$ = 1.29, p = 0.525).

Trustworthiness ratings

Changes in trust ratings for each position in the task-switching triads are shown in Figure 7.8. Adding time to the null model significantly improved the fit (β = 4.85, SE =

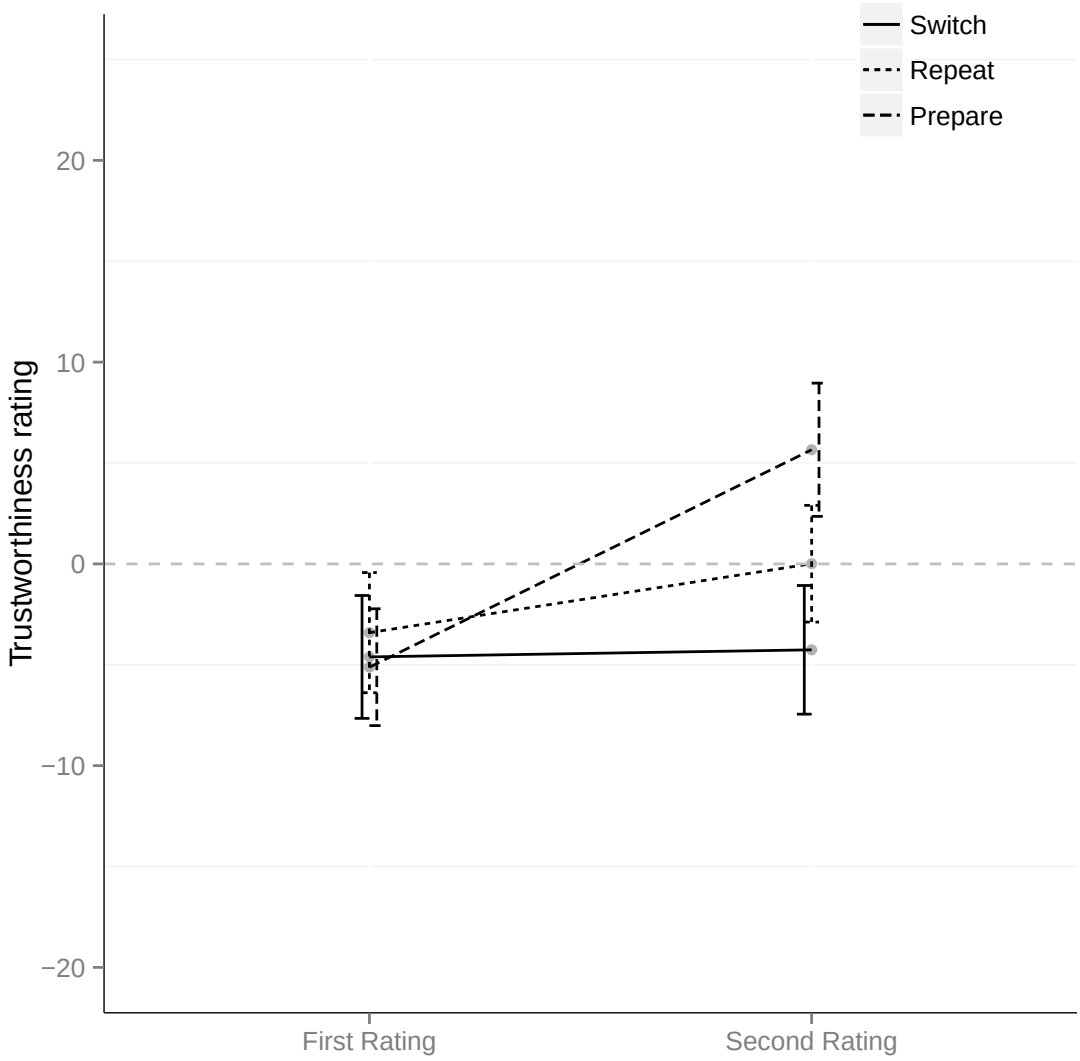

Figure 7.8: Time course of trustworthiness ratings over the course of Experiment 7.3 for switch (solid), repeat (dotted) and prepare (dashed line) trials between first and second ratings. Error bars show standard error.

2.06, $\chi^2(1) = 5.52$, $p = 0.019$), but including trial type did not (β = -1.71, SE = 2.53, $\chi^2(2) = 3.97$, $p = 0.138$). In this experiment, the interaction model (time x trial type) did fit the data marginally better than the full model (time + trial type) ($\beta = 0.20$, SE = 3.83, $\chi^2(2) = 4.72$, $p = 0.094$).

We ran further analysis of the changes in trustworthiness as a function of time for switch, repeat, and prepare faces separately. These models found that time did not significantly improve the model fit for either switch faces ($\beta = 0.35$, $SE = 3.34$, $\chi^2(1) = 0.01$, $p = 0.916$) or for repeat faces ($\beta = 3.42$, $SE = 3.61$, $\chi^2(1) = 0.90$, $p = 0.344$), but

did significantly improve the fit for prepare faces ($\beta = 10.78$, $SE = 3.39$, $\chi^2(1) = 9.91$, $p = 0.002$).

Linear mixed effects models suggested that there was a marginal interaction of trial type (switch/ repeat/ prepare) and time, such that faces that appeared on prepare trials were rated more trustworthy after the experiment than those that appeared on switch or repeat trials. However, the fact that the model fits the data does not mean that this result is interpretable: it is difficult to imagine why prepare faces should increase in trustworthiness so much more than repeat faces when both types of face are associated with greater visuomotor fluency than are switch faces (as evidenced by the RT cost, which is comparable to that seen in gaze cueing). It should also be noted that we collected data from five other task-switching studies (not reported here) and found highly significant task switching effects in RT and accuracy, but never observed any consistent effect of this visuomotor fluency on trust ratings of associated faces. For example, one such experiment was identical to Experiment 7.3 but used a different set of faces and failed to replicate the trust rating results reported here. As such, we feel confident in rejecting task-switching as an alternative methodology to gaze cueing, as over the course of eight experiments (the three here and the five unreported ones) no consistent patterns of results emerge.

7.4 Chapter Discussion

It is now well established that the eye movements of another person automatically shift attention and whether they consistently look towards or away from objects affects incidental learning of trust. The shifts of attention of another person certainly can be used to deceive, and hence it might be predicted that if there are no such behaviours in a

face, then learning of trust does not take place, even though particular face identities are associated with different levels of visuomotor fluency. Therefore this chapter examined whether learning of trust could be generated in the absence of any physical changes to the faces through a task-switching procedure. We found that in the absence of any physical changes, disruptions to participants' sense of visuomotor fluency were not sufficient to generate changes in trustworthiness, despite the RT costs associated with task-switching being comparable to those associated with gaze-cueing. This finding also held true regardless of whether the faces were distractors (Experiment 7.1) or targets (Experiments 7.2 and 7.3), and whether repeat trials included anticipation of an imminent switch (Experiment 7.2) or if this anticipation was removed (Experiment 7.3).

The results of Experiments 7.1, 7.2 and 7.3 indicate that the task-switching procedure − one that approximates the gaze-cueing procedure without physical changes to the face and retains only disruptions to visuomotor fluency − is unable to elicit reliable trust effects.It is also worth noting that this cannot be explained by the faces being more resilient to devaluation in Experiments 7.2 and 7.3, as Experiment 7.1 used the same faces as distractors that appear before the target object. While we must be cautious when interpreting null effects, the fact that multiple experiments that use different variations of the paradigm all fail to find evidence for trust learning means that the results of this chapter make a strong case against disruptions to visuomotor fluency being sufficient for incidental learning of trust.

These findings contrast with previous work that has shown that perceptual fluency does increase liking of objects (e.g., Reber & Schwarz, 1999; Burgess & Sales, 1971). There is also previous literature that does report that processing fluency can affect judgements of trust. For example, Winkielman and Olszanowski (2015) found that

increasing the disfluency associated with certain faces in a task requiring the identification of face emotion led to decreased ratings of trust in later judgements, and that the effect of this disfluency was unrelated to face valence. However, it is important to note that our experiments examined learning in the absence of explicit judgements of physical cues to trustworthiness (such as changes in expression) and as such these results are not necessarily inconsistent with this previous literature.

Learning of trust from patterns of eye-gaze is probably effective because joint attention can be positively reinforcing (Schilbach et al., 2010) and because gaze direction can be used by primates and humans to misdirect the attention of others (e.g., Klein, Shepherd & Platt, 2009). Hence the invalid gaze-cue when the face looks away from the highly salient target will be perceived as an act of deception. In contrast, the static faces in the task-switching procedure presented here do not provide such socially relevant information. Hence we can conclude that visuomotor fluency is not sufficient to produce changes in trust, but leave the question for future research to address what other mechanisms might exist for incidental social learning.

Chapter 8. General Discussion

The aim of this thesis has been to explore the mechanisms and properties of incidental learning of trust from gaze cues. The original effect outlined by Bayliss and Tipper (2006) showed that valid faces were selected as more trustworthy than invalid faces in a 2AFC measure, and that invalid faces were selected as having appeared more frequently than their valid counterparts. The experiments in this thesis have replicated and extended this original work using a dual scalar rating paradigm initially developed by Manssuer, Roberts and Tipper (2015).

Chapter 2 replicated the original result and showed that participants adjusted their ratings of trustworthiness in line with previous cueing behaviour. For faces expressing neutral emotions, this was characterised by a decrease in trust for invalid faces, but when they smiled this effect was bidirectional as it also showed an increase in trust for valid faces. However, given some evidence from later experiments in the thesis (Experiment 3.1 in particular, which is identical to Experiment 2.1, but shows a different change profile for valid faces), it is difficult to infer precisely what this difference between experiments might mean for valid faces. Across the thesis, and shown below in Figure 8.1, changes in trustworthiness to valid faces were much more varied than were changes to invalid faces and appear to be sensitive to factors other than expression. This point is discussed later in more detail.

We also explored some boundaries of this learning. Given that trustworthiness is a broad social dimension that can be applied in many different ways (e.g. trustworthiness as a proxy for warmth, in that trustworthy people are generally thought of as approachable, vs. trustworthiness as a measure of statistical reliability), it was surprising to see that this incidental learning from gaze cues appears to be quite specific, as it did not generalise to judgements of likeability. As such, while trustworthiness can be

considered a broad term, it appears that participants made their social judgements according to a particular sub-dimension of trustworthiness, and this dimension did not relate to how much participants liked the faces involved.

Another important boundary of the effect was that participants could not explicitly report the cueing behaviour of the faces after they had learned about them. This was a surprising result − the faces were clearly a salient feature, as evidenced by reliable cueing in RTs, and face-cueing associations were 100% reliable (in that a valid face would always be valid and never invalid), and yet the majority of participants (26 out of 30) performed at chance when asked to report which way the faces looked. This points to the mechanisms that underlie this social learning being implicit in nature, supporting previous observations of Bayliss and Tipper (2006) and Rogers et al. (2014). This fits with some evidence that unconscious processes play an important role in social information processing, from monitoring deception to making social evaluations of other people's personalities (ten Brinke et al., 2016; Pawling, Kirkham, Tipper & Over, 2016).

In Chapter 3 we explored how durable this memory for faces was. Despite several replications of the original effect, the question of how long this learning may last has never been explored. We found that when a short interference task was introduced between cueing and the final trustworthiness ratings, incidental trust learning was weakened, but not entirely eradicated. However, if participants were given the opportunity to familiarise themselves with the faces beforehand then this learning was much more stable, and could even be seen up to an hour later. Interestingly, the crucial element of this interference appeared to be a change in the task that participants were instructed to perform, as when we included an additional block of gaze cueing that differed only in terms of the faces' cueing behaviour (reversed such that valid faces were

now invalid and vice versa), memory for the original cueing behaviour was left relatively intact. It may be that an avenue for future research would be to investigate how different types of interference affect incidental social learning.

The nature of these stored representations of identities was explored in Chapter 4. One way in which participants may remember the trustworthiness of faces could be to update their memory for the physical appearance of the face to appear more or less trustworthy, to facilitate access to stored trait information. However, over two experiments − one where participants were asked to morph the faces and another where they made a 2AFC between trustworthy and untrustworthy morphed exemplars − we found that they did not seem to update these representations. This was further supported by the fact that participants did not give more favourable aesthetic ratings to images presented alongside previously valid faces than invalid faces: Evidence suggests that people will adjust ratings of coincident objects when faces are more attractive (Strick et al., 2008) or displaying a positive emotion (Bayliss et al., 2007), but this is apparently not the case in the absence of visual cues in the face, which again suggests that learned information does not interfere with how the physical properties of faces are remembered or perceived. In hindsight this seems advantageous − when building a representation of a person it is important to remember both static features (e.g. their physical appearance) and more fluid features (their behaviour or trustworthiness) separately. If you were to update your memory for how a face looked every time that person did something trustworthy, even if the changes were subtle, this could quickly stack up to the point where the memory for that face's physical appearance was warped beyond recognition, rendering it useless. Rather it is more sensible to keep these representations of the physical properties of a face and the associated emotional

responses or feelings largely independent of each other.

Chapters 5 and 6 looked at how higher order social information about a face affects how we learn about them from observing their gaze behaviour. In Chapter 5 we assigned participants to minimal groups, designated by a blue or yellow t-shirt, and completed the typical gaze-cueing paradigm. Interestingly, participants did not seem to learn about the cueing behaviour of individual faces anymore, and instead defaulted to making group-level distinctions where they judged in-group members as more trustworthy than out-group members (regardless of their cueing behaviour, or 'actual' trustworthiness). This extinction appears to be due to the explicit nature of the group manipulation, as in Chapter 6 where we used the real-world social group of race without drawing attention to the distinction participants again showed reliable learning of individual behaviours. However, in this experiment, participants now showed better learning for in-group (own-race) members than out-group (other-race) members, which seems to be driven by poorer representations of other-race individuals (or out-group homogeneity; the phenomenon whereby out-group members are considered more similar to each other, and so less individually distinct, than are in-group members; Park & Rothbart, 1982). The results of these chapters show that the influence of social group categories on incidental social learning is influenced not only by explicit saliency of group categories but also participants' own beliefs about the different groups.

The final chapter of this thesis aimed to explore the role of visuomotor fluency in incidental social learning. During gaze cueing, participants experience a rewarding sense of joint attention from valid faces as they look at the same targets as the face, where this is not the case for invalid faces. However, throughout the thesis the key finding of incidental trust learning has been that a decrease in trust to invalid faces is the more

reliable change than an increase in trust for valid faces, which suggests that this effect may not be driven by rewarding joint attention. Chapter 7 aimed to see if it could instead be explained by disruptions to visuomotor fluency; when an invalid face cues the wrong location, it causes a cost to fluent processing as participants need to reorient to the correct location, resulting in longer reaction times. To explore this we developed three task-switching procedures, where certain faces were consistently associated with high (comparable to valid faces) or low (comparable to invalid) visuomotor fluency. However, we found no reliable effects of fluency on trustworthiness ratings, and it seems that this paradigm may not be sufficient to elicit these changes in social judgements – perhaps because without a physical change in the face, such as a gaze shift, participants do not spontaneously anchor their sense of disfluency to the identity.

Over the course of eighteen experiments this thesis has explored a variety of features of this incidental trust learning from gaze cues. However, the number and variety of results presented here belie the consensus of several different experiments on what has proven to be a reliable and clear effect of incidental trust learning. When looking at the results of the thesis as a whole, an interesting picture of this mechanism emerges. For this reason, the next section reports a meta-analysis across all experiments in this thesis where trustworthiness judgements were collected before and after gaze cueing, to outline some of the recurring and important features of this effect.

8.1 Incidental learning across all experiments: a meta- analysis

It can be difficult when examining experiments and chapters individually to determine how strong and reliable incidental trust learning from gaze cues actually is. While several experiments in this thesis have replicated the effect reliably (Experiments 2.1,

2.2, and 3.3, for example), others have been weaker or shown little evidence of any trust learning (such as the minimal groups experiment in Chapter 5). In order to make inferences about the true size of the effect, we report the results of a meta-analysis of all experiments included in this thesis that can yield an effect size of incidental trust learning from gaze cues.

8.1.1 Meta-analysis protocols

Not all experiments in this thesis were included in the meta-analysis. Experiments were selected only if gaze cueing were the means by which social judgements were manipulated (therefore the task-switching experiments of Chapter 7 were not included), and only if *trustworthiness* judgements were collected at the beginning and at the end of the experiment (therefore not only were experiments such as Experiments 2.4, 4.1 and 4.2 excluded for not including both ratings, but also Experiment 2.3, which asked about liking rather than trust). Eleven experiments were included in this analysis, including data from 357 participants.

In order to analyse these results in a meaningful way, we needed to reduce them to a metric that was comparable across studies. Using the ANOVA tables listed in Appendix A, we were able to calculate r values from the F-ratios and degrees of freedom (as outlined in Field, 2007) for all time x validity interactions. While this meant that some nuances were ignored − for example, where such two-way interactions were modulated by a third factor, as in Experiment 6.1 − this allows us to directly compare the effect size for this interaction (as a measure of incidental trust learning) across all experiments.

These data were then analysed using the `meta` and `MAc` packages in the statistical software `R`. We selected a random-effects meta-analysis model, as while all of our

experiments were similar in how data were collected (they all used the same stimuli, were tested on the same participant pool, and run by the same experimenter), several experiments included manipulations designed a priori to weaken or enhance the effects, which means that we could not justifiably say that the true effect size for each study would not vary.

8.1.2 Results of meta-analysis

The results of the meta-analysis are shown in Figure 8.1. Across all eleven studies, a random-effects meta-analysis model found a moderate overall effect size ($r = 0.38$, 95% CI [0.31 to 0.45], $z = 10.39$, $p<.001$). As is evident from the forest plot in Figure 8.1, there was substantial variation across experiments, as certain manipulations were effective at disrupting learning (e.g. introducing an interference task in Experiment 3.2, or assigning participants to minimal groups in Experiment 5.1), while other manipulations yielded apparently stronger learning (e.g. when faces smiled in Experiment 2.2, or when participants were familiar with the faces in Experiment 3.3). A few experiments where individual results were somewhat ambiguous to interpret (such as Experiment 3.4, where participants left for an hour between cueing and rating, or Experiment 5.2, where faces still wore coloured shirts but were not assigned to discrete groups) now seem mostly in line with the overall effect size.

An important point to note when looking at the results of all experiments combined is that the consistent element of trust learning across this thesis is a decrease in trustworthiness in response to invalid faces. As evident in the bar plot (bottom) of Figure 8.1, while the magnitude of this decrease may vary, it is much more stable than any changes in valid faces across experiments. In fact it seems to be convincingly absent

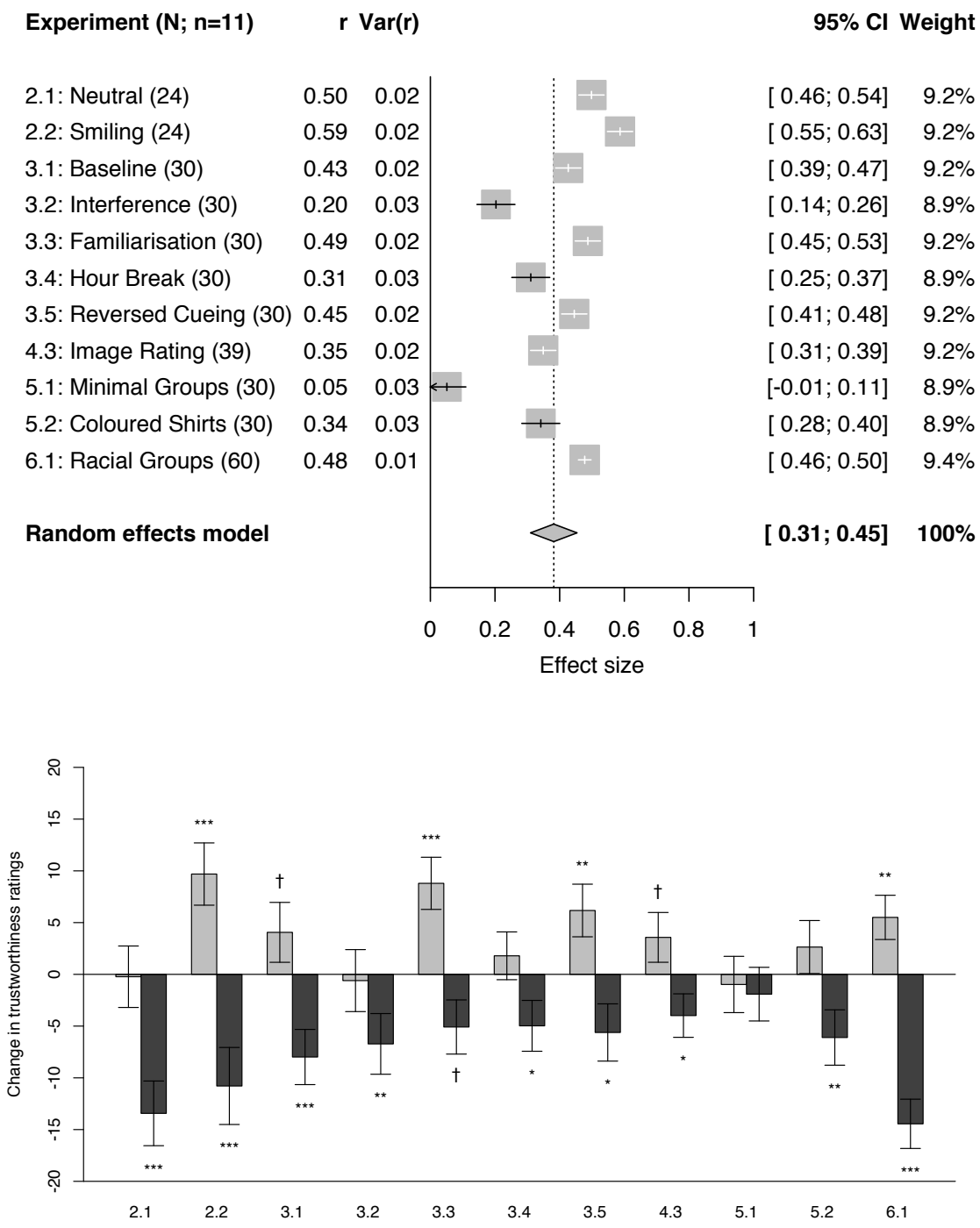

Figure 8.1: Results of all experiments included in a meta-analysis of all eleven results in the thesis that explore an interaction of time and cueing validity on gaze cues. The top plot shows a forest plot where the effect sizes (r, calculated from ANOVA outputs listed in Appendix A) are plotted along with the weights assigned to each by the random effects meta-analysis. Error bars show 95% confidence intervals for effect sizes. Black bars show those experiments that are underweighted in the analysis. The bottom plot shows the data from all experiments as change in trustworthiness scores (second rating minus first rating) for valid (light grey) and invalid (dark grey) faces. Significance markers denote significance of the effect of time on ratings as calculated with linear mixed effects models. Error bars show standard error. $^{***}p<.001$; $^{**}p<.01$; $^{*}p<.05$; $^{\dagger}p<.10$.

in only one experiment (Experiment 5.1, where participants were assigned to minimal groups), but in all other cases either trends or significantly changes in the predicted direction. This provides strong evidence that this incidental trust learning is part of a mechanism that is specialised for detecting invalid or deceptive faces. On the other hand, learning of valid faces may be more susceptible to other factors, such as the emotion that the face is posing or how familiar we are with the identities. It may also be that there are confounds that are outside the control of the experimental design − for example, the time in the academic year that data are collected could play a role, as participants who sign up for studies early in the academic calendar are more motivated than those who leave it until the end, particularly when they are volunteering for course credit (Nicholls, Loveless, Thomas, Loetscher & Churches, 2014).

8.2 Implications for a model of incidental social learning

In Chapter 1, we introduced a tentative model of how incidental social learning from gaze cues may be processed within the cognitive system. This model relies on there being two streams of information about faces − one that processes invariant information such as identity and is likely subserved by neural regions in the fusiform gyrus, and the other that processes variant information such as eye gaze and is likely processed more dorsally by regions in the superior temporal sulcus − that feed into a stable identity representation for that face, which is likely stored in a region in the anterior temporal lobe (Haxby et al., 2000). The model is shown again in Figure 8.2. We can now begin to consider the components of this model in light of the evidence presented in this thesis.

We know that eye gaze information is encoded along the variant pathway (in blue) and eventually feeds into the stored identity representation. The evaluative filter A.

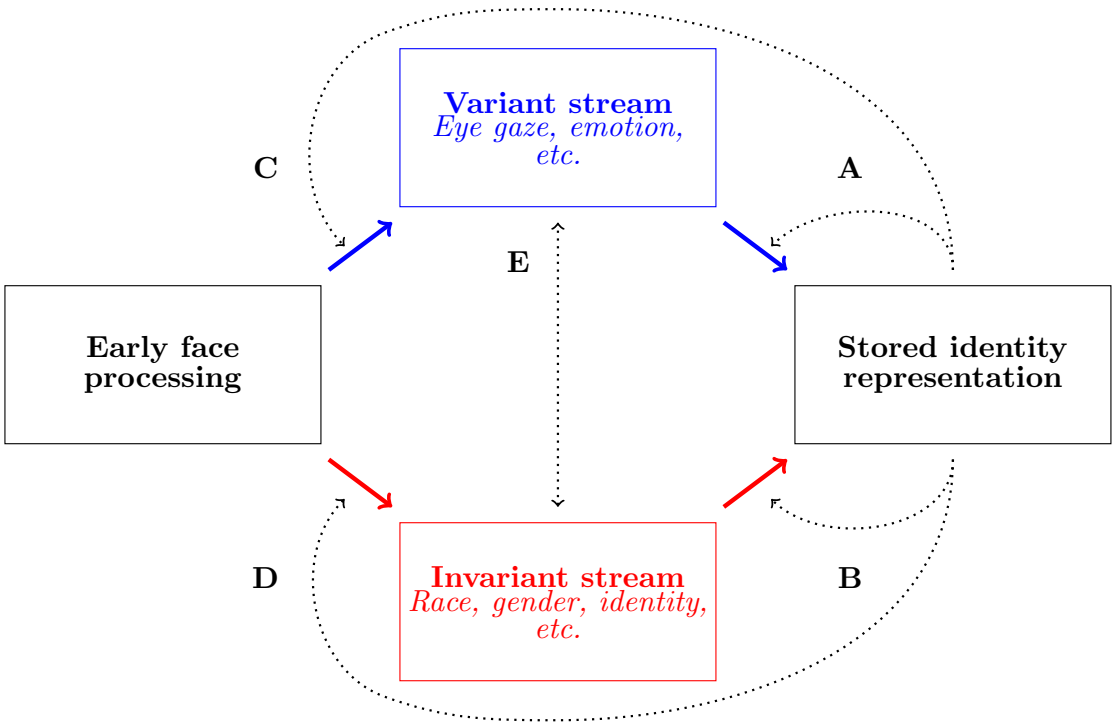

Figure 8.2: A model of incidental social learning from gaze cues presented in Chapter 1. Visual information enters the model from the left, through early face processing systems that identify the face-like configuration and structure. Information is then processed by separate streams: an invariant stream, in red, which processes information that is unlikely to change over the course of an interaction (e.g. identity), and a variant stream, in blue, which processes information that is likely to change (e.g. gaze direction). These streams then feed into a stored representation of the individual's identity, which can be used later to process incoming information. Some examples of feedback communications are shown: A. and B. processing of variant (A.) and invariant (B.) information is not affected by person knowledge, but the integration of this information is affected by what is already known about that identity; C. person knowledge affects processing of variant information such as eye gaze or emotion; D. person knowledge affects processing of invariant information such as gender or race; E. either variant or invariant information is affected by the content of the other.

evaluates incoming information based on how it fits with the stored identity representation and other sources of available information. Evidence suggests that over time, information from the stored identity representation could be used to filter this information (that is, eye gaze information is processed and followed by parietal attentional modules, but the social implications for this are then weighted according to prior knowledge of that person). As faces become more familiar, the filtering criteria may shift from monitoring for deception and untrustworthy individuals, to identifying and remembering cooperative or helpful behaviour (c.f. Experiment 3.1 compared with Experiment 3.3). Further research is needed to identify the features of this shift in order

to answer the question of why it would be easier to learn about prosocial behaviour for familiar faces, but antisocial behaviour for unfamiliar faces.

We now know that evidence for valid or invalid behaviour appears to be acquired cumulatively, as when faces change their cueing behaviour in the final block this does not override what participants have already learned about the face (c.f. Experiment 3.5). Once this system has learned about cueing validity, then, it appears to be somewhat resilient to contradictory information. In fact, it may be that continued exposure to eye gaze information helps to sustain these representations − when participants experienced an unrelated interference task in Experiment 3.2, they showed much weaker trust learning, but when a similar interference task was used with the same faces shifting their gaze (Experiment 4.3) or reversed gaze cueing (Experiment 3.5), the learning was much stronger.

The other filter, pathway B., which filters information from the invariant stream, is more difficult to identify from the current thesis. However, Experiment 5.1 suggest that when alternative sources of social information are made salient − such as when participants are explicitly assigned to minimal social groups − that this causes a shift in the weights assigned to different inputs, as participants default to judging the face based on who they are in the social group context, rather than considering their individual past gaze behaviour. This could be due to the fact that this information is a more straightforward heuristic tool for making their decision, rather than the computationally more expensive process of learning from gaze shifts. This also means that this experiment found no evidence to support pathway D. in the model, which suggests that stored identity information (e.g. knowledge of an individual's behaviour) might affect processing of invariant aspects of their identity (such as their social group).

On the other hand, evidence from using more natural social groups such as race gives some interesting insights into pathway E., which suggests communication between the two streams where processing of one type of information is affected by the content of the other. This was partly contradicted by the results of Experiment 6.1, which found that while group membership did affect learning of trust from gaze cues (which suggests that information from the invariant stream is affecting how gaze cues are integrated into the stored identity representation) this did not affect actual cueing, as cueing costs were similar for own- and other-race faces. As such, it seems that information about race is more likely to affect pathway A. (the variant filter), rather than pathway E. (earlier interference between the two streams).

Pathway C., which suggests that information from the stored identity representation may be able to feed back to affect how eye gaze information is initially processed, is also not supported. Although only Experiments 2.1 and 3.5 report cueing effects broken down by block within the main text, Appendix A reports results from all experiments broken down to see whether there is any evidence that cueing costs recover over time − that is, as participants learn the cueing behaviour of faces more and more, if this results in their being able to anticipate and compensate for invalid cues in the later blocks. However, no reliable evidence to support this interpretation emerges, which suggests that as participants are learning these cueing behaviours they are unable to influence earlier gaze processing. This is despite evidence that real world gaze cueing (e.g. feinting in basketball) can be moderated by experience (Weigelt, Güldenpenning, Steggemann-Weinrich, Alaboud & Kunde, 2016). There is also the point that cueing times in Experiment 3.5, where the relationship between valid and invalid faces was reversed in the sixth block, did not show an overall cost, which one might expect if

participants had been implicitly learning the pattern to be able to compensate for it (Knopman & Nissen, 1991; Reed & Johnson, 1994).

This model was generated as a tentative framework for describing how gaze information might be incorporated into identity representations and used to inform later judgements, and was intentionally oversimplified to give scope for expansion. The results of this thesis clearly show that this process is much more complex than described here, and likely more complex than can be easily summarised in a box-and-arrows figure, as the social learning system not only weights its modules differently according to the social context, but how the memories are formed, stored, and accessed appear to be important questions to which this thesis can provide only hints. There are several questions left outstanding before a coherent model of incidental social learning can be generated.

For example, one question is the issue of set capacity. Experiments throughout this thesis have used a set of 16 faces, as this is in line with previous findings, but we do not know if this accurately reflects how many faces people can learn about in situ. Manssuer, Roberts and Tipper (2015) reported ERP results that suggest that participants need five to six exposures to a face before the late positive potential (LPP) begins to discriminate validity, and so we have an idea of how much exposure we need to learn about a face, but no idea of how many faces we can learn about. Falvello et al. (2015) found that participants were able to learn up to 400 items when asked to learn associations of faces with explicit behaviour descriptions. However, this latter task required explicit active learning; in contrast, it seems unlikely that the in-the-moment implicit learning in our gaze-cueing studies, where faces are irrelevant and to-be-ignored, would have such a large capacity.

Another outstanding question relates to the extinction of trust learning in

Experiment 5.1. This seems to be driven by the explicit nature of the minimal groups manipulation. We did in fact collect some additional data from a similar experiment (not reported in this thesis) where we kept all references to separate groups of faces but simply did not assign participants to be a part of them, thereby creating two 'out-groups' in the stimuli, and we found that this too resulted in extinction of trust learning. This could suggest that making a group-level representation salient is enough to make participants use it as a heuristic, even when it is not personally relevant. Given that trust learning was observed when race was the key manipulation (and throughout the thesis when faces of different sexes were used), a potentially interesting follow-up study could be to replicate Experiment 6.1 but make the distinction of the two races experimentally salient, to see if this is also capable of extinguishing individual trust learning.

An avenue for future research might be to investigate how incidental learning of gaze cues might emerge in different populations, such as individuals with autism. Autism is characterised by a set of symptoms relating to poor social and communicative skills, as well as repetitive or obsessive behaviours or interests (Baron-Cohen, 2004). The condition is also associated with gaze processing deficits, that appear to be due to the interpretation and use of gaze information to coordinate with others (Pelphrey, Morris & McCarthy, 2005). It would be interesting to see how individuals who meet the clinical criteria for Autism Spectrum Disorders (ASD) would behave in the context of this gaze-cueing paradigm. There is evidence that individuals with ASD are sensitive to gaze direction and do show cueing effects (Kuhn et al., 2010), but these are atypical compared with controls (Freeth, Chapman, Ropar & Mitchell, 2010). As for incidental learning from these gaze cues, it is difficult to predict. While Bayliss and Tipper (2006) did show a negative correlation between incidental trust learning from gaze cues and autistic-like

traits, there is other evidence that individuals with ASD can use other social cues such as emotion to inform trustworthiness judgements (Caulfield, Ewing, Burton, Avard & Rhodes, 2014), and can also use gaze direction to bias memory and preference judgements for features of a scene in a similar way to controls (Freeth, Ropar, Chapman & Mitchell, 2010). It would therefore be interesting to investigate this in a clinical population, particularly using the more sensitive scalar ratings measures employed in the current experiments.

8.3 Summary

This thesis investigated incidental learning of trust from gaze cues. Over the course of six experimental chapters, eighteen experiments probed the limitations and features of this effect and found several key and some surprising points. Firstly, this learning is driven primarily by decreases in trustworthiness for invalid faces, and this is particularly prevalent when using neutral faces, which suggests that this learning is geared towards detecting cheaters and deceptive interaction partners. Learning is specific to trust, and appears to happen outside of conscious awareness. Memory for trustworthiness is surprisingly durable and is sensitive to the initial familiarity with the faces, but this memory does not impact subsequent perception of or memory for the physical features of faces. Learning is sensitive to the context of social groups − when faces belong to naturally occurring in- and out-groups participants appear to be sensitive to this and learn better about in-groups than out-groups, but if this distinction is part of an explicit experimental manipulation as when participants are assigned to minimal groups then this can override learning about individual behaviours. We also found that this phenomenon cannot be explained purely by visuomotor fluency, as learning did not emerge in a

task-switching procedure where there were no physical changes to the faces. Finally, even though trust learning is driven by patterns of eye-gaze, as trust learning develops this does not feedback and affect the attentional orienting triggered by the eye-gaze.

Appendices

A Results of conventional statistics

This appendix contains the tables of conventional ANOVAs, intended to supplement the analysis used throughout the thesis and to help interpretation of effects. All ANOVAs were run using the `ez` package in `R`. Partial eta squared (η_P^2) has been calculated by hand using the output of this analysis. For those analyses where the results of ANOVAs are included in the main text (i.e. trustworthiness ratings in Chapters 5 and 6), these are not repeated.

All analyses performed here include block as a factor, and this is done to show that although experiments may occasionally demonstrate an interaction of validity and block that may suggest participants are learning to overcome invalid cues (c.f. Experiment 3.5), when looking at the larger picture this is clearly not the case. We also provide figures of RTs and accuracy rates broken down across experimental blocks where these figures have not been included in the main text, to aid clarity of communication.

Note that where corrections for the violation of sphericity assumptions have been used (i.e. where degrees of freedom are not integers), the correction is Greenhouse-Geisser.

Chapter 2

Table A.1: Experiment 2.1: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	144496394.70	4749159.60	699.79	0.00	*	0.97
Block	1.92	44.16	896491.10	958578.90	21.51	0.00	*	0.48
Validity	1	23	100831.70	48378.60	47.94	0.00	*	0.68
Block:Validity	4	92	10619.70	215111	1.14	0.34		0.05

Table A.2: Experiment 2.1: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	1908167	16934.90	25915620483	0.00	*	0.99
Block	4	92	1527.34	10441.41	3.36	0.01	*	0.13
Validity	1	23	10.42	1575.52	0.15	0.70		0.01
Block:Validity	4	92	210.94	3320.31	1.46	0.22		0.06

Table A.3: Experiment 2.1: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	185680.04	34248.96	124.69	0.00	*	0.84
Time	1	23	1134.38	4917.62	10.36	0.00	*	0.19
Validity	1	23	3725.04	8266.96	10.36	0.00	*	0.31
Time:Validity	1	23	1584.38	4792.62	7.60	0.01	*	0.25

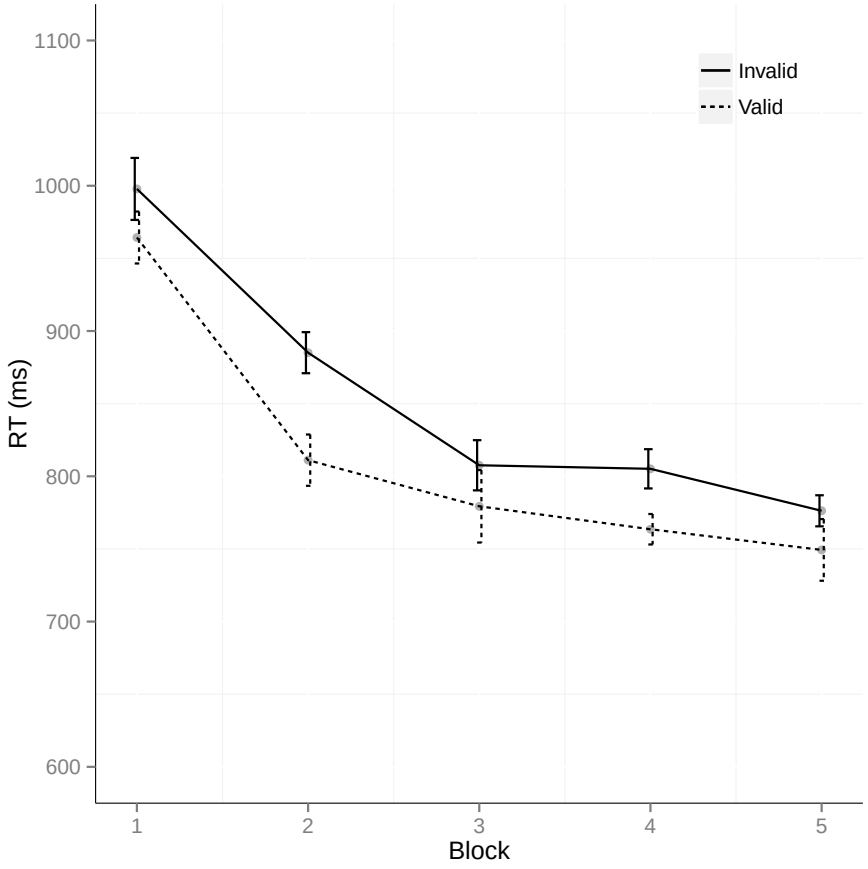

Figure A.1: Timecourse of reaction times in milliseconds across all five blocks in Experiment 2.2 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.4: Experiment 2.2: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	166918011.09	3030587.30	1266.79	0.00	*	0.98
Block	1.86	42.72	1488731.71	1092908.38	31.33	0.00	*	0.58
Validity	1	23	100035.26	89951.03	25.58	0.00	*	0.53
Block:Validity	4	92	18072.67	339735.19	1.22	0.31		0.05

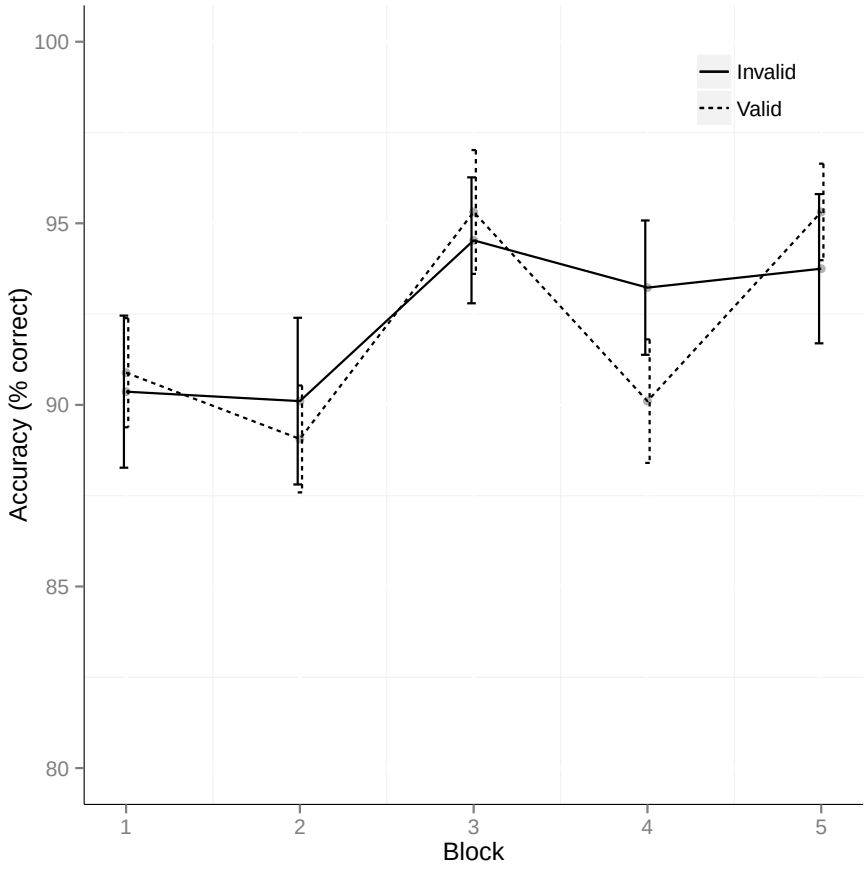

Figure A.2: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 2.2 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.5: Experiment 2.2: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	2043107	6443.85	7292.45	0.00	*	1.00
Block	4	92	1076.82	11970.05	2.07	0.09		0.08
Validity	1	23	4.07	1062.34	0.09	0.77		0.00
Block:Validity	4	92	166.02	3005.86	1.27	0.29		0.05

Table A.6: Experiment 2.2: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	5835.96	20220.54	6.64	0.02	*	0.22
Time	1	23	7.18	10691.01	0.02	0.90		0.00
Validity	1	23	1542.01	6467.89	5.48	0.03	*	0.19
Time:Validity	1	23	2516.38	4791.83	12.08	0.00	*	0.34

Table A.7: Experiments 2.1 and 2.2: Results of a 2x2 mixed factorial ANOVA on trustworthiness ratings across expression (smiling/neutral; between subjects) and validity (valid/invalid; within subjects)

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	46	1228.19	25744.41	2.19	0.15		0.05
Expression	1	46	872.27	25744.41	1.56	0.22		0.03
Validity	1	46	7356.56	17863.54	18.94	0.00	*	0.29
Expression:Validity	1	46	321.29	17863.54	0.83	0.37		0.02

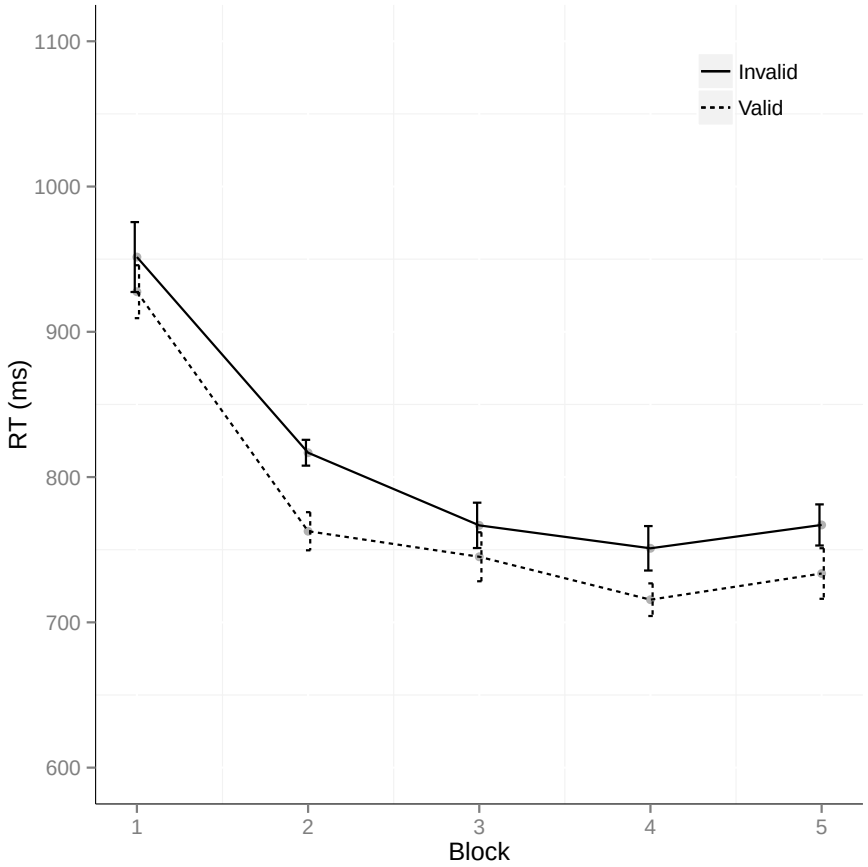

Figure A.3: Timecourse of reaction times in milliseconds across all five blocks in Experiment 2.3 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.8: Experiment 2.3: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	151217500	3158464.60	1101.17	0.00	*	0.98
Block	2.18	50.24	1355386	826177.10	37.73	0.00	*	0.62
Validity	1	23	68011.92	108231.10	14.45	0.00	*	0.39
Block:Validity	2.44	56.13	7913.63	334116.30	0.54	0.70		0.02

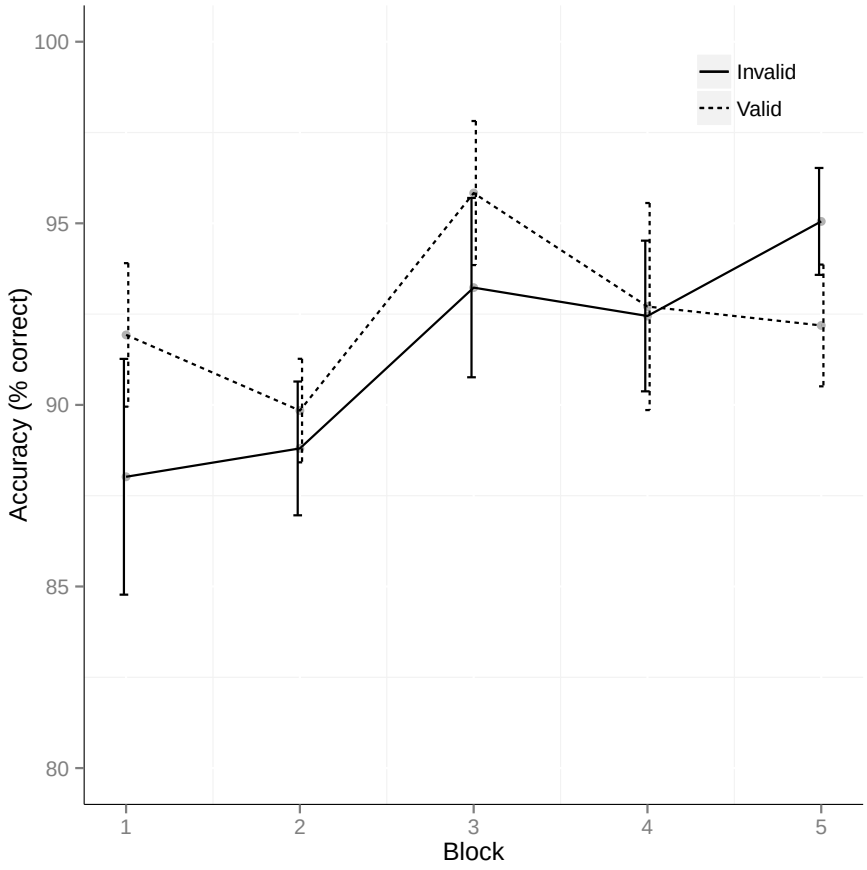

Figure A.4: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 2.3 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.9: Experiment 2.3: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	2031590	7413.90	6302.56	0.00		1.00
Block	2.52	58.00	990.56	18579.75	1.23	0.31		0.05
Validity	1	23	58.76	1288.90	1.05	0.32		0.04
Block:Validity	4	92	318.03	3627.28	2.02	0.10		0.08

Table A.10: Experiment 2.3: Results of a 2x2 (time x validity) factorial ANOVA on likeability ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	1345.32	5201.53	5.95	0.02	*	0.21
Time	1	23	13.59	1527	0.20	0.66		0.01
Validity	1	23	3.14	1671.84	0.04	0.84		0.00
Time:Validity	1	23	125.18	1276.56	2.26	0.15		0.09

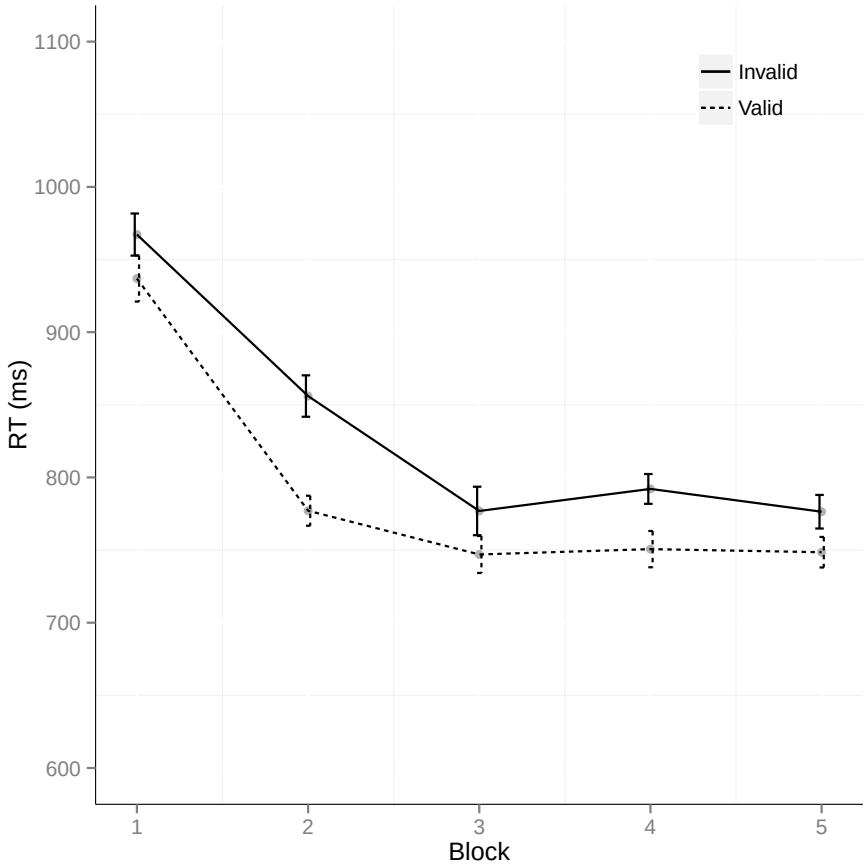

Figure A.5: Timecourse of reaction times in milliseconds across all five blocks in Experiment 2.4 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.11: Experiment 2.4: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	194028837.20	4605779.65	1221.69	0.00	*	0.98
Block	2.70	78.30	1532746.84	1423146.06	31.23	0.00	*	0.52
Validity	1	29	136805.24	177644.04	22.33	0.00	*	0.44
Block:Validity	4	116	28615.24	411962.29	2.01	0.10		0.06

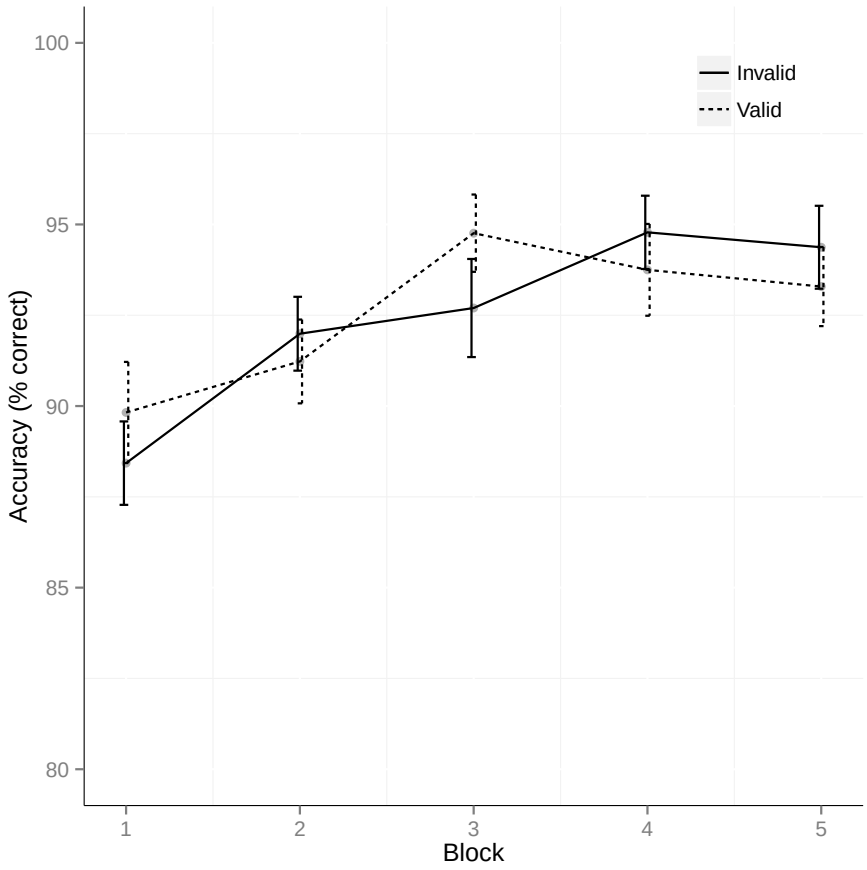

Figure A.6: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 2.4 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.12: Experiment 2.4: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	2530128.59	12347.50	5942.39	0.00	*	1.00
Block	4	116	1001.61	12793.80	2.27	0.07		0.07
Validity	1	29	2.71	922	0.09	0.77		0.00
Block:Validity	4	116	205.82	4845.69	1.23	0.30		0.04

Chapter 3

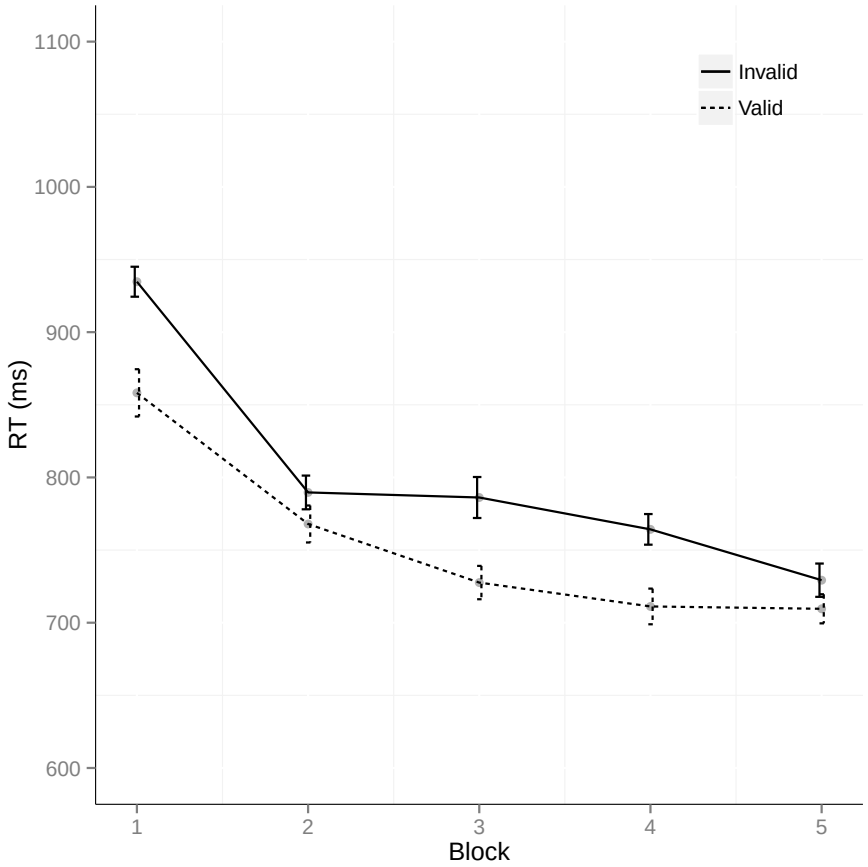

Figure A.7: Timecourse of reaction times in milliseconds across all five blocks in Experiment 3.1 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.13: Experiment 3.1: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	181702558.03	3066083.22	1718.60	0.00	*	0.98
Block	2.01	58.18	1187893.28	1541979.38	22.34	0.00	*	0.44
Validity	1	29	161094.12	124785.72	37.44	0.00	*	0.56
Block:Validity	4	116	39311.80	453493.56	2.51	0.05	*	0.08

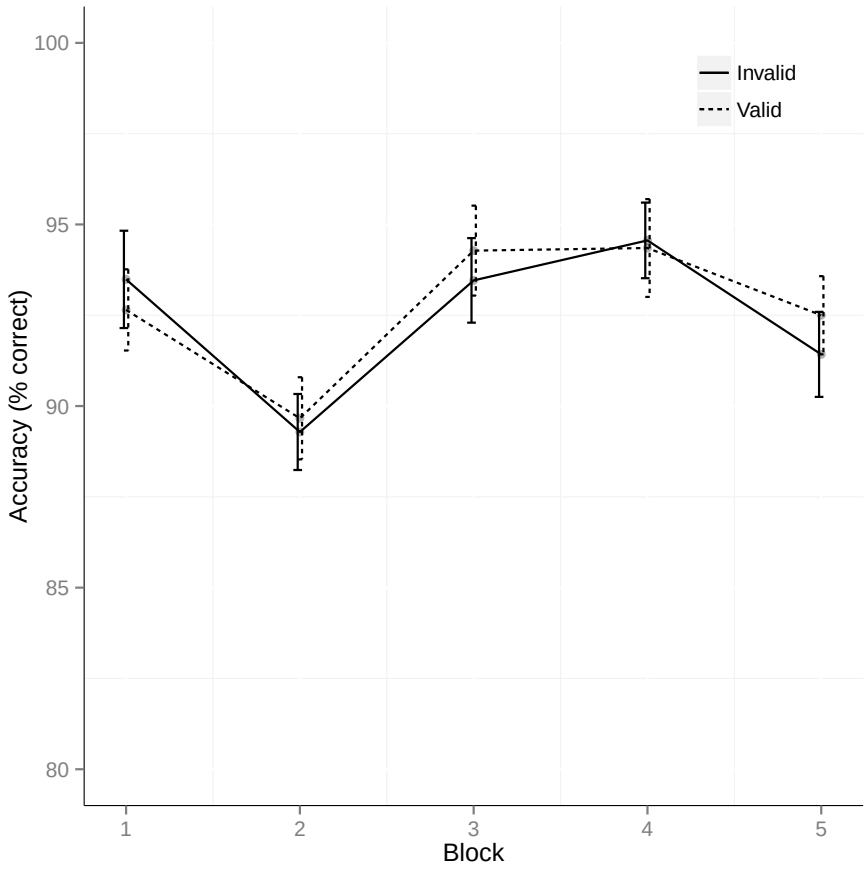

Figure A.8: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 3.1 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.14: Experiment 3.1: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	2565663.83	6281.76	11844.49	0.00	*	1.00
Block	4	116	952.72	11556.49	2.39	0.05		0.08
Validity	1	29	5.62	1109.83	0.15	0.70		0.01
Block:Validity	3.35	97.23	38.08	3491	0.32	0.83		0.01

Table A.15: Experiment 3.1: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	556.31	36335.03	0.44	0.51		0.02
Time	1	29	115.79	3248.28	1.03	0.32		0.03
Validity	1	29	1429.16	7190.22	5.76	0.02	*	0.17
Time:Validity	1	29	1088.27	4893.39	6.45	0.02	*	0.18

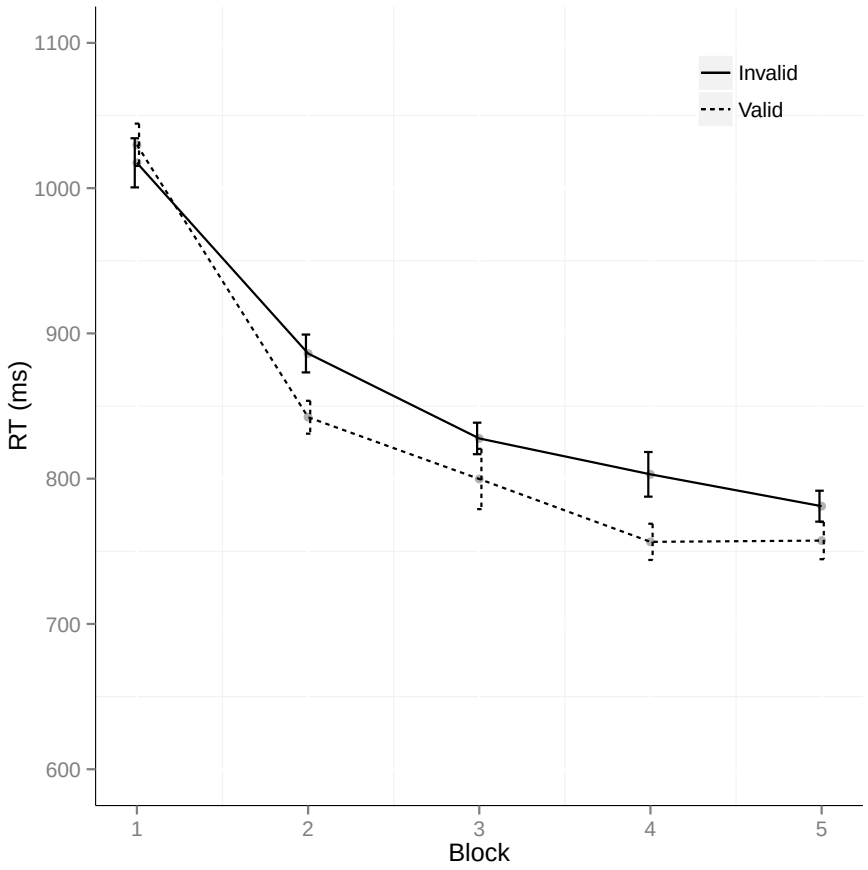

Figure A.9: Timecourse of reaction times in milliseconds across all five blocks in Experiment 3.2 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.16: Experiment 3.2: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	216425386.21	6905925.56	908.83	0.00	*	0.97
Block	2.66	77.23	2565372.43	1594537.15	46.66	0.00	*	0.62
Validity	1	29	50569.08	194808.87	7.53	0.01	*	0.21
Block:Validity	2.81	81.42	35177.11	671721.96	1.52	0.22		0.05

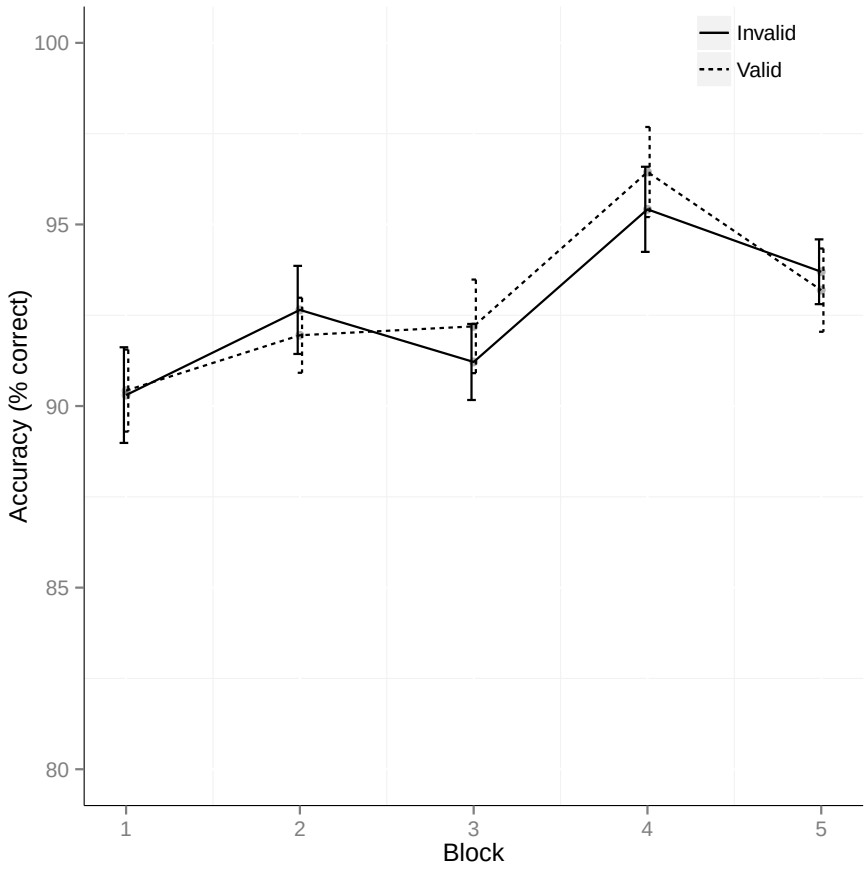

Figure A.10: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 3.2 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.17: Experiment 3.2: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	2571865.82	7484.33	9965.37	0.00	*	1.00
Block	3.34	96.94	1153.84	10185.58	3.29	0.02	*	0.10
Validity	1	29	1.17	1438.28	0.02	0.88		0.00
Block:Validity	4	116	45.06	4245.98	0.31	0.87		0.01

Table A.18: Experiment 3.2: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	266.26	62697.53	0.12	0.73		0.00
Time	1	29	402.42	5461.38	2.14	0.15		0.07
Validity	1	29	1038.41	6261.17	4.81	0.04	*	0.14
Time:Validity	1	29	280.60	6532.82	1.25	0.27		0.04

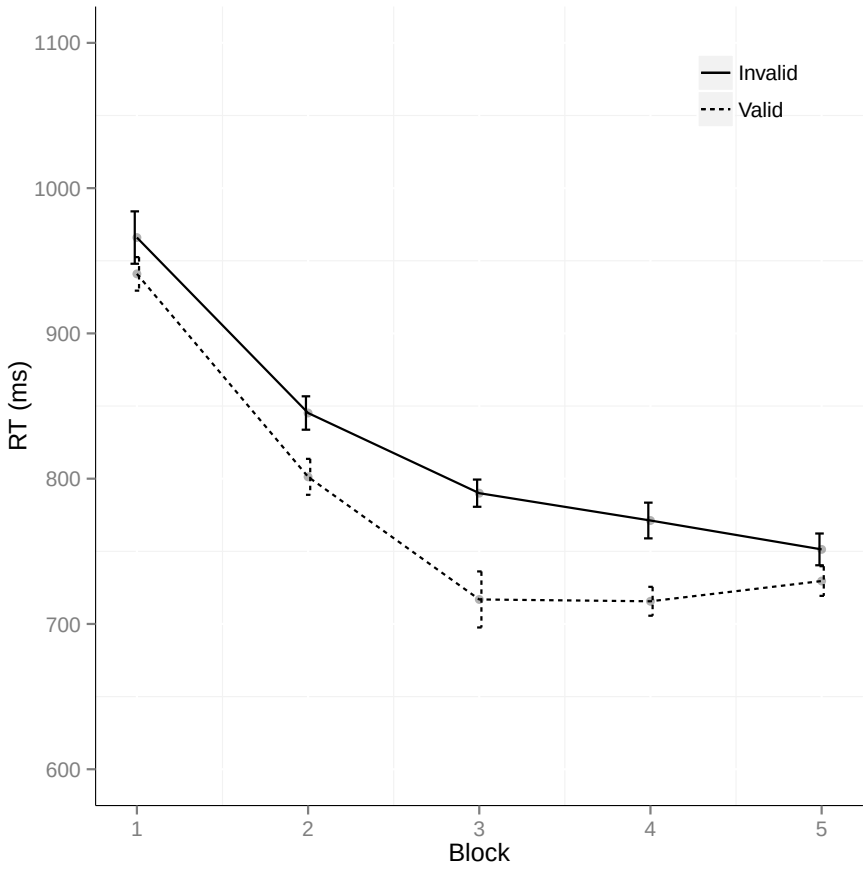

Figure A.11: Timecourse of reaction times in milliseconds across all five blocks in Experiment 3.3 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.19: Experiment 3.3: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	193230210.02	4113244.89	1362.35	0.00	*	0.98
Block	2.50	72.52	1980669.40	1141249.64	50.33	0.00	*	0.63
Validity	1	29	142786.66	124887.58	33.16	0.00	*	0.53
Block:Validity	2.73	79.18	28178.58	673214.09	1.21	0.31		0.04

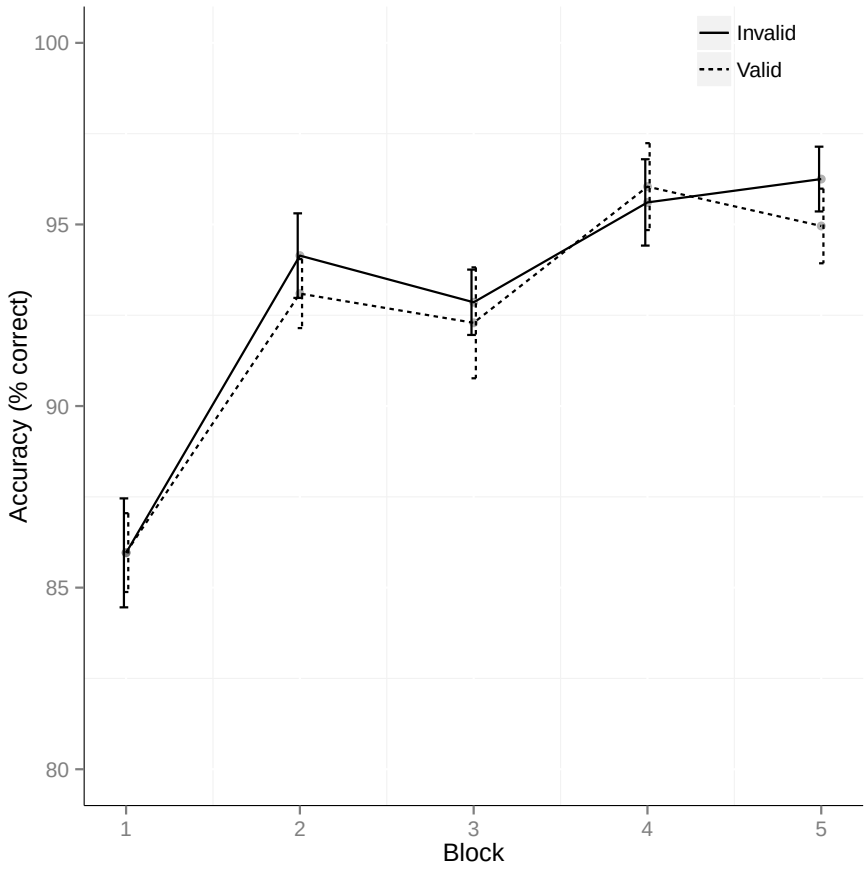

Figure A.12: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 3.3 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.20: Experiment 3.3: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	2568031.38	3975.41	18733.37	0.00	*	1.00
Block	2.79	80.86	4520.88	14287.55	9.18	0.00	*	0.24
Validity	1	29	25.69	991.52	0.75	0.39		0.03
Block:Validity	4	116	32.27	3794.76	0.25	0.91		0.01

Table A.21: Experiment 3.3: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	656.84	10802.60	1.76	0.19		0.06
Time	1	29	103.14	3135.11	0.95	0.34		0.03
Validity	1	29	1456.03	10567.39	4	0.06		0.12
Time:Validity	1	29	1445.60	4644.79	9.03	0.01	*	0.24

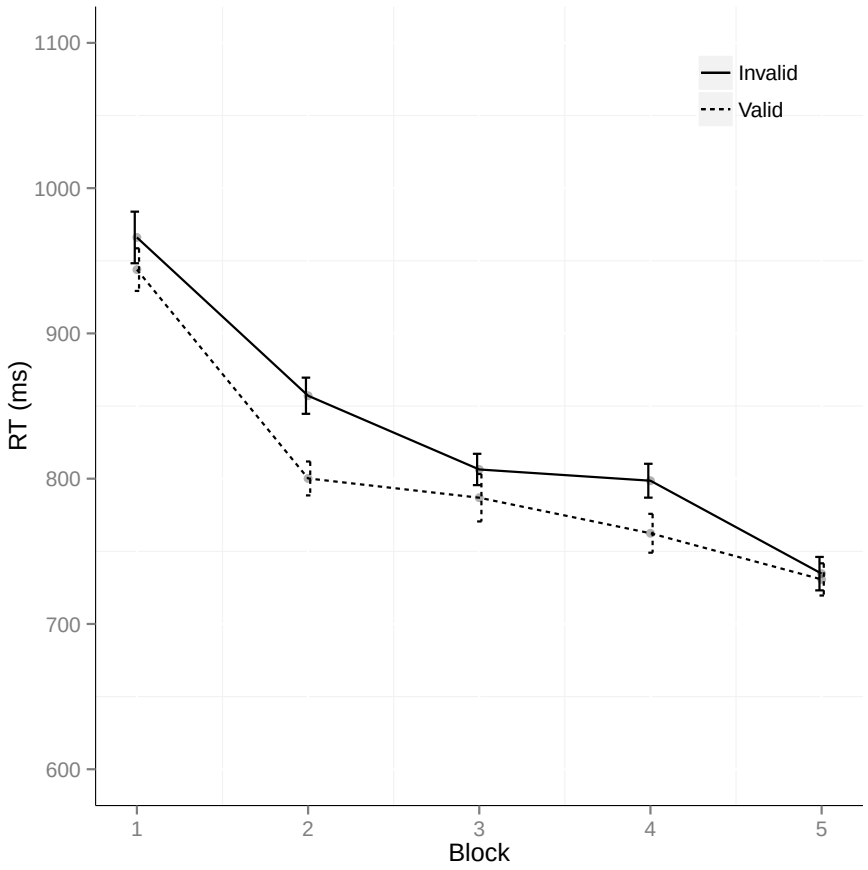

Figure A.13: Timecourse of reaction times in milliseconds across all five blocks in Experiment 3.4 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.22: Experiment 3.4: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	201528160.75	5112404.03	1143.16	0.00	*	0.98
Block	4	116	1722498.23	1197859.08	41.70	0.00	*	0.59
Validity	1	29	63196.84	88401.25	20.73	0.00	*	0.42
Block:Validity	4	116	28268.28	469156.19	1.75	0.14		0.06

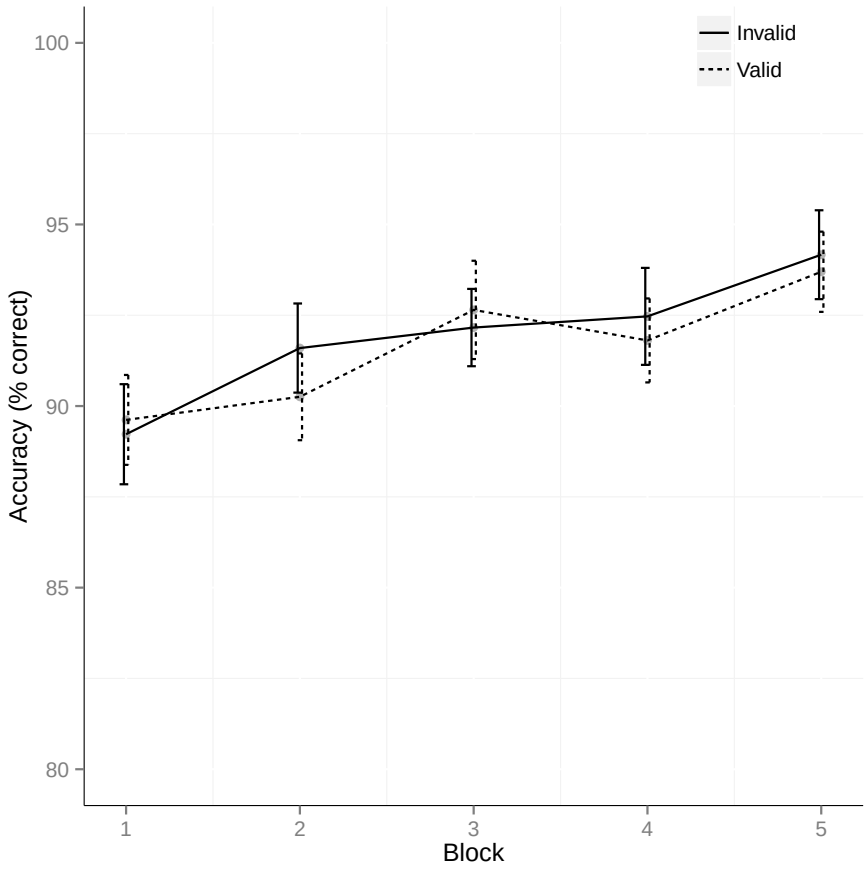

Figure A.14: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 3.4 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.23: Experiment 3.4: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	2520778.77	6005.78	12172.03	0.00	*	1.00
Block	4	116	733.93	9169.22	2.32	0.06		0.07
Validity	1	29	8.84	1446.21	0.18	0.68		0.01
Block:Validity	4	116	42.01	4223.39	0.29	0.88		0.01

Table A.24: Experiment 3.4: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	758.78	18637.01	1.18	0.29		0.04
Time	1	29	76	3781.65	0.58	0.45		0.02
Validity	1	29	245.82	7935.10	0.90	0.35		0.03
Time:Validity	1	29	343.41	3215.24	3.10	0.09		0.10

Table A.25: Experiment 3.5: Results of a 2x5 (validity x block) factorial ANOVA on reaction times across all six blocks

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	236791305.16	6599722.71	1040.49	0.00	*	0.97
Block	2.57	74.53	1421389.09	2295398.17	17.96	0.00	*	0.38
Validity	1	29	133569.96	193472.64	20.02	0.00	*	0.41
Block:Validity	3.88	112.62	67146.71	554109.08	3.51	0.01	*	0.11

Table A.26: Experiment 3.5: Results of a 2x5 (validity x block) factorial ANOVA on reaction times over the final two blocks of gaze cueing, where cueing behaviour reversed

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	67869665.45	2021687.68	973.55	0.00	*	0.97
Block	1	29	11728.69	167398.36	2.03	0.16		0.07
Validity	1	29	4988.85	112268.57	1.29	0.27		0.04
Block:Validity	1	29	1034.73	73528.27	0.41	0.53		0.01

Table A.27: Experiment 3.5: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates across all six blocks

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	1238.83	3.12	11515.46	0.00	*	1.00
Block	5	145	0.54	5.79	2.70	0.02	*	0.09
Validity	1	29	0.02	0.56	0.88	0.36		0.03
Block:Validity	3.79	109.91	0.20	2	2.83	0.03	*	0.09

Table A.28: Experiment 3.5: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates over the final two blocks of gaze cueing, where cueing behaviour reversed

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	426.55	1.87	6602.13	0.00	*	1.00
Block	1	29	0.04	0.80	1.54	0.23		0.05
Validity	1	29	0.00	0.23	0.14	0.71		0.00
Block:Validity	1	29	0.04	0.64	1.62	0.21		0.05

Table A.29: Experiment 3.5: Results of a 2x2 (time x validity) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	1597.79	28862.52	1.61	0.22		0.05
Time	1	29	2.30	5325.24	0.01	0.91		0.00
Validity	1	29	262.92	4961.78	1.54	0.23		0.05
Time:Validity	1	29	1042.09	4217.09	7.17	0.01	*	0.20

Chapter 4

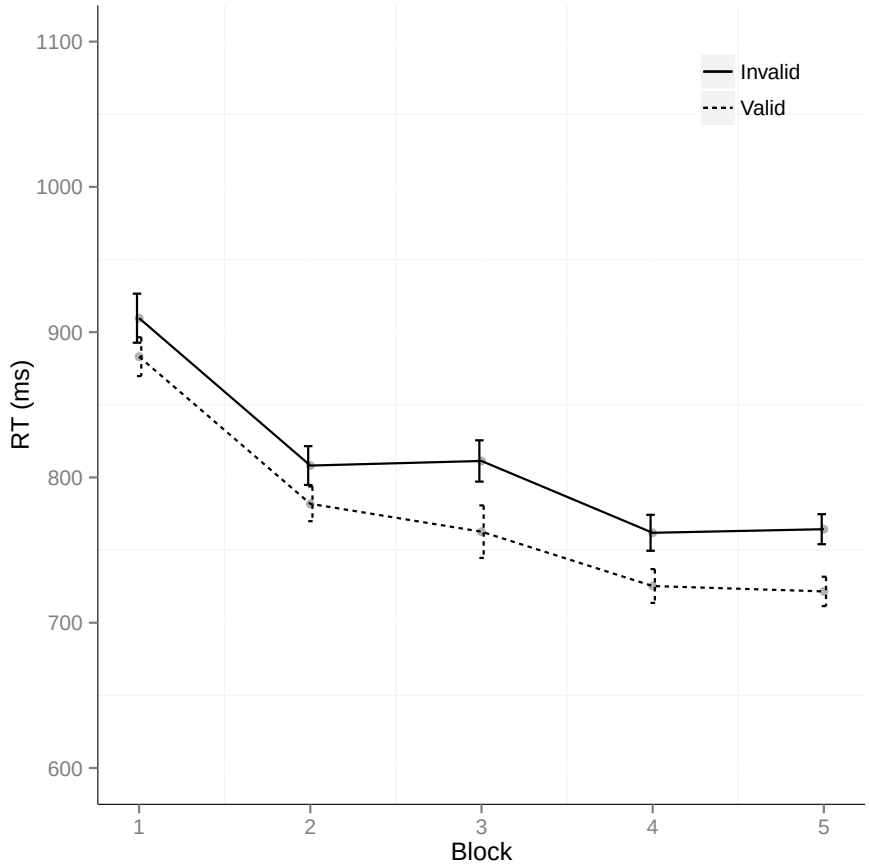

Figure A.15: Timecourse of reaction times in milliseconds across all five blocks in Experiment 4.1 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.30: Experiment 4.1: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	150908973.74	2709866.85	1280.84	0.00	*	0.98
Block	1.84	42.41	755947.70	976305.87	17.81	0.00	*	0.44
Validity	1	23	77245.22	97889.84	18.15	0.00	*	0.44
Block:Validity	4	92	5411.05	428946.77	0.29	0.88		0.01

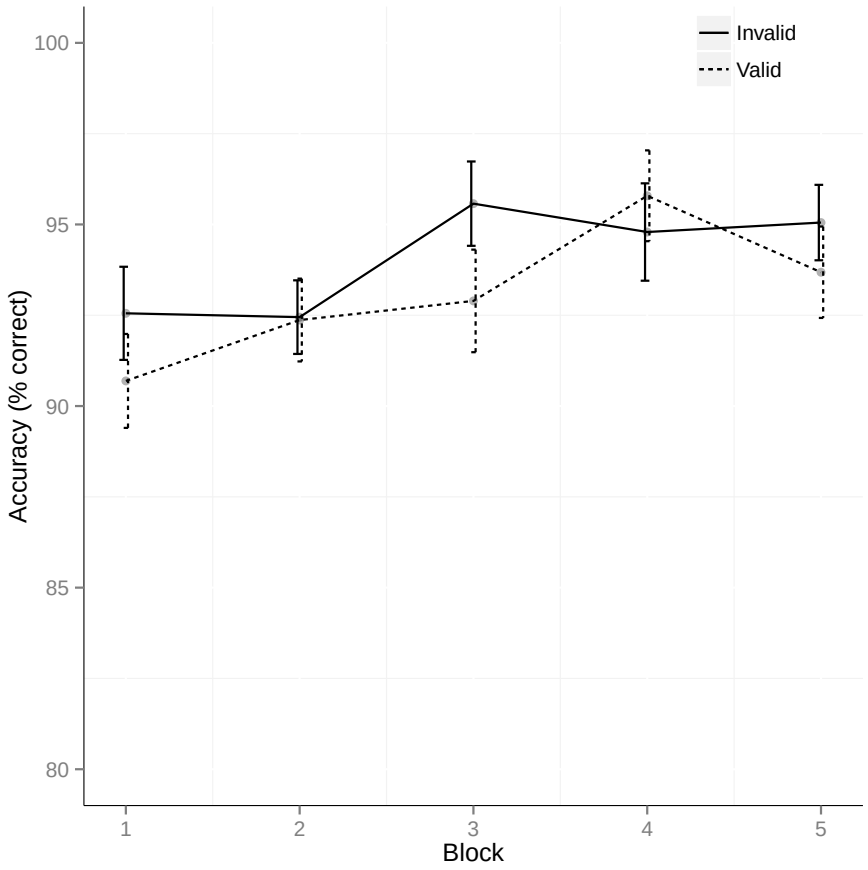

Figure A.16: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 4.1 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.31: Experiment 4.1: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	838.28	2.22	8671.39	0.00	*	1.00
Block	2.71	62.38	0.22	3.37	1.47	0.23		0.06
Validity	1	23	0.03	0.34	1.70	0.21		0.07
Block:Validity	4	92	0.04	1.23	0.70	0.59		0.03

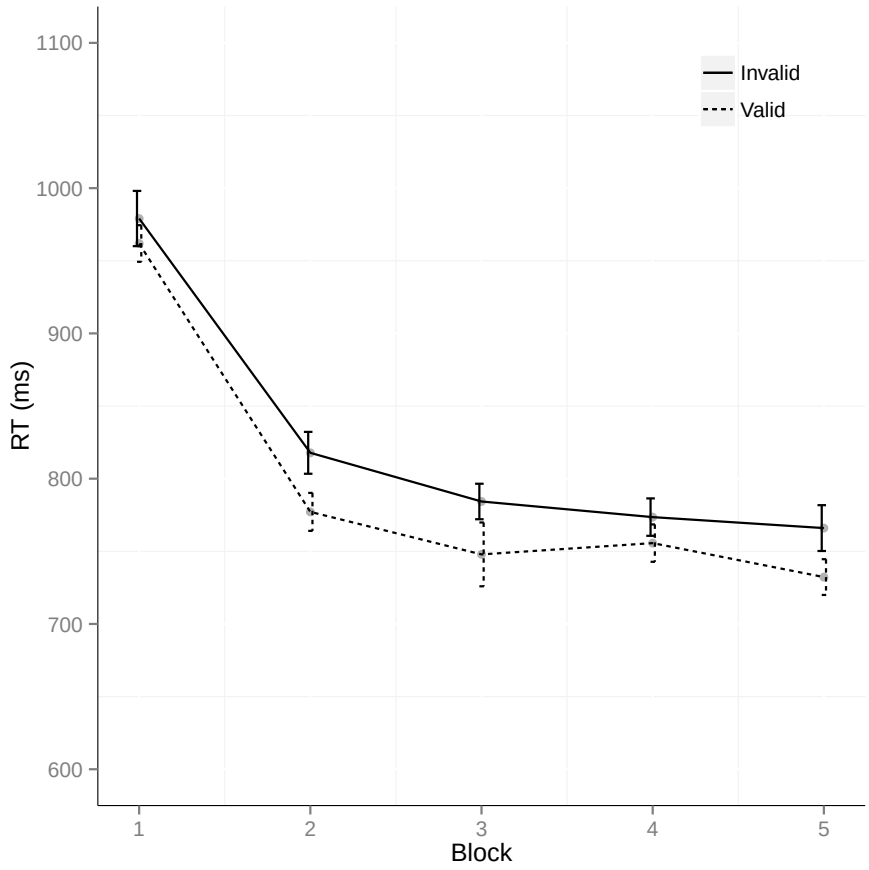

Figure A.17: Timecourse of reaction times in milliseconds across all five blocks in Experiment 4.2 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.32: Experiment 4.2: Results of a 2x5 (validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	22	150813672.15	4954058.99	669.73	0.00	*	0.97
Block	2.57	56.58	1566289.61	965350.49	35.70	0.00	*	0.62
Validity	1	22	44151.50	129782.10	7.48	0.01	*	0.25
Block:Validity	2.09	46	5973.05	383126.35	0.34	0.72		0.02

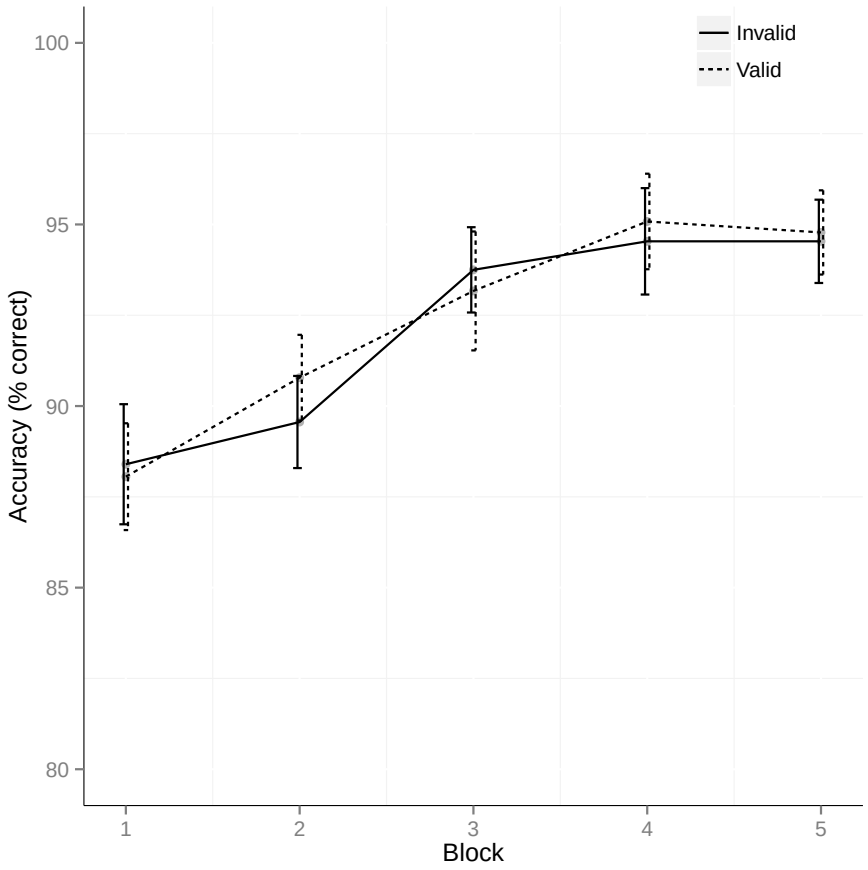

Figure A.18: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 4.2 in response to valid (dotted) and invalid (solid line) trials. Error bars show standard error.

Table A.33: Experiment 4.2: Results of a 2x5 (validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	22	781.46	1.82	9426.83	0.00	*	1.00
Block	4	88	0.67	3.17	4.64	0.00	*	0.17
Validity	1	22	0.00	0.18	0.04	0.84		0.00
Block:Validity	4	88	0.01	1.40	0.11	0.98		0.00

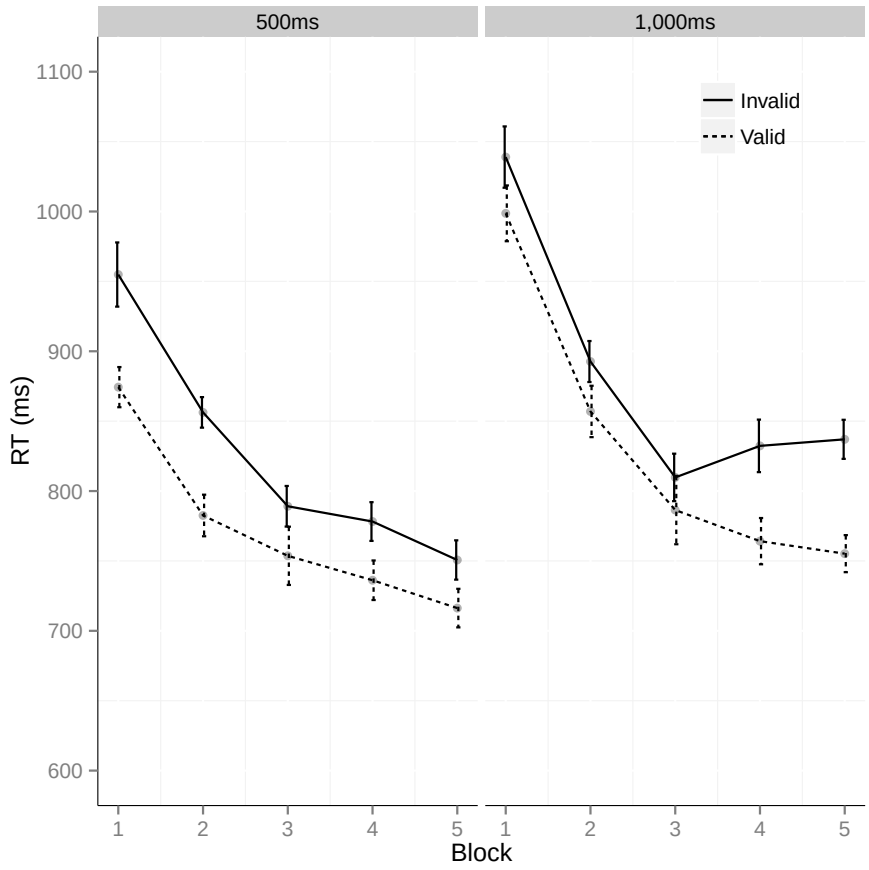

Figure A.19: Timecourse of reaction times in milliseconds across all five blocks in Experiment 4.3 in response to valid (dotted) and invalid (solid line) trials in the short SOA (500ms, left plot) and long SOA conditions (1,000ms, right plot). Note that SOA as a factor only affected the paradigm after cueing had ended, and so these conditions were identical at the point that these data were collected. Error bars show standard error.

Table A.34: Experiment 4.3: Results of a 2x2x5 (SOA x validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	38	2.74e+08	4121431.19	2529.72	0.00	*	0.99
SOA	1	38	349204.07	4121431.19	3.22	0.08		0.08
Block	2.29	86.88	2.24e+06	1953630.34	43.60	0.00	*	0.53
Validity	1	38	269136.86	227815.07	44.89	0.00	*	0.54
SOA:Block	2.29	86.88	66317.96	1953630.34	1.29	0.28		0.03
SOA:Validity	1	38	378.09	227815.07	0.06	0.80		0.00
Block:Validity	3.25	123.36	12725.64	677135.74	0.71	0.56		0.02
SOA:Block:Validity	3.25	123.36	33154.19	677135.74	1.86	0.14		0.05

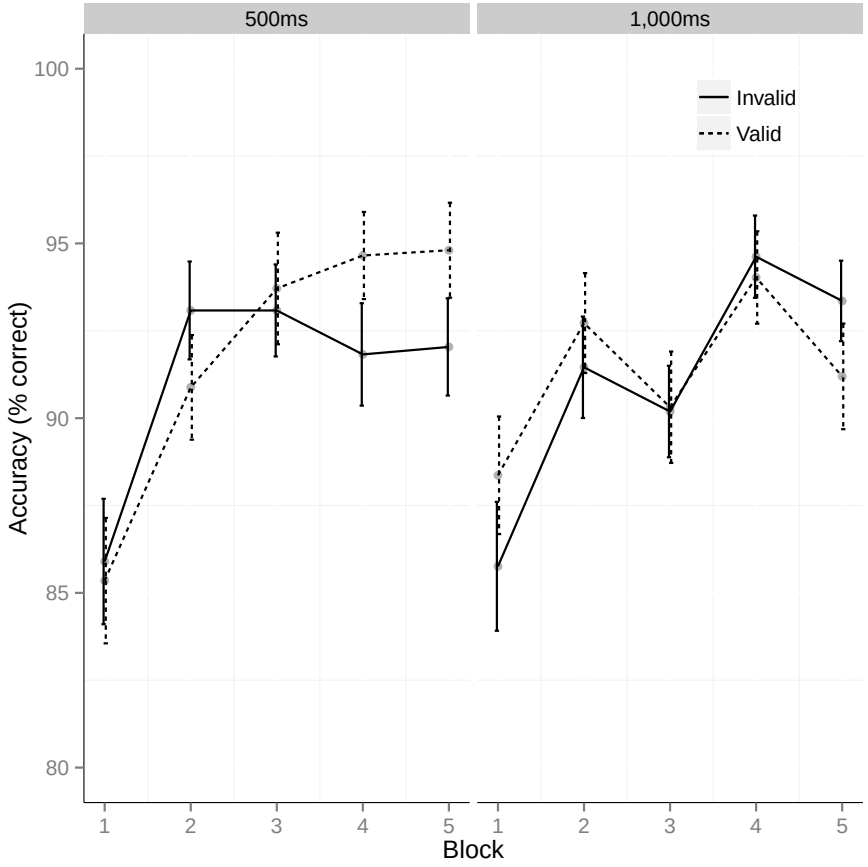

Figure A.20: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 4.3 in response to valid (dotted) and invalid (solid line) trials in the short SOA (500ms, left plot) and long SOA conditions (1,000ms, right plot). Note that SOA as a factor only affected the paradigm after cueing had ended, and so these conditions were identical at the point that these data were collected. Error bars show standard error.

Table A.35: Experiment 4.3: Results of a 2x2x5 (SOA x validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	38	1331.21	3.68	13755.06	0.00	*	1.00
SOA	1	38	0.00	3.68	0.03	0.87		0.00
Block	4	152	1.08	6.48	6.33	0.00	*	0.14
Validity	1	38	0.01	0.53	0.73	0.40		0.02
SOA:Block	4	152	0.11	6.48	0.66	0.62		0.02
SOA:Validity	1	38	0.00	0.53	0.07	0.80		0.00
Block:Validity	4	152	0.02	2.70	0.25	0.91		0.01
SOA:Block:Validity	4	152	0.10	2.70	1.41	0.23		0.04

Table A.36: Experiment 4.3: Results of a 2x2x2 (SOA x validity x time) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	38	5428.90	26030.87	7.93	0.01	*	0.17
SOA	1	38	0.62	26030.87	0.00	0.98		0.00
Time	1	38	1.70	8815.69	0.01	0.93		0.00
Validity	1	38	1003.75	6262.50	6.09	0.02	*	0.14
SOA:Time	1	38	411.20	8815.69	1.77	0.19		0.04
SOA:Validity	1	38	174.83	6262.50	1.06	0.31		0.03
Time:Validity	1	38	570.97	4110.82	5.28	0.03	*	0.12
SOA:Time:Validity	1	38	25.40	4110.82	0.23	0.63		0.01

Table A.37: Experiment 4.3: Results of a 2x2 (SOA x validity) factorial ANOVA on image ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	38	1559.82	71.30	831.37	0.00	*	0.96
SOA	1	38	2.91	71.30	1.55	0.22		0.04
Validity	1	38	0.13	4.40	1.14	0.29		0.03
SOA:Validity	1	38	0.41	4.40	3.57	0.07		0.09

Chapter 5

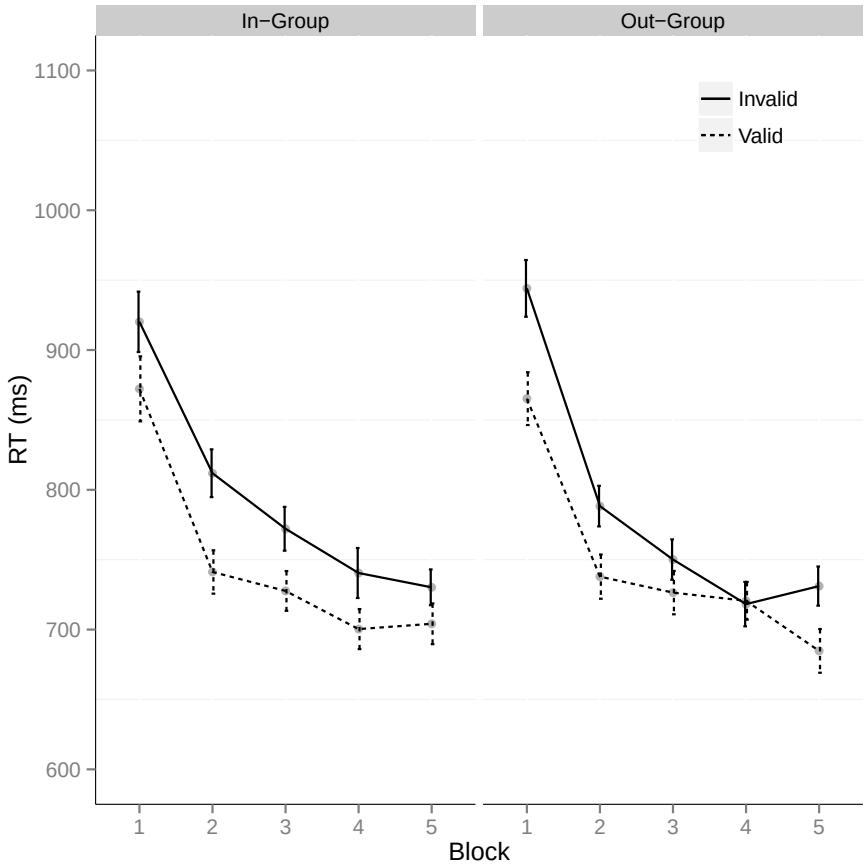

Figure A.21: Timecourse of reaction times in milliseconds across all five blocks in Experiment 5.1 in response to valid (dotted) and invalid (solid line) trials with in-group (left plot) and out-group members (right plot) as the cueing faces. Error bars show standard error.

Table A.38: Experiment 5.1: Results of a 2x2x5 (group x validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	3.55e+08	7.87e+06	1308.39	0.00	*	0.98
Block	2.65	76.79	2.78e+06	1.81e+06	44.50	0.00	*	0.61
Validity	1	29	256878.42	185140.81	40.24	0.00	*	0.58
Group	1	29	4600.28	137859.94	0.97	0.33		0.03
Block:Validity	2.76	80.05	36207.44	964282.39	1.09	0.36		0.04
Block:Group	2.93	84.99	10861.30	834019.30	0.38	0.76		0.01
Validity:Group	1	29	1274.22	145607.86	0.25	0.62		0.01
Block:Validity:Group	2.90	83.99	30575.44	1.22e+06	0.73	0.53		0.02

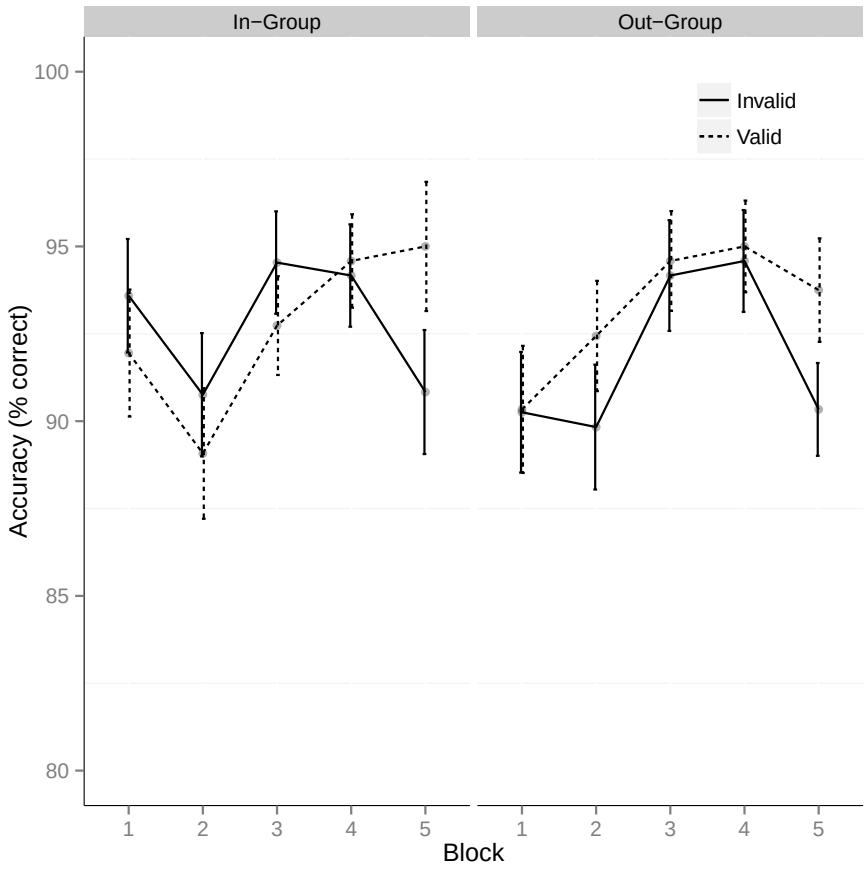

Figure A.22: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 5.1 in response to valid (dotted) and invalid (solid line) trials with in-group (left plot) and out-group members (right plot) as the cueing faces. Error bars show standard error.

Table A.39: Experiment 5.1: Results of a 2x2x5 (group x validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	5.14e+06	13315.86	11194	0.00	*	1.00
Block	4	116	1419.16	18653.76	2.21	0.07		0.07
Validity	1	29	56.02	2411	0.67	0.42		0.02
Group	1	29	4.17	4233.68	0.03	0.87		0.00
Block:Validity	4	116	428.07	6610.13	1.88	0.12		0.06
Block:Group	4	116	314.99	9223.21	0.99	0.42		0.03
Validity:Group	1	29	104.17	2276.04	1.33	0.26		0.04
Block:Validity:Group	4	116	135.71	7978.88	0.49	0.74		0.02

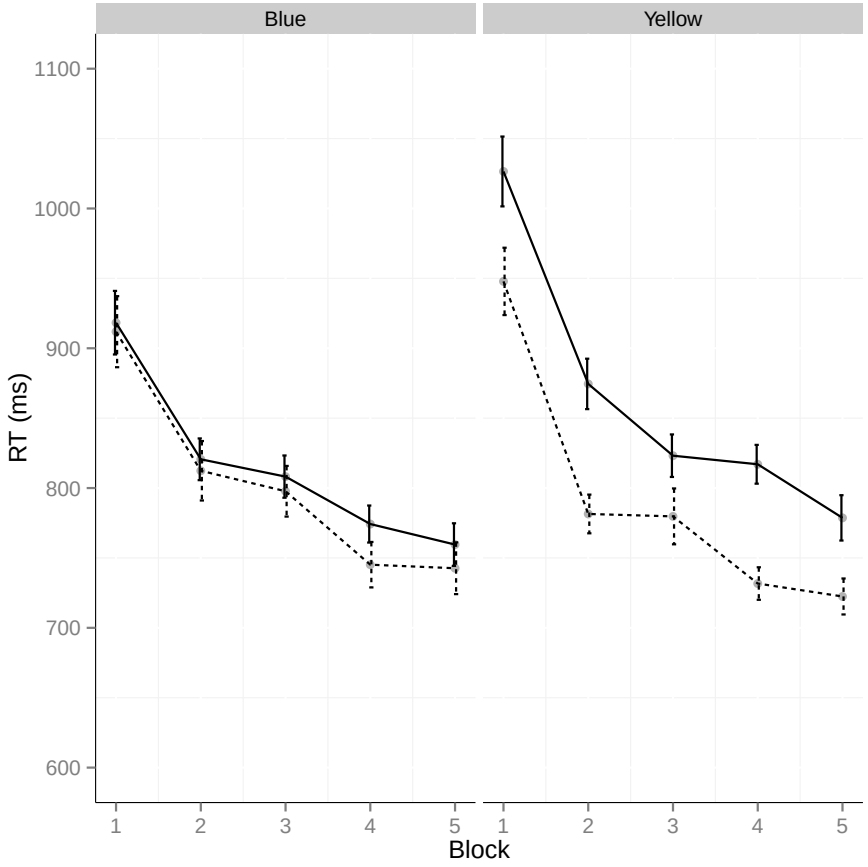

Figure A.23: Timecourse of reaction times in milliseconds across all five blocks in Experiment 5.2 in response to valid (dotted) and invalid (solid line) trials with faces wearing blue (left plot) and yellow shirts (right plot) as the cueing faces. Error bars show standard error.

Table A.40: Experiment 5.2: Results of a 2x2x5 (shirt colour x validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	4.02e+08	1.12e+07	1043.93	0.00	*	0.97
Block	2.26	65.50	3.10e+06	2.31e+06	38.91	0.00	*	0.57
Validity	1	29	269363.75	356110.60	21.94	0.00	*	0.43
Colour	1	29	60683.60	186645.20	9.43	0.00	*	0.25
Block:Validity	2.48	71.96	16215.89	1.29e+06	0.37	0.74		0.01
Block:Colour	4	116	113579.60	985068.90	3.34	0.01	*	0.10
Validity:Colour	1	29	117868.49	262405.70	13.03	0.00	*	0.31
Block:Validity:Colour	4	116	11771.29	835460.60	0.41	0.80		0.01

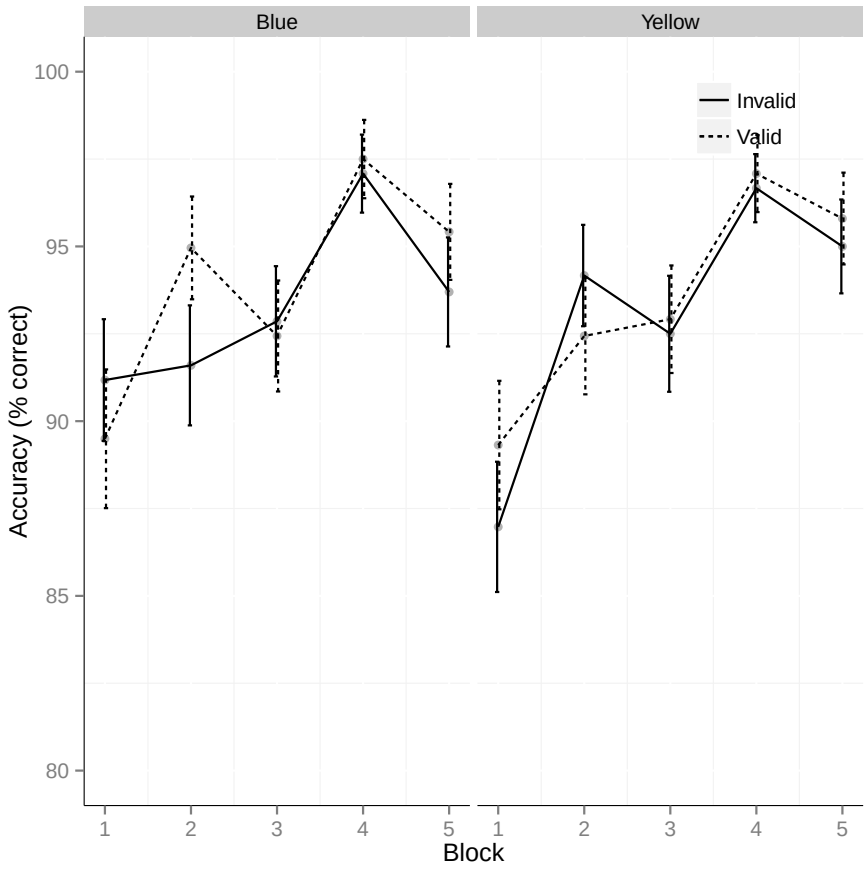

Figure A.24: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 5.2 in response to valid (dotted) and invalid (solid line) trials with faces wearing blue (left plot) and yellow shirts (right plot) as the cueing faces. Error bars show standard error.

Table A.41: Experiment 5.2: Results of a 2x2x5 (shirt colour x validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	29	5.23e+06	9718.63	15614.73	0.00	*	1.00
Block	3.01	87.27	4191.15	22253.30	5.46	0.00	*	0.16
Validity	1	29	46.30	1566.55	0.86	0.36		0.03
Colour	1	29	16.67	1891.32	0.26	0.62		0.01
Block:Validity	4	116	21.12	7020.54	0.09	0.99		0.00
Block:Colour	4	116	154.92	9213.14	0.49	0.74		0.02
Validity:Colour	1	29	4.17	2483.68	0.05	0.83		0.00
Block:Validity:Colour	4	116	338.14	7391.03	1.33	0.26		0.04

Chapter 6

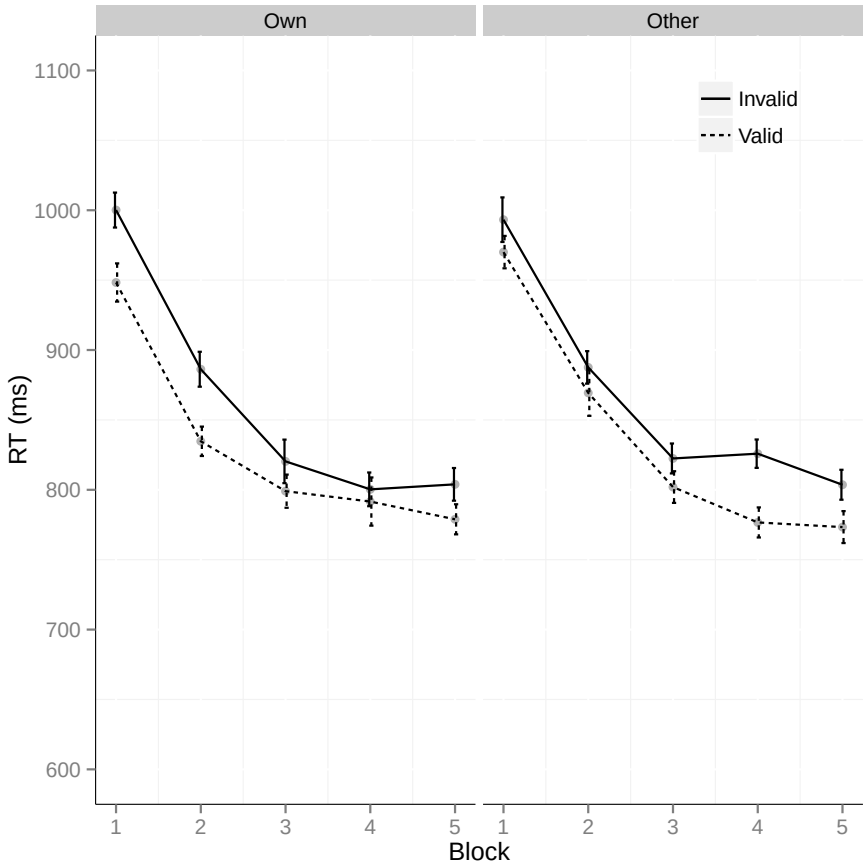

Figure A.25: Timecourse of reaction times in milliseconds across all five blocks in Experiment 6.1 in response to valid (dotted) and invalid (solid line) trials with own-race (left plot) and other-race individuals (right plot) as the cueing faces. Error bars show standard error.

Table A.42: Experiment 6.1: Results of a 2x2x5 (race x validity x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	59	8.66e+08	2.80e+08	1824.55	0.00	*	0.97
Block	2.80	165.13	6.00e+06	5.74e+06	61.59	0.00	*	0.51
Validity	1	59	285313.70	449543.20	37.45	0.00	*	0.39
Race	1	59	14212.36	613819.40	1.37	0.25		0.02
Block:Validity	4	236	8628.08	1.55e+06	0.33	0.86		0.01
Block:Race	3.08	181.80	16428.97	1721763	0.56	0.64		0.00
Validity:Race	1	59	3.91	528676.50	0.00	0.98		0.00
Block:Validity:Race	3.45	203.83	48620.95	2.25e+06	1.27	0.28		0.02

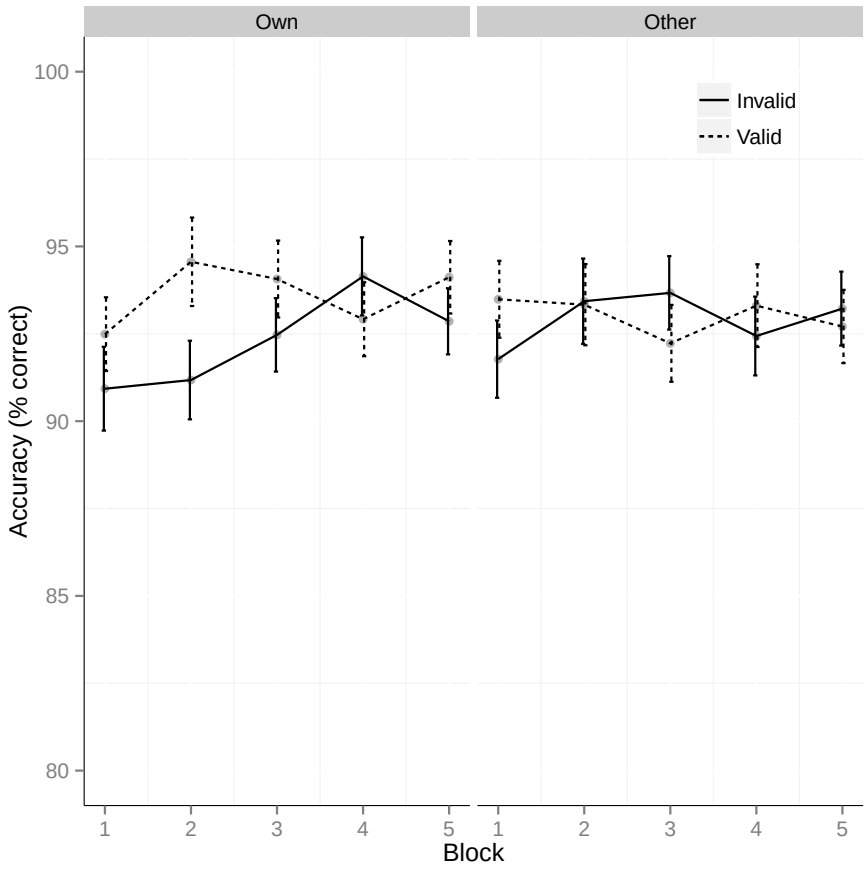

Figure A.26: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 6.1 in response to valid (dotted) and invalid (solid line) trials with own-race (left plot) and other-race individuals (right plot) as the cueing faces. Error bars show standard error.

Table A.43: Experiment 6.1: Results of a 2x2x5 (race x validity x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	59	1.03e+07	32094.10	19011.48	0.00	*	1.00
Block	4	236	260.91	35091.52	0.44	0.78		0.01
Validity	1	59	194.68	4779.28	2.40	0.12		0.04
Race	1	59	0.93	3497.34	0.02	0.90		0.00
Block:Validity	4	236	205.93	16747.19	0.73	0.58		0.01
Block:Race	4	236	127.11	16409.35	0.46	0.77		0.01
Validity:Race	1	59	75	3693.40	1.23	0.27		0.02
Block:Validity:Race	3.30	194.90	346.441	16019.88	1.28	0.28		0.02

Chapter 7

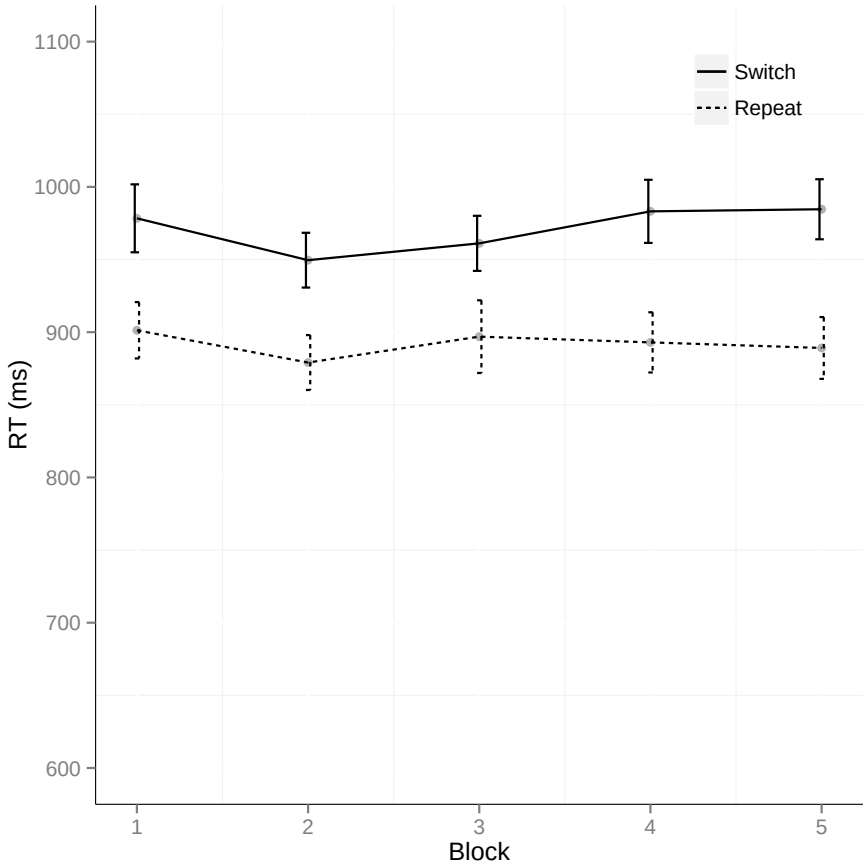

Figure A.27: Timecourse of reaction times in milliseconds across all five blocks in Experiment 7.1 in response to switch (solid) and repeat (dotted line) trials. Error bars show standard error.

Table A.44: Experiment 7.1: Results of a 2x5 (trial x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	2.11e+08	4.09e+06	1186.49	0.00	*	0.98
Block	2.53	58.10	59014.34	2.65e+06	0.51	0.64		0.02
Trial	1	23	468567.03	284170	37.92	0.00	*	0.62
Block:Trial	2.90	66.77	8315.41	759859.50	0.25	0.85		0.01

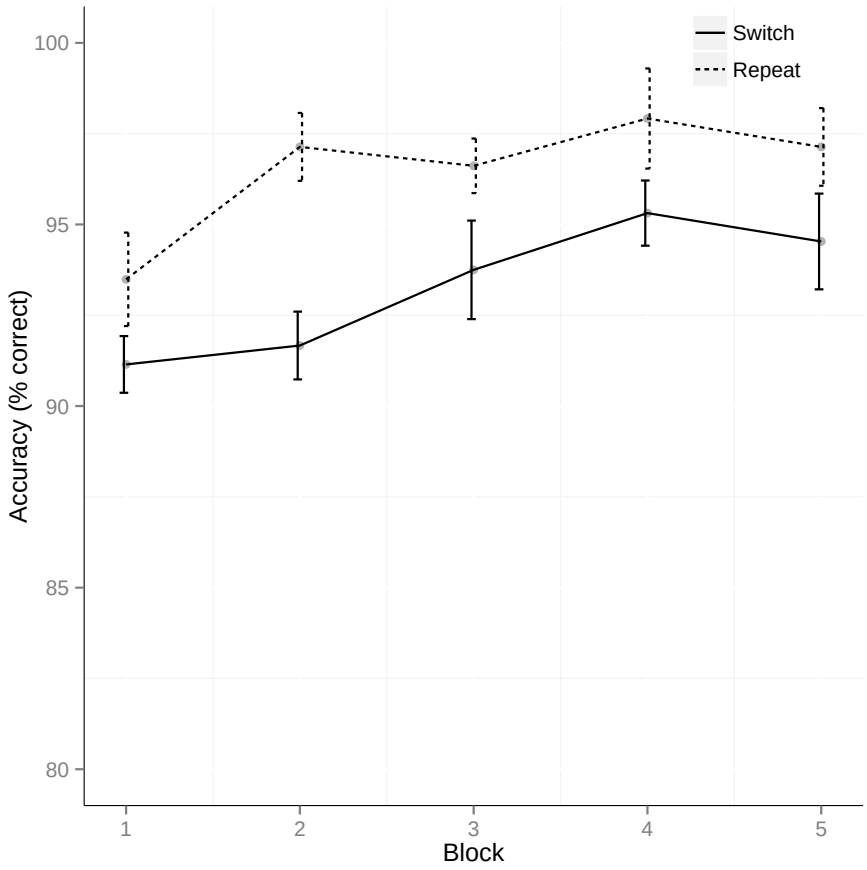

Figure A.28: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 7.1 in response to switch (solid) and repeat (dotted line) trials. Error bars show standard error.

Table A.45: Experiment 7.1: Results of a 2x5 (trial x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	2.16e+06	1648.27	30141.54	0.00	*	1.00
Block	4	92	518.55	3567.38	3.34	0.01	*	0.13
Trial	1	23	605.63	281.09	49.56	0.00	*	0.68
Block:Trial	4	92	80.40	2333.66	0.79	0.53		0.03

Table A.46: Experiment 7.1: Results of a 2x2 (time x trial) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	2512.54	25014.03	2.31	0.14		0.09
Time	1	23	77.27	2559.29	0.69	0.41		0.03
Trial	1	23	410.96	1645.65	5.74	0.03	*	0.20
Time:Trial	1	23	7.11	887.89	0.18	0.67		0.01

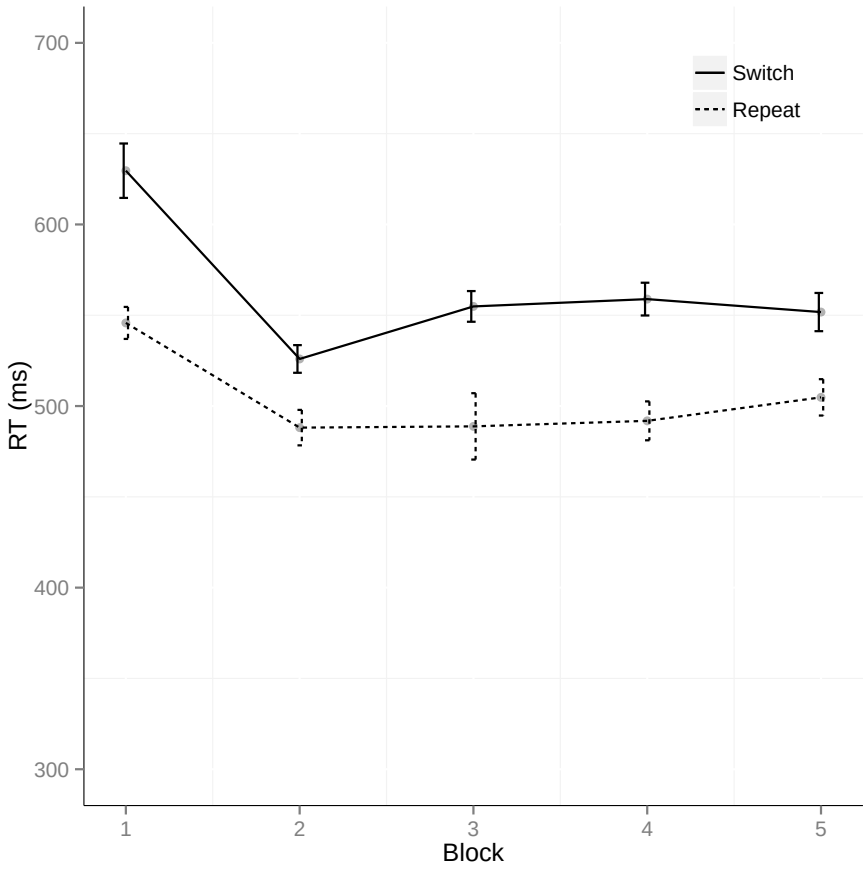

Figure A.29: Timecourse of reaction times in milliseconds across all five blocks in Experiment 7.2 in response to switch (solid) and repeat (dotted line) trials. Error bars show standard error.

Table A.47: Experiment 7.2: Results of a 2x5 (trial x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	6.87e+07	2.94e+06	536.74	0.00	*	0.96
Block	2.69	61.79	202119.71	822782.90	5.65	0.00	*	0.20
Trial	1	23	215415.87	183275	27.03	0.00	*	0.54
Block:Trial	4	92	16099.07	212855.70	1.74	0.15		0.07

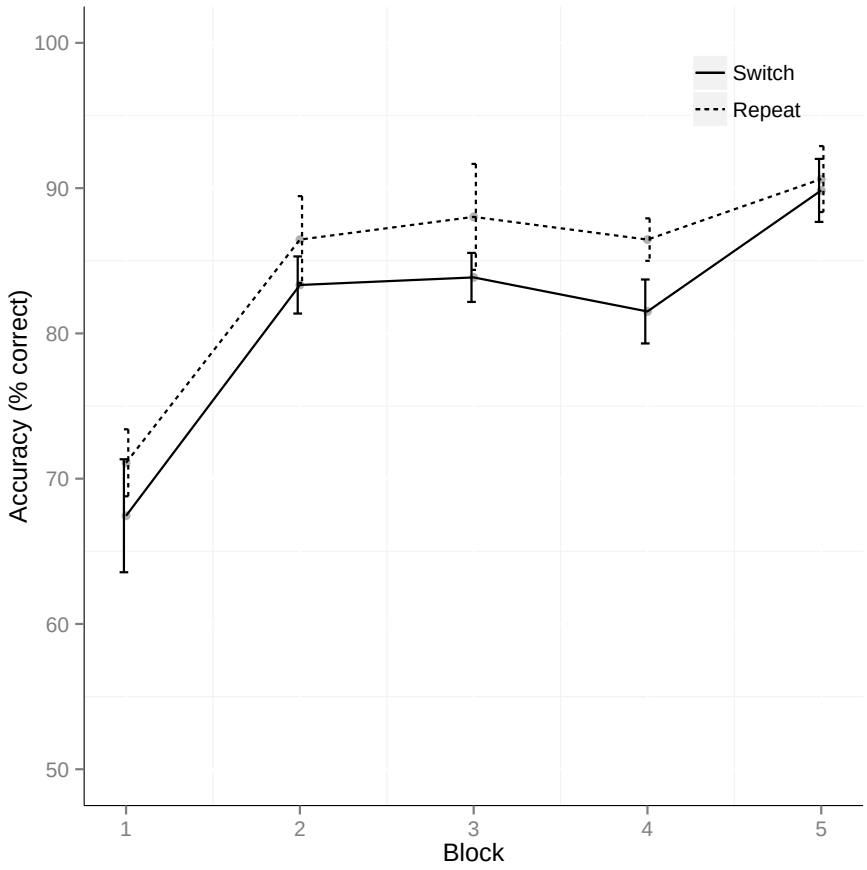

Figure A.30: Timecourse of accuracy rates as percentage correct across all five blocks in Experiment 7.2 in response to switch (solid) and repeat (dotted line) trials. Error bars show standard error.

Table A.48: Experiment 7.2: Results of a 2x4 (trial x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	1.65e+06	11382.16	3330.06	0.00	*	0.99
Block	2.58	59.39	12188.48	30061.52	9.33	0.00	*	0.29
Trial	1	23	666.67	872.40	17.58	0.00	*	0.43
Block:Trial	4	92	119.47	1974.28	1.39	0.24		0.06

Table A.49: Experiment 7.2: Results of a 2x2 (time x trial) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	1210.37	25720.25	1.08	0.31		0.04
Time	1	23	891.97	4510.76	4.55	0.04	*	0.17
Trial	1	23	33.40	1679.03	0.46	0.51		0.02
Time:Trial	1	23	26.96	471.36	1.32	0.26		0.05

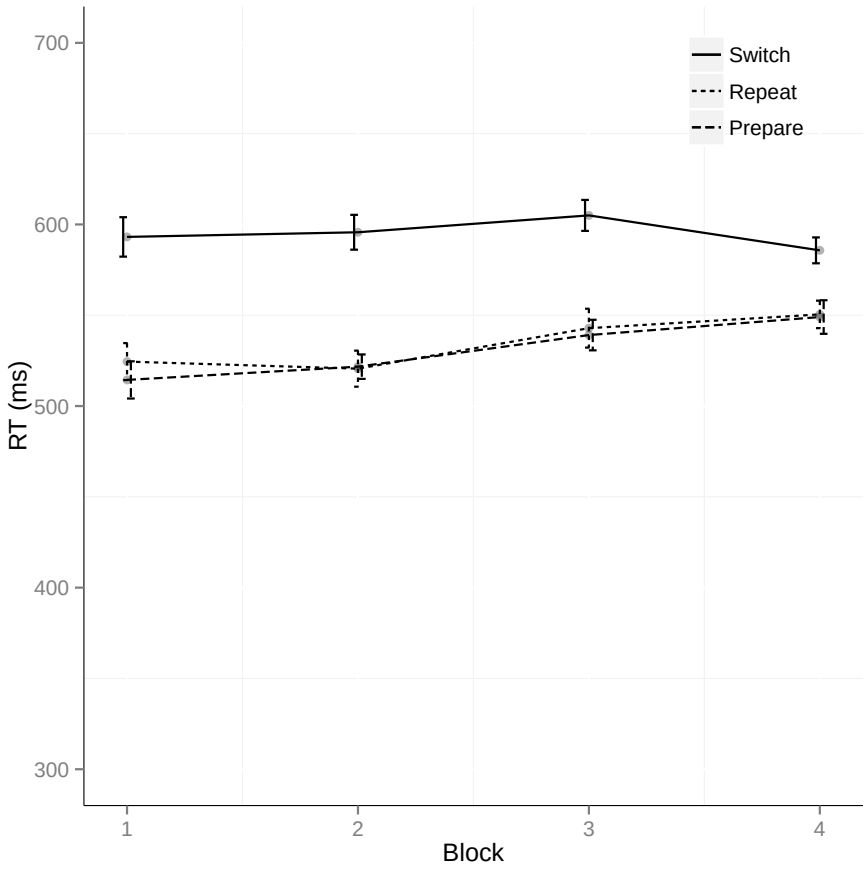

Figure A.31: Timecourse of reaction times in milliseconds across all four blocks in Experiment 7.3 in response to switch (solid), repeat (dotted), and prepare (dashed line) trials. Error bars show standard error.

Table A.50: Experiment 7.3: Results of a 3x4 (trial x block) factorial ANOVA on reaction times

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	8.82e+07	2.05e+06	989.37	0.00	*	0.98
Block	2.10	48.39	21048.52	767548.42	0.63	0.54		0.03
Trial	1.26	28.90	247142.71	175776.35	32.34	0.00	*	0.58
Block:Trial	4.34	99.78	16427.32	143166.86	2.64	0.03	*	0.10

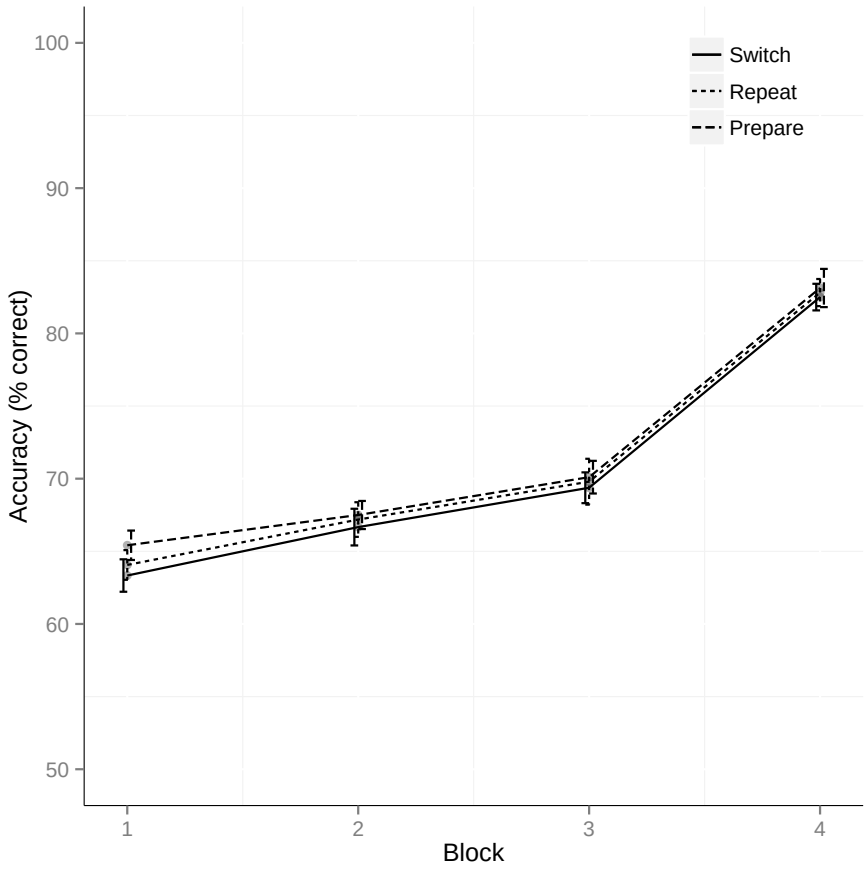

Figure A.32: Timecourse of accuracy rates as percentage correct across all four blocks in Experiment 7.3 in response to switch (solid), repeat (dotted), and prepare (dashed line) trials. Error bars show standard error.

Table A.51: Experiment 7.3: Results of a 3x4 (trial x block) factorial ANOVA on accuracy rates

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	23	1.45e+06	8121.09	4110.50	0.00	*	0.99
Block	2.22	50.98	14503.04	6402.17	52.10	0.00	*	0.69
Trial	2	46	54.82	107.68	11.71	0.00	*	0.34
Block:Trial	4.16	95.79	18.45	348.22	1.22	0.31		0.05

Table A.52: Experiment 7.3: Results of a 2x3 (time x trial) factorial ANOVA on trustworthiness ratings

Effect	DFn	DFd	SSn	SSd	F	p	*	η_P^2
(Intercept)	1	26	619.32	56295.17	0.29	0.60		0.01
Time	1	26	952.18	5338.69	4.64	0.04	*	0.15
Trial	2	52	602.69	14410.54	1.09	0.34		0.04
Time:Trial	2	52	775.26	6905.47	2.92	0.06		0.10

B Interference task used in Experiments 3.2 and 3.3

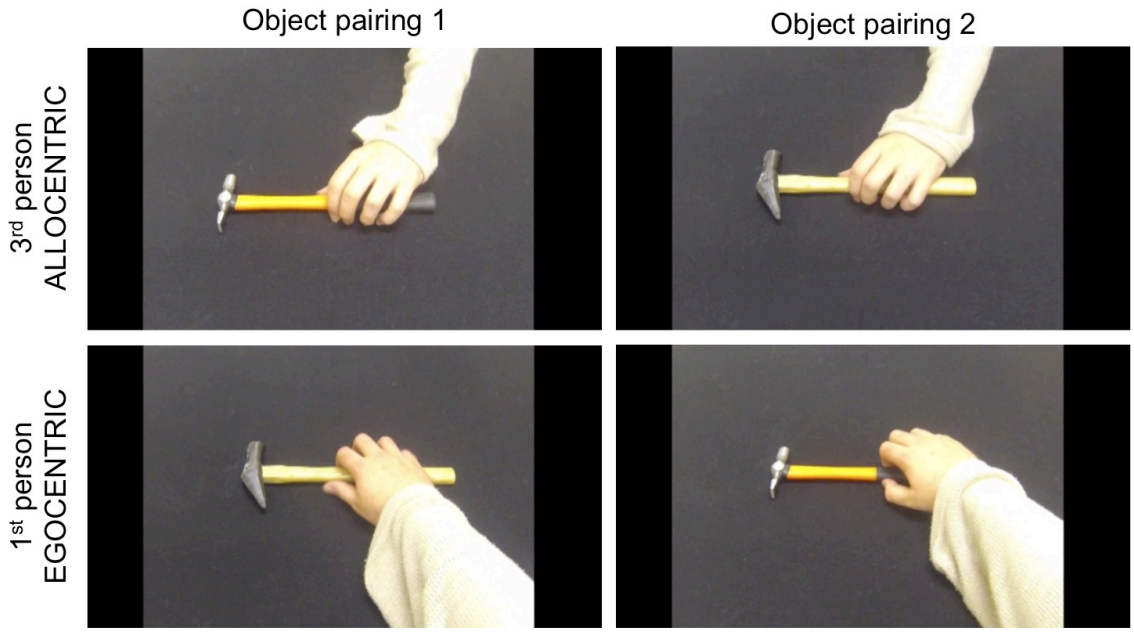

Figure A.33: Examples of object-directed actions in the allocentric (third person; top row) and egocentric (first person; bottom row) perspectives. Each participant would only see either object pairing 1 (with object 1 in the allocentric perspective and object 2 in the egocentric) or object pairing 2 (vice versa), although the order of which perspective orientation was presented first was counterbalanced across objects and participants.

Stimuli were videos of pairs of everyday household objects - see Appendix C for full list. Each pair of objects was broadly matched to appear physically similar to the other object as possible while still remaining visually distinct enough to differentiate between the two. Videos showed the object sitting on a plain black surface before a hand appeared from either the top (allocentric perspective, third person) or bottom (egocentric, first person) of the screen, picked it up and brought it back off screen and out of sight. See Figure A.33 for screen caps of the four different video conditions for one object pair.

Videos were captured using a GoPro Hero 3 Black Edition video camera mounted on a tripod and capturing images with a resolution of 1080 and recording speed of 25fps. The camera occupied the same position for all videos regardless of action orientation. Each object was filmed once from the ego- and allocentric perspectives, resulting in 48

videos in total. Each video was cropped and edited using Adobe Premiere Elements 12 such that each video began 3 seconds before the hand first appeared, and ended 3 seconds after the object had disappeared from view. They were also visually cropped such that the table was all that could be seen and each video was visually the same size. All audio recordings were removed from the recordings as well.

Due to a black border around the edges of the videos, these were presented on a black background, a clearly distinct visual experience from the white-backed surrounding tasks. Participants were told that they would see two objects that would be picked up and moved off screen, and their task was simply to judge which of the two they preferred. Before each video a white fixation cross appeared in the centre of the screen, and then the video would play.

Participant response conditions

IIn the first condition, participants saw images of the objects (the first frame of the video clips) presented side by side alongside key pairings. In this condition, the keys were the numbers 1 and 2, which are located next to each other on a normal QWERTY keyboard, meaning that participants' responses did not differ in terms of gross action mapping between the two conditions. This is the Low action mapping condition.

In the second condition, participants saw the images side by side, but instead of pressing 1 and 2, they pressed Z and M to identify whether their preferred image was presented on the left or right side of the screen. As these keys are located at opposite ends of a QWERTY keyboard, these two responses would be dissociable in terms of response actions, as they would be mapped onto different sides of space. As such this is the Medium action mapping condition.

In the third condition, participants saw the images side by side, but instead of using the keyboard they moved the mouse to the image that they preferred. In each trial, participants had to click on a START location at the bottom of the screen, and then the images would appear in the top two corners, meaning that participants had to simulate a reaching motion with the mouse in order to click on the preferred object. The mouse speed was slowed to encourage more motor movement, and after each trial participants returned the mouse to the bottom of the screen to click on a NEXT button in order to advance. This is the Moderate action mapping condition.

The final condition had no keyboard responses. Instead, participants were presented with the images and they were instructed to point to their preferred image. The image (left or right) was then registered by the experimenter and coded later. Participants pointed at the object, and this action closely (although not exactly) mirrored the grasping action in broad motor representations. This is the High action mapping condition.

Results and Discussion

Filler tasks

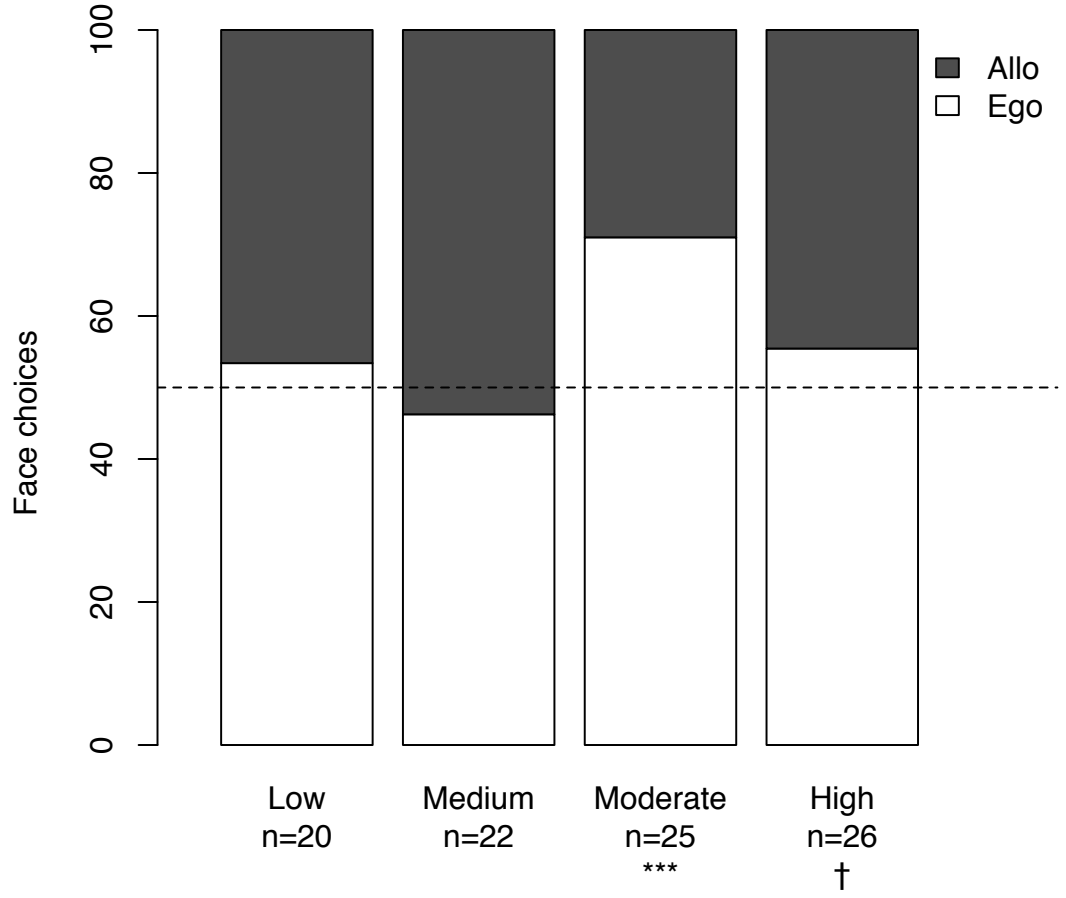

Figure A.34: Stacked bar chart to show the average proportions of objects chosen as preferred when grasped from an egocentric orientation (white) and those chosen when grasped from an allocentric orientation (grey). The three bars show the results for the four different types of condition. Subject *n*s are shown beneath the x-axis labels. Dashed line shows point of equal preference for ego- and allo- videos (50%). ***p<1; †p<.10

The results of the four experiments are shown in Figure A.34. In the High condition participants chose egocentrically grasped objects more than allocentrically grasped, but a binomial test found this only approached conventional levels of significance ($p = 0.062$). Participants chose the object grasped from the egocentric perspective on only 55.45% of

trials.

There did not appear to be any bias in the Low ($p = 0.272$), or Medium ($p = 0.295$) conditions. However, there was a significant bias towards selecting egocentrically grasped images over allocentrically grasped images in the Moderate condition − where participants responded using response keys that mapped onto the preferred image's location in space (Z and M for left and right, respectively; $p < .001$). However, that this bias emerges in this condition but only marginally in the High action mapping condition (where participants have to actively reach out and point to their preferred object) is a curious point, and one that might undermine our confidence in this finding.

C List of objects used in filler task in Experiments 3.2 and 3.3

- Coloured wooden block
- Cafetière
- Corkscrew
- Glass
- Hammer
- Computer mouse
- Mug
- Hole punch
- Scissors
- Screwdriver
- Tape measure
- Torch

D Images used in Experiment 4.3

Arrows

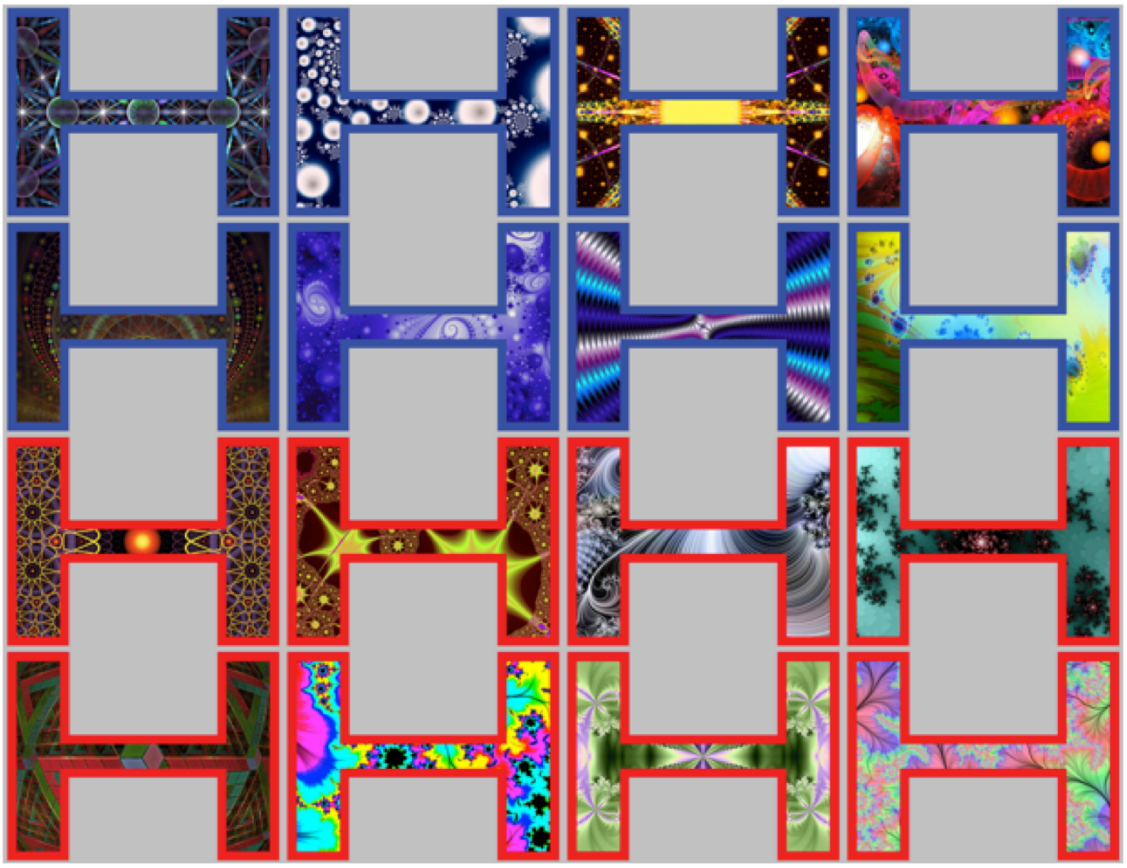

Figure A.35: 16 images used as 'arrow' stimuli in Experiment 4.3. These were described as decorated images of the letter H during the experiment.

Fractals

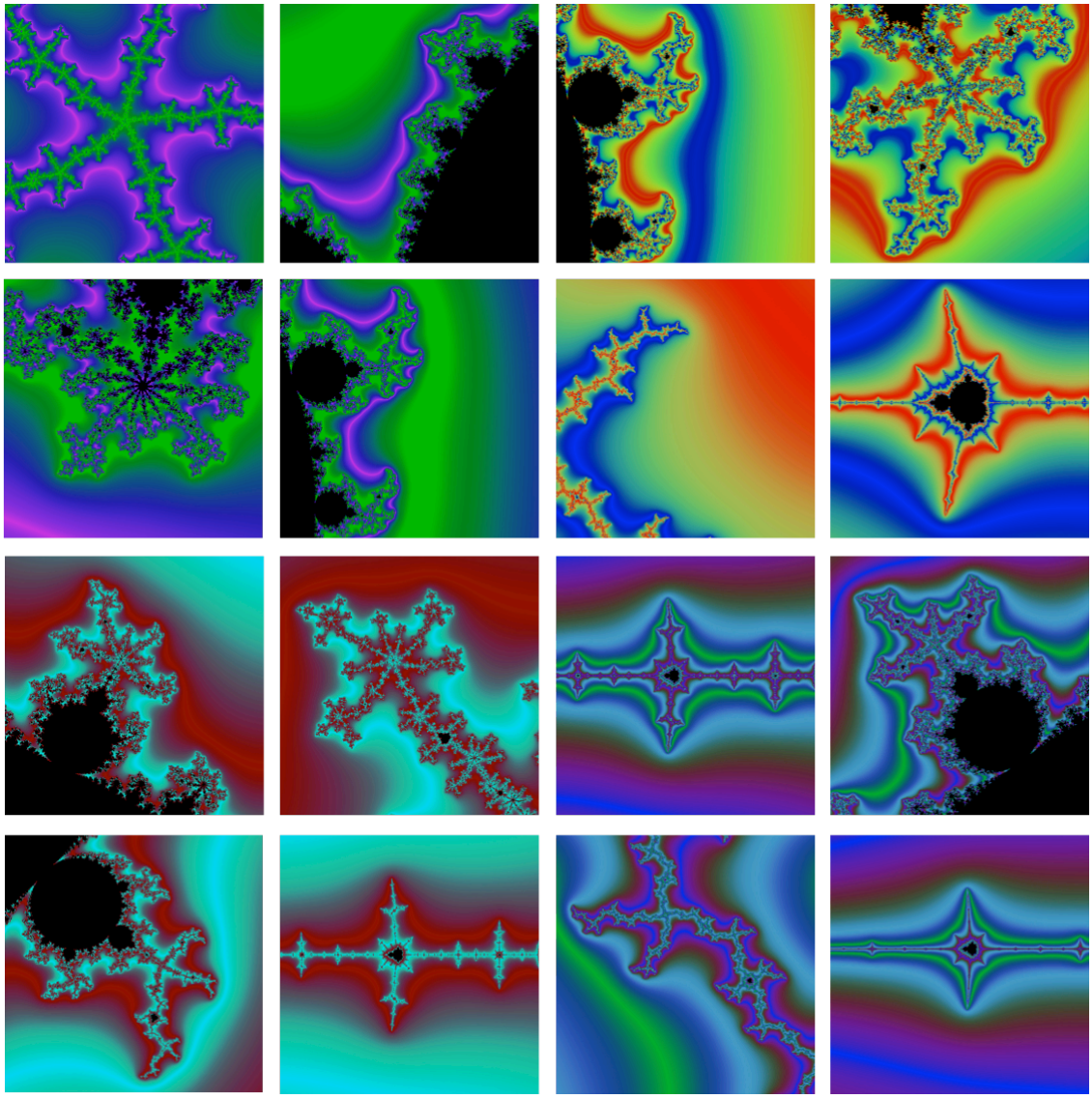

Figure A.36: 16 images of Mandelbrot fractal images used as stimuli in Experiment 4.3.

Kandinsky

Figure A.37: 16 images of Processing-generated Kandinsky-style artwork used as stimuli in Experiment 4.3.

E Results of EMG recording in Experiment 6.1

Experiment 6.1b was a replication and extension of Experiment 6.1a where in addition to behavioural analysis we record electromyographic (EMG) activity. EMG is an electrophysiological technique that measures electrical signals in muscles when they contract. Facial EMG can be used to measure small contractions in muscles associated with particular emotional expressions such as the corrugator supercilii (which runs across the brow and is associated with frowning) and the zygomaticus major (which runs from the corner of the mouth up the cheek towards the temple and is associated with smiling). Studies have associated activity in these muscles with emotional responses to stimuli − both muscles show heightened activity during observations of particular emotional expressions (corrugator when frowning, zygomaticus when smiling; Dimberg, Thunberg & Elmehed, 2000), but this emotional mimicry is influenced by the social context of incoming visual information such that mimicry of smiles is inhibited when they are perceived to be gloating (Kirkham, Hayes & Tipper, 2015). This points to facial EMG signals as a measurement not of automatic motor imitation but of participants embodied emotional states as they experience them.

To date, only one study has investigated EMG activity in incidentally learned trust from gaze cues. Manssuer, Pawling et al. (2015) recorded from the corrugator and zygomaticus, and found that signals in the corrugator increased when a participant experienced invalid gaze cues. This points to an embodied emotional response to deception, as the contraction of this muscle is an action coding unit for anger in the Facial Action Coding System (FACS; Ekman & Friesen, 1978). Crucially, this embodied response only emerged in participants who subsequently rated invalid faces as less trustworthy than valid faces − in a selection of those participants who did not show this

learning, there was no embodied emotional response.

In Experiment 6.1b, we record from these same facial muscles (corrugator and zygomaticus) and measure activity in response to all four stimulus conditions (White Valid; White Invalid; Chinese Valid; and Chinese Invalid), to see not only if invalid cues lead to greater corrugator response, but also to investigate whether this response is moderated by race. If the prediction regarding expectancy violation − that learning will be stronger for events that are more surprising − it may be that this is reflected in the EMG signal. Alternatively, the use of this electrophysiological technique may find such a coding that behavioural measures are not sensitive to.

EMG Parameters and Preprocessing

Two pairs of 4mm Ag/AgCl electrodes filled with conductive electrolyte gel were secured upon the left-hand side of the face of each participant using adhesive discs. The electrodes were sited according to the guidance of van Boxtel (2010). A ground electrode was also placed upon the forehead. Prior to the application of the electrodes, each site was prepared by cleaning and exfoliating the skin, before wiping with an alcohol swab. EMG activity was obtained at 2000Hz using a combination of BioPac systems (MP150 and EMG100C), and recorded using AcqKnowledge software.

Following the completion of each recording, the raw signal from each muscle was filtered using a bandpass filter (20Hz - 500Hz) and a notch filter of 50Hz, before being rectified and smoothed with an integration window of 50ms (100 samples).

EMG amplitude was then normalised across all trials around the average of the activity during the final 500ms of the fixation screen, and then binned into 250ms windows, and these are used to show signal timecourses. For the purposes of analysis,

these data were then averaged across all bins and trials for each epoch and compared across participants.

Corrugator Supercilii

Results of EMG recording from the corrugator are broken down by different periods of the experiment. For the analysis of the gaze-cueing portion of the experiment, trials were broken down into separate events: trial event 1 was where the face appeared showing direct gaze for the first time for 1,500ms; trial event 2 was where the face shifted its gaze for 500ms; trial event 3 was where the target appeared and participants made their response, which lasted 2,500ms; and trial event 4 was where the face returned to direct gaze for a further 1,000ms. These are examined separately and are shown in Figure A.38.

Trial event 1: Direct gaze

Adding validity to the null model did not explain significantly more of the variance than did the null model (β = 0.00, SE = 0.00, $\chi^2(1) = 0.55$, $p = 0.458$), nor did including race (β = -0.00, SE = 0.00, $\chi^2(1) = 0.77$, $p = 0.380$), and comparison of the two-fixed-factor models found that an interaction term fit the data only marginally better than when the two factors were modelled without an interaction (β = -0.00, SE = 0.00, $\chi^2(1) = 3.07$, $p = 0.080$).

Trial events 2 and 3: Target appears

As trial event 2, where the face shifted its gaze, was the shortest of all trial events and only yielded two data points per participant per condition (due to the data being binned into 250ms chunks), trial events 2 and 3 (where the target appeared) were collapsed together for analysis. Adding validity to the null model did not explain significantly

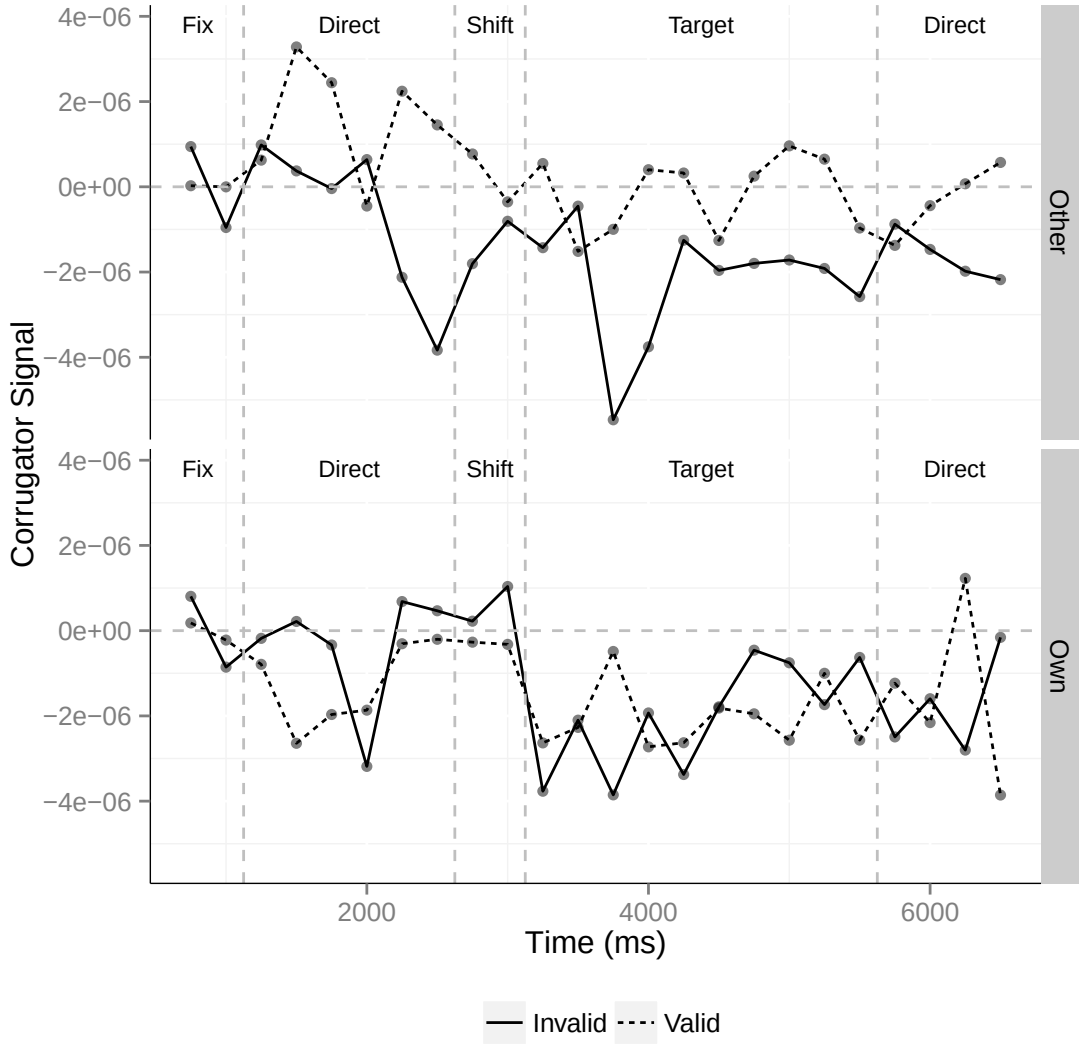

Figure A.38: Timecourse of EMG signal recording in Experiment 6.1 in the corrugator supercilii in response to valid (dotted) and invalid (solid line) British (top plot) and Chinese faces (bottom plot) during gaze cueing.

more of the variance than did the null model (β = 0.00, SE = 0.00, $\chi^2(1) = 0.58$, p = 0.445), nor did including race (β = -0.00, SE = 0.00, $\chi^2(1) = 0.00$, $p = 0.953$), and there was no evidence of any interaction between these two factors (β = -0.00, SE = 0.00, $\chi^2(1) = 2.16$, $p = 0.141$).

Trial event 4: Return to direct gaze

For trial event 4, where the face returned to direct gaze after the participants' response had been made, adding validity to the null model did not explain significantly more of

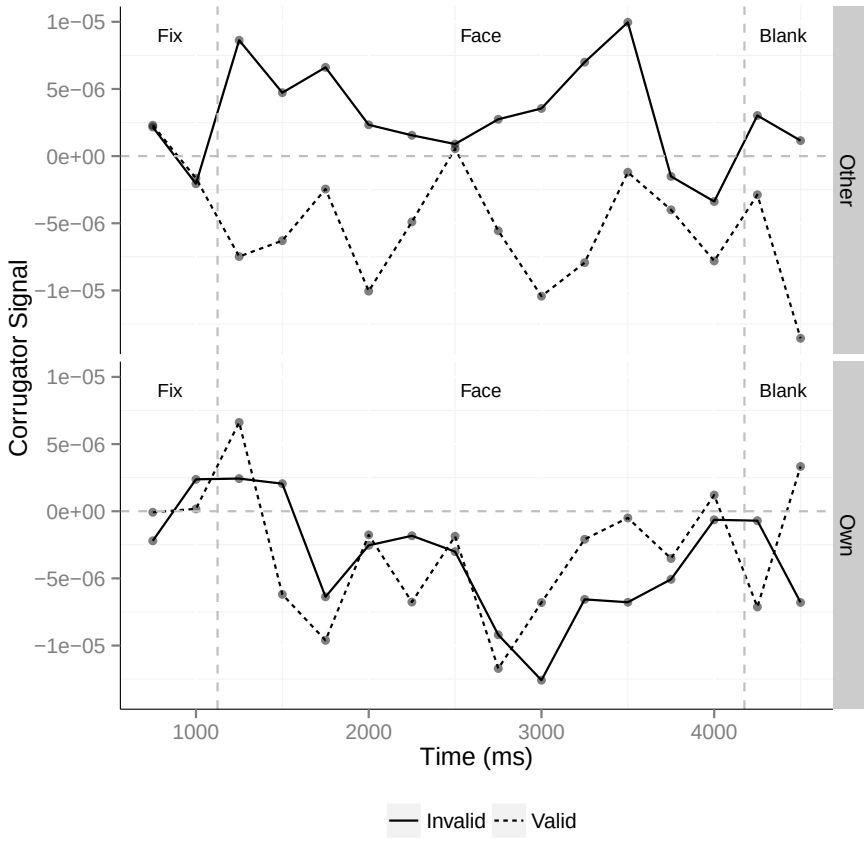

Figure A.39: Timecourse of EMG signal recording in Experiment 6.1 in the corrugator supercilii in response to valid (dotted) and invalid (solid line) British (top plot) and Chinese faces (bottom plot) during pre-experiment trustworthiness ratings.

the variance than did the null model ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 0.27$, $p = 0.600$), nor did including race ($\beta =$ -0.00, $SE = 0.00$, $\chi^2(1) = 0.06$, $p = 0.803$), and there was no evidence of any interaction between these two factors ($\beta =$ -0.00, $SE = 0.00$, $\chi^2(1) =$ 0.19, $p = 0.664$).

Trustworthiness ratings

For analysis of the trustworthiness ratings, EMG signal to all four conditions of faces in the pre-experiment rating (see Figure A.39) and post-experiment rating (see Figure A.40) were analysed together and time included in the modelling as a fixed factor.

Adding time to the null model did not explain significantly more of the variance than did the null model ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 0.27$, $p = 0.605$), nor did

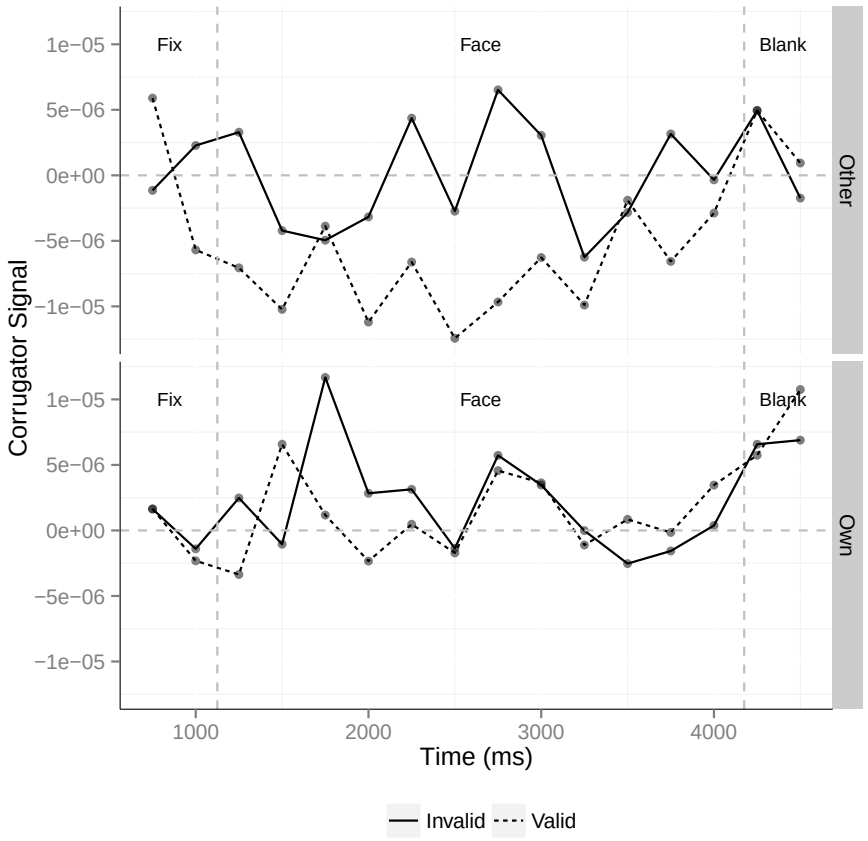

Figure A.40: Timecourse of EMG signal recording in Experiment 6.1 in the corrugator supercilii in response to valid (dotted) and invalid (solid line) British (top plot) and Chinese faces (bottom plot) during post-experiment trustworthiness ratings.

including race (β = 0.00, SE = 0.00, $\chi^2(1) = 0.15$, $p = 0.697$), but adding validity did marginally improve the fit (β = -0.00, SE = 0.00, $\chi^2(1) = 2.93$, $p = 0.087$). However, comparing time and validity in both a two-factor and interaction model found no evidence of an interaction (β = 0.00, SE = 0.00, $\chi^2(1) = 0.00$, $p = 0.946$), and nor did a three-factor comparison of models with time, validity and race (β = -0, SE = 0.00, $\chi^2(4) = 5.89$, $p = 0.207$).

Zygomaticus Major

Results of EMG recordings from the zygomaticus major are examined separately and shown in Figure A.41.

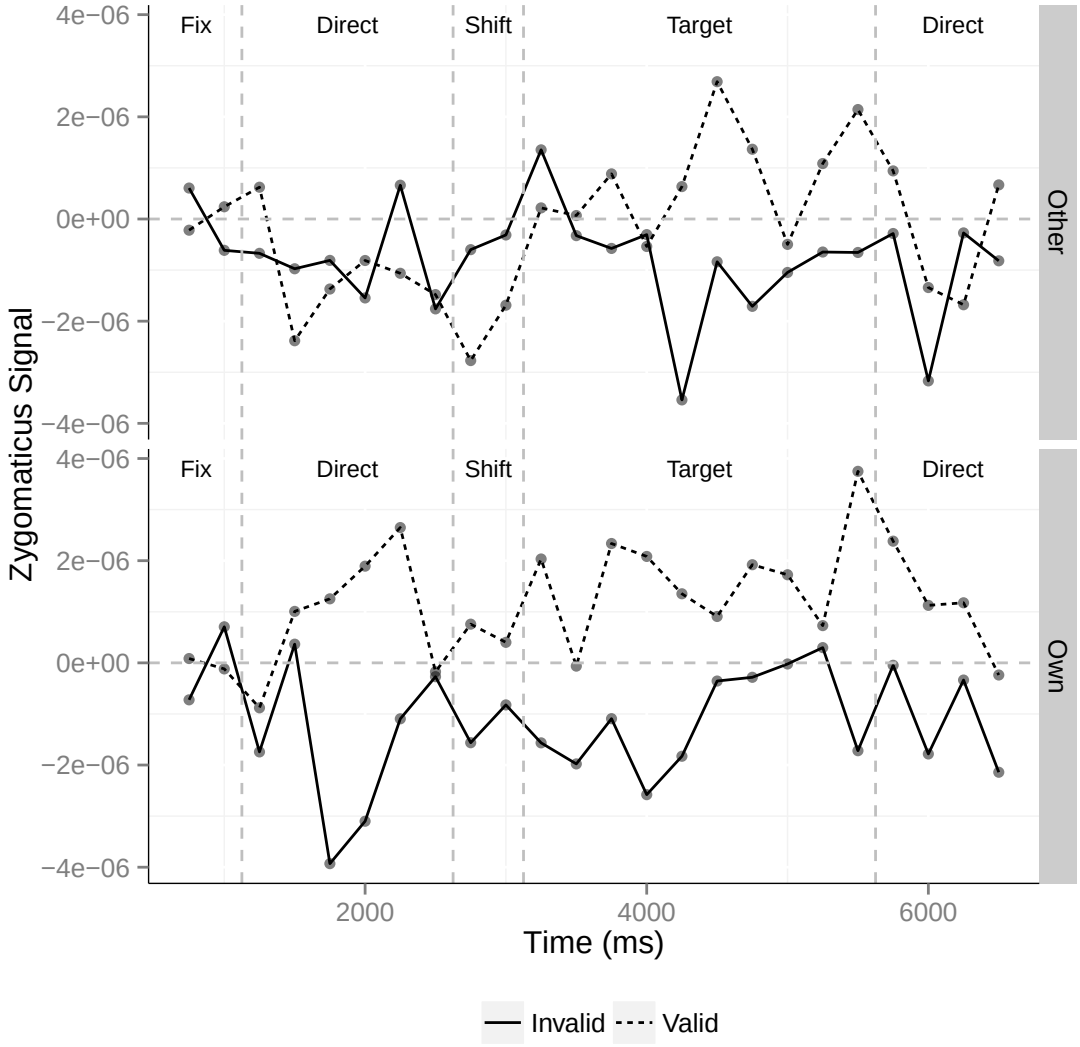

Figure A.41: Timecourse of EMG signal recording in Experiment 6.1 in the zygomaticus major in response to valid (dotted) and invalid (solid line) British (top plot) and Chinese faces (bottom plot) during gaze cueing.

Trial events 1 and 2: Direct gaze

Adding validity to the null model did not explain significantly more of the variance than did the null model ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 1.70$, $p = 0.192$), nor did including race ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 0.25$, $p = 0.618$), and there was no evidence of any interaction between these two factors ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 2.62$, $p = 0.105$).

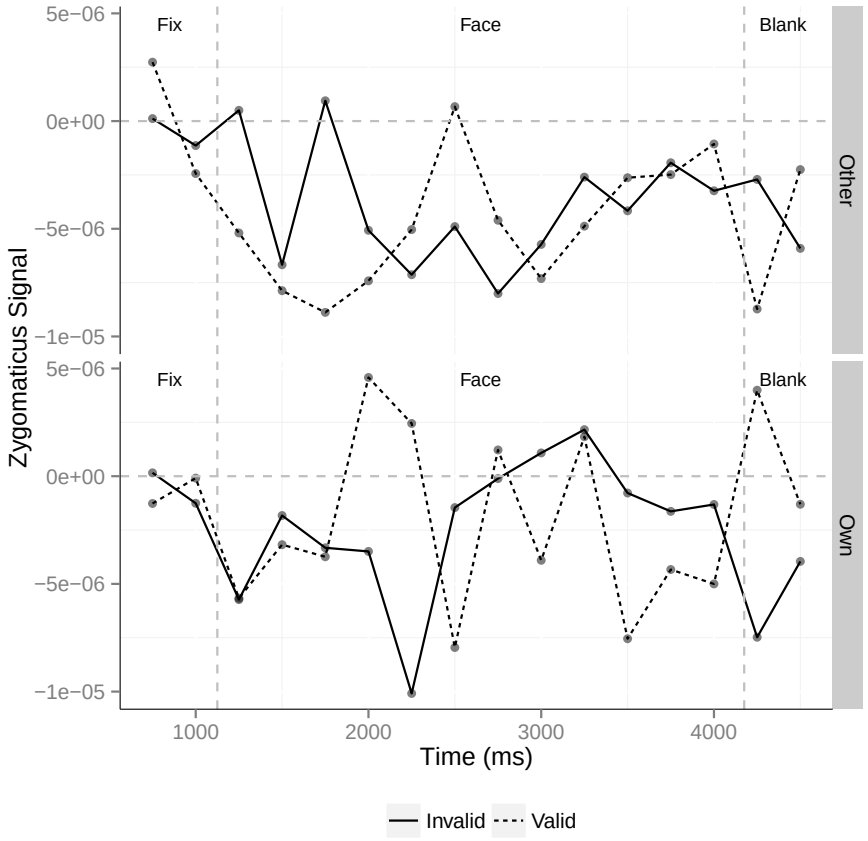

Figure A.42: Timecourse of EMG signal recording in Experiment 6.1 in the zygomaticus major in response to valid (dotted) and invalid (solid line) British (top plot) and Chinese faces (bottom plot) during pre-experiment trustworthiness ratings.

Trial events 2 and 3: Target appears

For trial events 2 and 3, where the face shifted its gaze and the target object appeared in either the cued or uncued location, adding validity to the null model did not explain significantly more of the variance than did the null model ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 1.23$, $p = 0.268$), nor did including race ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 0.40$, $p = 0.529$), and there was no evidence of any interaction between these two factors ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 2.04$, $p = 0.153$).

Trial event 4: Return to direct gaze

For trial event 4, where the face returned to direct gaze after the participants' response had been made, adding validity to the null model did not explain significantly more of

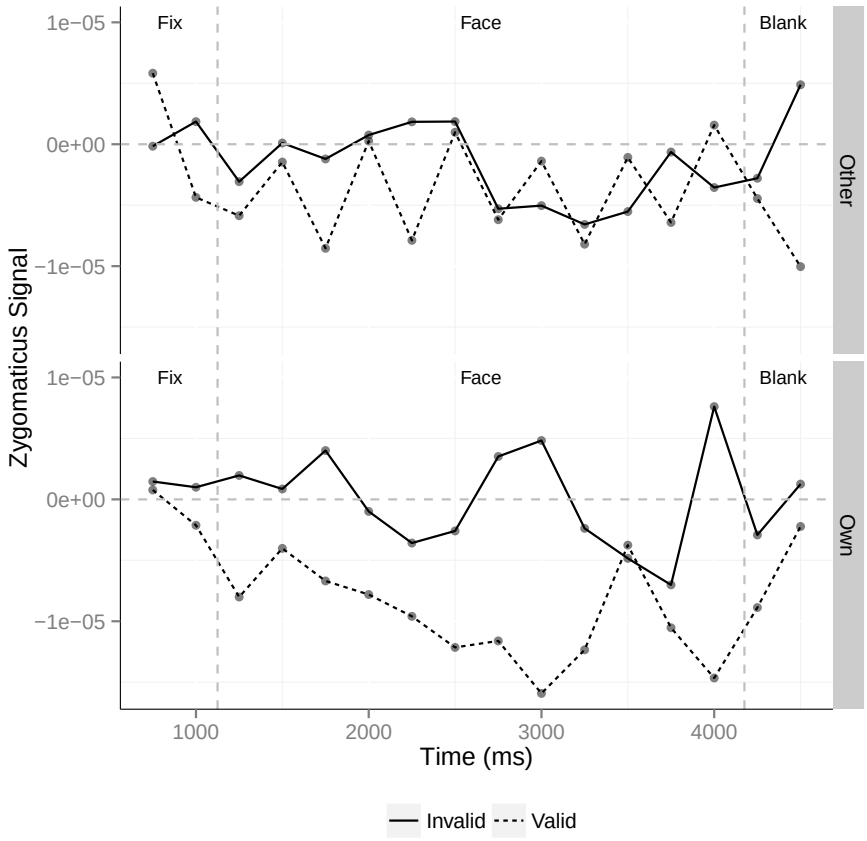

Figure A.43: Timecourse of EMG signal recording in Experiment 6.1 in the zygomaticus major in response to valid (dotted) and invalid (solid line) British (top plot) and Chinese faces (bottom plot) during post-experiment trustworthiness ratings.

the variance than did the null model ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 1.99$, $p = 0.159$), nor did including race ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 0.33$, $p = 0.564$), and there was no evidence of any interaction between these two factors ($\beta = 0.00$, $SE = 0.00$, $\chi^2(1) = 0.47$, $p = 0.494$).

Trustworthiness ratings

For analysis of the trustworthiness ratings, EMG signal to all four conditions of faces in the pre-experiment rating (see Figure A.42) and post-experiment rating (see Figure A.43) were analysed together and time included in the modelling as a fixed factor.

Adding time to the null model explained significantly more of the variance than did the null model ($\beta = -0$, $SE = 0.00$, $\chi^2(1) = 0.05$, $p = 0.824$), but including race did not

(β = 0.00, SE = 0.00, $\chi^2(1)$ = 0.00, p = 0.948), and nor did including validity (β = -0, SE = 0.00, $\chi^2(1)$ = 1.04, p = 0.309). Similarly, comparing time and validity in both a two-factor and interaction model found no evidence of an interaction (β = -0, SE = 0.00, $\chi^2(1)$ = 1.44, p = 0.230), and nor did a three-factor comparison of models with time, validity and race (β = -0, SE = 0.00, $\chi^2(4)$ = 3.63, p = 0.459).

References

Adams, R. B., Pauker, K. & Weisbuch, M. (2010). Looking the other way: The role of gaze direction in the cross-race memory effect. *Journal of Experimental Social Psychology*, *46*(2), 478–481.

Allen, V. L. & Wilder, D. A. (1975). Categorization, belief similarity, and intergroup discrimination. *Journal of Personality and Social Psychology*, *32*(6), 971–977.

Andrews, S., Jenkins, R., Cursiter, H. & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*(September), 1–10.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baron-Cohen, S. (2004). The cognitive neuroscience of autism. *Journal of neurology, neurosurgery, and psychiatry*, *75*(7), 945–948.

Bayliss, A. P., Frischen, A., Fenske, M. J. & Tipper, S. P. (2007). Affective evaluations of objects are influenced by observed gaze direction and emotional expression. *Cognition*, *104*(3), 644–653.

Bayliss, A. P., Griffiths, D. & Tipper, S. P. (2009). Predictive gaze cues affect face evaluations: The effect of facial emotion. *European Journal of Cognitive Psychology*, *21*(7), 1072–1084.

Bayliss, A. P., Paul, M. a., Cannon, P. R. & Tipper, S. P. (2006). Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic bulletin & review*, *13*(6), 1061–1066.

Bayliss, A. P. & Tipper, S. P. (2006). Predictive gaze cues and personality judgments: Should eye trust you? *Psychological Science*, *17*(6), 514–520.

Beaupré, M. G. & Hess, U. (2006). An ingroup advantage for confidence in emotion recognition judgments: The moderating effect of familiarity with the expressions of outgroup members. *Personality & social psychology bulletin*, *32*(1), 16–26.

Bell, R., Buchner, A., Erdfelder, E., Giang, T., Schain, C. & Riether, N. (2012). How specific is source memory for faces of cheaters? Evidence for categorical emotional tagging. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 457–472.

Bell, R., Buchner, A., Kroneisen, M. & Giang, T. (2012). On the flexibility of social source memory: A test of the emotional incongruity hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1512–1529.

Bell, R., Buchner, A. & Musch, J. (2010). Enhanced old-new recognition and source memory for faces of cooperators and defectors in a social-dilemma game. *Cognition*, *117*(3), 261–275.

Birmingham, E., Bischof, W. F. & Kingstone, A. (2009). Get real! Resolving the debate about equivalent social stimuli. *Visual Cognition*, *17*(6-7), 904–924.

Bruce, V. & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(Pt 3), 305–327.

Buchner, A., Bell, R., Mehl, B. & Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evolution and Human Behavior*, *30*(3), 212–224.

Burgess, T. D. G. & Sales, S. M. (1971). Attitudinal effects of "mere exposure": A reevaluation. *Journal of Experimental Social Psychology*, *7*(4), 461–472.

Carré, J. M. & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1651), 2651–2656.

Caulfield, F., Ewing, L., Burton, N., Avard, E. & Rhodes, G. (2014). Facial trustworthiness judgments in children with ASD are modulated by happy and angry emotional cues. *PLoS ONE*, *9*(5), e97644.

Chen, Y. C. & Yeh, S. L. (2012). Look into my eyes and I will see you: Unconscious processing of human gaze. *Consciousness and Cognition*, *21*(4), 1703–1710.

Chow, V., Poulin-Dubois, D. & Lewis, J. (2008). To see or not to see: Infants prefer to follow the gaze of a reliable looker: PAPER. *Developmental Science*, *11*(5), 761–770.

Constable, M. D., Bayliss, A. P., Tipper, S. P. & Kritikos, A. (2013). Self-generated cognitive fluency as an alternative route to preference formation. *Consciousness and Cognition*, *22*(1), 47–52.

Dasgupta, N., Mcghee, D. E., Greenwald, A. G. & Banaji, M. R. (2000). Automatic Preference for White Americans: Eliminating the Familiarity Explanation. *Journal of Experimental Social Psychology*, *328*(3), 316–328.

Destrebecqz, A. & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic bulletin & review*, *8*(2), 343–50.

Dimberg, U., Thunberg, M. & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological science : a journal of the American Psychological Society / APS*, *11*(1), 86–89.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E. & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, *6*(5), 509–540.

Dunsmoor, J. E., Kubota, J. T., Li, J., Coelho, C. A. O. & Phelps, E. A. (2016). Racial stereotypes impair flexibility of emotional learning. *Social cognitive and affective neuroscience*, *11*(9), 1363–1373.

Efferson, C. & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, *3*, 1047.

Eger, E., Schweinberger, S. R., Dolan, R. J. & Henson, R. N. (2005). Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *NeuroImage*, *26*(4), 1128–1139.

Ekman, P. & Friesen, W. V. (1978). *Manual for the facial action coding system.* Consulting Psychologists Press.

Elfenbein, H. A. & Ambady, N. (2002a). Is there an in-group advantage in emotion recognition? *Psychological Bulletin*, *128*(2), 243–249.

Elfenbein, H. A. & Ambady, N. (2002b). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, *128*(2), 203–235.

Engell, A. D., Todorov, A. & Haxby, J. V. (2010). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception*, *39*(7), 931–941.

Falvello, V., Vinson, M., Ferrari, C. & Todorov, A. (2015). The Robustness of Learning about the Trustworthiness of Other People. *Social Cognition*, *33*(5), 368–386.

Field, A. (2007). *Discovering Statistics Using SPSS* (3rd). SAGE Publications.

Firestone, C. & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences*, *4629*, 1–72.

Fiske, S. T., Cuddy, A. J. C. & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Freeth, M., Chapman, P., Ropar, D. & Mitchell, P. (2010). Do gaze cues in complex scenes capture and direct the attention of high functioning adolescents with asd? evidence from eye-tracking. *Journal of Autism and Developmental Disorders*, *40*(5), 534–547.

Freeth, M., Ropar, D., Chapman, P. & Mitchell, P. (2010). The eye gaze direction of an observed person can bias perception, memory, and attention in adolescents with and without autism spectrum disorder. *Journal of Experimental Child Psychology*, *105*(1-2), 20–37.

Friesen, C. K. & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490–495.

Frischen, A. & Tipper, S. P. (2004). Orienting attention via observed gaze shift evokes longer term inhibitory effects: implications for social interactions, attention, and memory. *Journal of experimental psychology. General*, *133*(4), 516–33.

Gómez-Valdés, J., Hünemeier, T., Quinto-Sánchez, M., Paschetta, C., de Azevedo, S., González, M. F., ... González-José, R. (2013). Lack of Support for the Association between Facial Shape and Aggression: A Reappraisal Based on a Worldwide Population Genetics Perspective. *PLoS ONE*, *8*(1), e52317.

Gordon, I., Eilbott, J. A., Feldman, R., Pelphrey, K. A. & Vander Wyk, B. C. (2013). Social, reward, and attention brain networks are involved when online bids for joint attention are met with congruent versus incongruent responses. *Social neuroscience*, *8*(6), 544–554.

Haselhuhn, M. P., Kennedy, J. A., Kray, L. J., Van Zant, A. B. & Schweitzer, M. E. (2015). Gender differences in trust dynamics: Women trust more than men following a trust violation. *Journal of Experimental Social Psychology*, *56*, 104–109.

Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233.

Hayes, A. E., Paul, M. A., Beuger, B. & Tipper, S. P. (2008). Self produced and observed actions influence emotion: The roles of action fluency and eye gaze. *Psychological Research*, *72*(4), 461–472.

Hehman, E., Flake, J. K. & Freeman, J. B. (2015). Static and Dynamic Facial Cues Differentially Affect the Consistency of Social Evaluations. *Personality and Social Psychology Bulletin*, *41*(8), 1–12.

Heuer, H., Schmidtke, V. & Kleinsorge, T. (2001). Implicit learning of sequences of tasks. *Journal of experimental psychology. Learning, memory, and cognition*, *27*(4), 967–983.

Hood, B. M., Willen, J. & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological science*, *9*(2), 131–134.

Hu, C. S., Wang, Q., Han, T., Weare, E. & Fu, G. (2015). Differential emotion attribution to neutral faces of own and other races. *Cognition and Emotion*, 1–9.

Hungr, C. J. & Hunt, A. R. (2012). Physical self-similarity enhances the gaze-cueing effect. *The Quarterly Journal of Experimental Psychology*, *65*(7), 1250–1259.

Jenkins, R., Lavie, N. & Driver, J. (2005). Recognition memory for distractor faces depends on attentional load at exposure. *Psychonomic Bulletin & Review*, *12*(2), 314–320.

Jenkins, R., White, D., Van Montfort, X. & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323.

Kaunitz, L. N., Rowe, E. G. & Tsuchiya, N. (2016). Large Capacity of Conscious Access for Incidental Memories in Natural Scenes. *Psychological Science*, *27*(9), 1266–77.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M. & Koch, I. (2010). Control and interference in task switching--a review. *Psychological bulletin*, *136*(5), 849–74.

Kirkham, A. J., Hayes, A. E. & Tipper, S. P. (2015). The effects of implicit and explicit emotion consistency on facial mimicry. *PloS one*, *10*(12), e0145731.

Klein, J. T., Shepherd, S. V. & Platt, M. L. (2009). Social Attention and the Brain. *Current Biology*, *19*(20), R958–62.

Knopman, D. & Nissen, M. J. (1991). Procedural learning is impaired in Huntington's disease: Evidence from the serial reaction time task. *Neuropsychologia*, *29*(3), 245–254.

Kościński, K. (2013). Perception of facial attractiveness from static and dynamic stimuli. *Perception*, *42*(2), 163–175.

Kramer, R. S. & Ward, R. (2010). Internal facial features are signals of personality and health. *Quarterly Journal of Experimental Psychology*, *63*(11), 2273–2287.

Kuhn, G., Benson, V., Fletcher-Watson, S., Kovshoff, H., McCormick, C. A., Kirkby, J. & Leekam, S. R. (2010). Eye movements affirm: Automatic overt gaze and arrow cueing for typical adults and adults with autism spectrum disorder. *Experimental Brain Research*, *201*(2), 155–165.

Kuhn, G., Tatler, B. W. & Cole, G. G. (2009). You look where I look! Effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition*, *17*(6-7), 925–944.

Kuzmanovic, B., Bente, G., von Cramon, D. Y., Schilbach, L., Tittgemeyer, M. & Vogeley, K. (2012). Imaging first impressions: Distinct neural processing of verbal and nonverbal social information. *NeuroImage*, *60*(1), 179–188.

Lachat, F., Conty, L., Hugueville, L. & George, N. (2012). Gaze Cueing Effect in a Face-to-Face Situation. *Journal of Nonverbal Behavior*, *36*(3), 177–190.

Lickel, B., Hamilton, D. L. & Sherman, S. J. (2001). Elements of a lay theory of groups: Types of groups, relational styles, and the perception of group entitativity. *Personality and Social Psychology Review*, *5*(2), 129–140.

Lundqvist, D., Flykt, A. & Öhman, A. (1998). The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91–630.

Manssuer, L. R. (2015). *Psychophysiological evidence of a role for emotion in the learning of trustworthiness from identity-contingent eye-gaze cues* (Doctoral dissertation, Bangor University).

Manssuer, L. R., Pawling, R., Hayes, A. E. & Tipper, S. P. (2015). The role of emotion in learning trustworthiness from eye-gaze: Evidence from facial electromyography. *Cognitive Neuroscience*, *8928*(October), 1–21.

Manssuer, L. R., Roberts, M. V. & Tipper, S. P. (2015). The late positive potential indexes a role for emotion during learning of trust from eye-gaze cues. *Social neuroscience*, *0919*(April), 1–16.

Marques, J. M., Yzerbyt, V. Y. & Leyens, J.-P. (1988). The 'black sheep effect': Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology*, *18*(1), 1–16.

McCall, C. & Singer, T. (2015). Facing Off with Unfair Others: Introducing Proxemic Imaging as an Implicit Measure of Approach and Avoidance during Social Interaction. *PLOS ONE*, *10*(2), e0117532.

Meissner, C. A. & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Ng, W.-J. & Lindsay, R. C. (1994). Cross-Race Facial Recognition: Failure of the Contact Hypothesis. *Journal of Cross-Cultural Psychology*, *25*(2), 217–232.

Nicholls, M. E. R., Loveless, K. M., Thomas, N. a., Loetscher, T. & Churches, O. (2014). Some participants may be better than others: Sustained attention and motivation are higher early in semester. *Quarterly journal of experimental psychology (2006)*, *68*(1), 10–18.

Oosterhof, N. N. & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–92.

Oosterhof, N. N. & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion (Washington, D.C.)* *9*(1), 128–133.

Park, B. & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068.

Pawling, R., Kirkham, A. J., Tipper, S. P. & Over, H. (2016). Memory for incidentally perceived social cues: Effects on person judgment. *British Journal of Psychology*, 1–22.

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Pelphrey, K. A., Morris, J. P. & McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain*, *128*(5), 1038–1048.

Porter, S., England, L., Juodis, M., Ten Brinke, L. & Wilson, K. (2008). Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science*, *40*(3), 171–177.

Raymond, J. (2009). Interactions of attention, emotion and motivation. *Progress in Brain Research*, *176*, 293–308.

Reber, R. & Schwarz, N. (1999). Effects of Perceptual Fluency on Judgments of Truth. *Consciousness and Cognition*, *8*(3), 338–342.

Reed, J. & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 585–594.

Reid, V. M., Striano, T., Kaufman, J. & Johnson, M. H. (2004). Eye gaze cueing facilitates neural processing of objects in 4-month-old infants. *Neuroreport*, *15*(16), 2553–2555.

Rescorla, R. A. & Wagner, A. R. (1972). 3 A Theory of Pavlovian Conditioning : Variations in the Effectiveness of Reinforcement and Nonreinforcement lIT.

Rezlescu, C., Duchaine, B., Olivola, C. Y. & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, *7*(3), e34293.

Roberts, S. C., Saxton, T. K., Murray, A. K., Burriss, R. P., Rowland, H. M. & Little, A. C. (2009). Static and dynamic facial images cue similar attractiveness judgements. *Ethology*, *115*(6), 588–595.

Rogers, R. D. & Monsell, S. (1995). Costs of a Predictable Switch Between Simple Cognitive Tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231.

Rogers, R. D., Bayliss, A. P., Szepietowska, A., Dale, L., Reeder, L., Pizzamiglio, G., ... Tipper, S. P. (2014). I want to help you, but I am not sure why: gaze-cuing induces altruistic giving. *Journal of experimental psychology. General*, *143*(2), 763–77.

Rubenstein, A. J. (2005). Variation in Perceived Attractiveness: Differences Between Dynamic and Static Faces. *Society*, *16*(10), 759–762.

Rule, N. O., Slepian, M. L. & Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition*, *125*(2), 207–218.

Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G., ... Vogeley, K. (2010). Minds made for sharing: initiating joint attention recruits

reward-related neurocircuitry. *Journal of cognitive neuroscience*, *22*(12), 2702–2715.

Slone, A. E., Brigham, J. C. & Meissner, C. A. (2000). Social and Cognitive Factors Affecting the Own Race Bias in Whites. *Basic and Applied Social Psychology*, *22*(2), 71–84.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R. & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7710–5.

Stirrat, M. & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: male facial width and trustworthiness. *Psychological science*, *21*(3), 349–54.

Strachan, J. W. A., Kirkham, A. J., Manssuer, L. R. & Tipper, S. P. (2016). Incidental Learning of Trust: Examining the Role of Emotion and Visuomotor Fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *Advance on*, 1–15.

Strachan, J. W. A., Kirkham, A. J. & Tipper, S. P. (n.d.). *Gaze cueing effects are the same for one's own face as for other faces.*

Strick, M., Holland, R. W. & van Knippenberg, A. (2008). Seductive eyes: Attractiveness and direct gaze increase desire for associated objects. *Cognition*, *106*(3), 1487–1496.

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C. & Heagins, T. B. (2015). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Methods*.

Süßenbach, F. & Schönbrodt, F. (2014). Not afraid to trust you: Trustworthiness moderates gaze cueing but not in highly anxious participants. *Journal of Cognitive Psychology*, (September 2015), 1–9.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D. & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118.

Suzuki, A. & Suga, S. (2010). Enhanced memory for the wolf in sheep's clothing: Facial trustworthiness modulates face-trait associative memory. *Cognition*, *117*(2), 224–229.

Tajfel, H., Billig, M. G., Bundy, R. P. & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178.

Takahashi, K. & Watanabe, K. (2013). Gaze cueing by pareidolia faces. *i-Perception*, *4*(8), 490–492.

ten Brinke, L., Vohs, K. D. & Carney, D. R. (2016). Can Ordinary People Detect Deception After All? *Trends in Cognitive Sciences*, *20*(8), 579–588.

Tiddeman, B., Burt, M. & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, *21*(5), 42–50.

Tipples, J. (2002). Eye gaze is not unique: Automatic orienting in response to uninformative arrows. *Psychonomic bulletin & review*, *9*(2), 314–318.

Todorov, A. T. & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological science*, *25*(7), 1404–1417.

Todorov, A., Baron, S. G. & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127.

Tomasello, M., Hare, B., Lehmann, H. & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, *52*(3), 314–320.

Tummeltshammer, K. S., Mareschal, D. & Kirkham, N. Z. (2014). Infants' selective attention to reliable visual cues in the presence of salient distractors. *Child Development*, *85*(5), 1981–1994.

Uleman, J. S., Adil Saribay, S. & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual review of psychology*, *59*, 329–60.

van Boxtel, A. (2010). *Facial EMG as a tool for inferring affective states*. Noldus Information Technology.

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W. & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(32), E3353–61.

Verosky, S. C., Todorov, A. & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia*, *51*(11), 2100–2108.

Wang, W. C., Ranganath, C. & Yonelinas, A. P. (2014). Activity reductions in perirhinal cortex predict conceptual priming and familiarity-based recognition. *Neuropsychologia*, *52*(1), 19–26.

Weibert, K. & Andrews, T. J. (2015). Activity in the right fusiform face area predicts the behavioural advantage for the perception of familiar faces. *Neuropsychologia*, *75*, 588–96.

Weigelt, M., Güldenpenning, I., Steggemann-Weinrich, Y., Alaboud, M. A. A. & Kunde, W. (2016). Control over the processing of the opponent's gaze direction in basketball experts. *Psychonomic Bulletin & Review*, 1–7.

Willis, J. & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598.

Wilson, J. P. & Rule, N. O. (2015). Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychological Science*, *26*(8), 1325–1331.

Winkielman, P. & Olszanowski, M. (2015). Faces In-Between : Evaluations Reflect the Interplay of Facial Features and Task-Dependent Fluency. *Emotion*, *15*(2), 232–242.

Wylie, G. & Allport, A. (2000). Task switching and the measurement of "switch costs". *Psychological Research*, *63*(3-4), 212–233.

Xu, S., Zhang, S. & Geng, H. (2011). Gaze-induced joint attention persists under high perceptual load and does not depend on awareness. *Vision Research*, *51*(18), 2048–2056.

Yeung, N. (2006). Between-Task Competition and Cognitive Control in Task Switching. *Journal of Neuroscience*, *26*(5), 1429–1438.