Exploring Social Metacognition: the role of confidence in updating estimates of advisor reliability with and without feedback.

Matt Jaquiery

Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2020

Abstract

This *R Markdown* template is for writing an Oxford University thesis. The template is built using Yihui Xie's `bookdown` package, with heavy inspiration from Chester Ismay's `thesisdown`, and the `OxThesis` LaTeX template (most recently adapted by John McManigle).

This template's sample content include illustrations of how to do the various things you need to write a thesis in R Markdown, and largely follow the structure from this R Markdown workshop.

Congratulations for taking a step further into the lands of open, reproducible science, by writing your thesis using a tool that allows you to transparently include tables and dynamically generated plots directly from the underlying data. Hip hooray!

Exploring Social Metacognition: the role of confidence in updating estimates of advisor reliability with and without feedback.

Matt Jaquiery

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

For TBC

Acknowledgements

This is where you will normally thank your advisor, colleagues, family and friends, as well as funding and institutional support. In our case, we will give our praises to the people who developed the ideas and tools that allow us to push open science a little step forward by writing plain-text, transparent, and reproducible theses in R Markdown.

We must be grateful to John Gruber for inventing the original version of Markdown, to John MacFarlane for creating Pandoc (`http://pandoc.org`) which converts Markdown to a large number of output formats, and to Yihui Xie for creating `knitr` which introduced R Markdown as a way of embedding code in Markdown documents, and `bookdown` which added tools for technical and longer-form writing.

Special thanks to Chester Ismay, who created the `thesisdown` package that helped many a PhD student write their theses in R Markdown. And a very special tahnks to John McManigle, whose adaption of Sam Evans' adaptation of Keith Gillow's original maths template for writing an Oxford University DPhil thesis in LaTeX provided the template that I adapted for R Markdown.

Finally, profuse thanks to JJ Allaire, the founder and CEO of RStudio, and Hadley Wickham, the mastermind of the tidyverse without whom we'd all just given up and done data science in Python instead. Thanks for making data science easier, more accessible, and more fun for us all.

Ulrik Lyngs
Linacre College, Oxford
2 December 2018

Abstract

This *R Markdown* template is for writing an Oxford University thesis. The template is built using Yihui Xie's `bookdown` package, with heavy inspiration from Chester Ismay's `thesisdown`, and the `OxThesis` LaTeX template (most recently adapted by John McManigle).

This template's sample content include illustrations of how to do the various things you need to write a thesis in R Markdown, and largely follow the structure from this R Markdown workshop.

Congratulations for taking a step further into the lands of open, reproducible science, by writing your thesis using a tool that allows you to transparently include tables and dynamically generated plots directly from the underlying data. Hip hooray!

Contents

List of Figures

List of Tables

List of Abbreviations

1-D, 2-D . . . One- or two-dimensional, referring in this thesis to spatial dimensions in an image.

Otter . . . . . One of the finest of water mammals.

Hedgehog . . . Quite a nice prickly friend.

Part I

Introduction

1 Introduction

!TODO[Style notes]:

- Graphs:
  - Continuous axes which show theoretical limits should have small expand, axes which do not show those limits should have arrowheads from broken_axis_*

!TODO[update evo models to use this terminology]

$a_{i,t}$ - advice agent i receives at time t

$b_{i,t}$ - bias of agent i at time t

$c_{i,t}^{I/A/F}$ - agent i's confidence in their initial/advisory/final estimate at time t

$e_{i,t}^{I/A/F}$ - agent i's initial/advisory/final estimate at time t

i, j - agent identifiers

$s_{i,t}$ - sensitivity of agent s at time t

u_i - fitness/utility of agent i

v - true value in the world

λ - learning rate

$\omega_{i,t}$ - weighting for agent i's estimate at time t

This thesis attempts to determine whether individual psychological processes in the seeking and taking of advice are sufficient to produce entrenched biases at the network level. The work includes empirical investigation into the individual psychological processes underlying advice seeking and advice taking using behavioural experiments; computational modelling and behavioural experimentation exploring the effects of contextual factors on advice-taking; agent-based computational modelling of the effects interactions between agents with psychological advisor evaluation processes that produce biases in assimilation of information and source selection; and a comparison between the structures of networks produced in these models and those naturally occurring social networks. The organisation is as follows: this introduction establishes the core concepts invoked in this thesis, and describes their treatment in the literature; the following sections each include a short chapter on the specific question addressed and the techniques used, a detailed description of the work conducted, and a short discussion of the conclusions drawn from the work; and a final section offers broader conclusions arising as a consequence of the presented work, alongside some suggestions for related research.

Overview

The first line of behavioural experiments report investigates whether people decide which advisors to consult using internal signals where external feedback is unavailable, extending a previous finding concerning the influence of advice (**pescetelliRoleDecisionConfidence2018**) into the domain of advisor choice. While little support is found for the metacognitive mechanisms underpinning the influence of advice, the frequency of agreement is shown to be a good indicator of advisor choice. Next, the effects of advisor agreement are explored in terms of influence using a different decision-making domain; a date estimation rather than perceptual decision task. Again, people denied objective feedback on the quality of advice are shown to attend to agreement. The work above demonstrates how the influence of advice depends upon the reputation of an advisor built up over time. In a second line of experiments, reputations of advisors are established at the outset

and the responses to different pieces of advice are explored. People consider both the properties of advice and the properties of advisors, placing less trust in advice which appears suspicious or which comes from a suspicious source. These behavioural results support implications from evolutionary models which demonstrate that the presence of even a few bad actors in a population can mean that distrusting all advice from all sources a little is adaptive. The evolutionary models are extended to show that this kind of general distrust, known as egocentric discounting, emerges as adaptive even where none of the agents deliberately mislead others, and even where all agents are equally skilled and fully cooperative. Finally, network effects arising from interconnected networks of agents are explored. Agents are based on the empirical results heretofore presented, and the consequences are explored for networks with various starting structures characterised by sparsity and homogeneity, under varying rates and reliabilities of external feedback. !TODO[results of this]. The resulting network characteristics are compared with network structures from real online networks from social media websites. !TODO[results of this].

Advice

Advice is broadly defined as information which comes from a social source. Advice is therefore different from other sources of information in that it is the result of (at least) mental processing of other information. In some cases, it may additionally include discussions among different group members (e.g. advice from the International Advisory Panel on Climate Change). Throughout this thesis, the focus is primarily on advice which comes from a single, stable source, as when we see a post by an acquaintance on social media, or when a stranger provides us with advice.

Advice-taking

Advice occurs in the context of a decision, and forms a part of the information which is integrated during the decision-making process to produce a decision. To the extent that the decision reached differs from the decision that would have occurred

had the advice not been presented, the advice has had an effect on the decision; to the extent that this difference changes the decision in a way consistent with the advice, the advice has been 'taken' (as opposed to 'rejected').

It is tacitly implied by many operationalizations of advice-taking that the informational content of the advice determines the extent to which it is taken or rejected !TODO[CITE]. Insofar as the identity of the advisor matters, it matters because it functions as a cue to the informational content of the advice. This is likely a major oversimplification, however, because in many real-world contexts advice-giving and advice-taking form part of a developing social relationship: being consulted for advice and having one's advice followed are inherently rewarding (**hertzIntrinsicValueSocial2018**; **hertzNeuralComputationsUnderpinning2017**); and taking advice can serve as a (sometimes costly) social signal of valuing a relationship with a person or group (**byrneOstracismReducesReliance2016**). Furthermore, some authors have argued that people may perceive taking advice as sacrificing their independence or autonomy !TODO[CITATIONNEEDED]. While this thesis follows previous literature in omitting to consider the wider social concerns influencing the taking of advice, it is nevertheless important to remember that the processes investigated herein take place in a variety of social contexts where complex social agents attempt to optimise over numerous goals over numerous timescales.

Three-factor model of trust

The degree to which advice is taken is proportional to the trust placed in the advisor by the decision-maker. Interpersonal trust, or the degree to which one is prepared to place one's fortune in the hands of another (e.g. by relying on their advice), is apportioned by Mayer et al. (**mayerIntegrativeModelOrganizational1995**) onto three properties of the advisor (as judged by the decision-maker): ability, benevolence, and integrity. To these three properties of the advisor we may add the decision-maker's general propensity to trust, as well as situational cues and

task cues (e.g. the phenomenon that advice is more readily taken for hard tasks than easy ones, (Gino and Moore 2007)).

Ability

Ability captures the expertise of an advisor: their raw ability to perform the task for which they are giving advice. In some cases this is relatively straightforward, as in the expertise of a General Practitioner in matters of health and disease, and in others more complex, as in the expertise of a hairdresser when deciding on a haircut (when matters of personal taste comingle with aesthetic considerations of facial structure, practical considerations of hair constitution, and social considerations of fashion). The greater the ability of an advisor, the greater the influence of their advice, as demonstrated by experiments showing that participants' decisions are more affected by the advice of advisors who are labelled as more expert in a relevant domain (**sahCheapTalkCredibility2013**; Schultze, Mojzisch, et al. 2017; Sniezek, Schrah, et al. 2004; Sniezek and Van Swol 2001; Soll and Mannes 2011), or are shown to be more expert empirically (**pescetelliUseMetacognitiveSignals2017**; **sahCheapTalkCredibility2013**; Ilan Yaniv and Kleinberger 2000).

Benevolence

Benevolence refers to the extent to which the advisor seeks to further the interests of the decision-maker. Where ability represents the absolute limit on the quality of advice, benevolence represents the extent to which the advice approaches this limit. The advice of even a renouned expert may be doubted if there is reason to believe their goal is to mislead, a vital lesson for medieval monarchs with their councils of politicking advisors. Experimental work has shown that psychology students relied more on the advice of their friends than on the advice of labelled experts !TODO[CITATIONNEEDED], and that participants are more inclined to reject advice when uncertainty is attributed to malice rather than ignorance (**schulInfluencesDistrustTrust2015**)}.

Integrity

Advisors with integrity exhibit adherence to principles which the decision-maker endorses. As with benevolence, integrity acts to determine the extent to which advice approaches the limit imposed by ability. While not mutually exclusive, integrity is typically important where relationships are less personal (e.g. we may place great trust in a General Practitioner because of their expertise in medical matters and their *integrity* in adhering to a set of professional ethical and conduct requirements). !TODO[Some description of the research]

Normative models of advice-taking

Advice-taking can be evaluated formally with reference to a normative model. The simplest and most common of these views the decision-making task as an estimation problem (or combination of estimation problems), and provides an approximately Bayesian variance-weighted integration of independent estimates. To borrow from Galton (**galtonVoxPopuliWisdom1907**)}, consider the task of judging the weight of a bullock. We can model any single guess (e) as the true weight (v) plus some error (ϵ):

$$e = v + \epsilon \tag{1.1}$$

The key insight is to observe that the error is drawn from a normal distribution ($\mathcal{N}(\mu = 0, \sigma^2)$)[1]. As the number of samples from this distribution increases, the mean of those samples tends towards the mean of the distribution. Thus, the more estimates are taken, the closer on average the sum of errors will be to 0.

[1]The normal distribution is well-supported by empirical evidence, but note that any symmetrical distribution around 0 will lead to the same conclusion.

$$\frac{\sum_i^N (e_i)}{N} = \frac{\sum_i^N (v + \mathcal{N}(\mu = 0,\, \sigma_i^2))}{N} \tag{1.2}$$

$$\frac{\sum_i^N (e_i)}{N} = \frac{\sum_i^N (v)}{N} + \frac{\sum_i^N (\mathcal{N}(\mu = 0,\, \sigma_i^2))}{N} \tag{1.3}$$

$$\frac{\sum_i^N (e_i)}{N} = \frac{Nv}{N} + \hat{0} \tag{1.4}$$

$$\frac{\sum_i^N (e_i)}{N} \approx v \tag{1.5}$$

Observe that this formulation is true no matter the value of N. On average, it is always better to have more estimates than fewer. This suggests that, even in the situation where there are only two estimates (the decision-maker's and the advisor's), the best policy will be to incorporate both estimates into the final decision.

The variance of the normal distribution from which errors are derived (σ_i^2) is, in the example above, drawn from a normal distribution itself ($\sigma_i^2 \sim \mathcal{N}(\mu = 0,\, \text{sd}^2)$meaning that it is also cancelled out on average over repeated samples). Where few estimates are taken, weighting those estimates by the variance of the error distributions will increase the accuracy of the estimates in proportion to the difference between the variances:

$$e^F = \frac{\frac{1}{N}\sum_i^N \omega_i e_i^I}{\sum_i^N \omega_i}$$

Where ω_i is $1/\sigma_i^2$.

Many experimental implementations of this model avoid weighting issues by calibrating decision-makers and advisors to be equally accurate on average ($\sigma^2_{\text{decision-maker}} = \sigma^2_{\text{advisor}}$). The result of this constraint is that the optimal policy is simply to average all estimates together:

$$e^F = \frac{1}{N}\sum_i^N e_i^I \approx v$$

Egocentric-discounting

From the perspective of the normative model above, decision-makers should weigh their own estimate equally with each other estimate they receive in the process of coming to their decision. One of the most robust findings in the literature on advice-taking is that people routinely underweight advisory estimates relative to their own estimates, a phenomenon known as *egocentric discounting*(**danaAdviceChoice2015**; **minsonCostCollaborationWhy2012**; **raderAdviceFormSocial**
Gino and Moore 2007; Hütter and Ache 2016; Liberman et al. 2012; Ronayne and Sgroi 2018; See et al. 2011; Soll and Mannes 2011; Trouche et al. 2018; Ilan Yaniv and Kleinberger 2000; Ilan Yaniv and Choshen-Hillel 2012; Ilan Yaniv and Milyavsky 2007). Egocentric discounting occurs in both feedback and no-feedback contexts (Ilan Yaniv and Kleinberger 2000). Explanations for egocentric discounting are usually framed in terms of personal-level psychology: decision-makers have better access to reasons for their decision (Ilan Yaniv and Kleinberger 2000); overrate their own competence (Sniezek, Schrah, et al. 2004); may have a desire to appear consistent (Ilan Yaniv and Milyavsky 2007); may see opinions as possessions (Soll and Mannes 2011); may be loss-averse to providing a worse final estimate due to advice-taking (Soll and Mannes 2011); or have difficulty avoiding anchoring (Schultze, Mojzisch, et al. 2017) or repetition bias effects (Trouche et al. 2018). None of these explanations has survived rigorous empirical testing, however, and recently suggestions have widened to include consideration of aggregate-level rather than personal-level causes, with Trouche et al. (2018) arguing that the potential for misaligned incentives between decision-maker and advisor motivate discounting of advice. In the course of this thesis `crossref needed`, I demonstrate that egocentric discounting may be a stable metastrategy which protects against exploitation, carelessness, incompetence, and miscommunication. From this perspective, the normative model pertains to a particular instantiation of a problem with questionable ecological validity given the typical ethology of advice-taking in humans. While such considerations affect the conclusions one draws from egocentric discounting relative to the normative model,

they do not detract substantially from the practice of using the normative model as an optimum 'set point' from which to evaluate advice-taking behaviour.

Homophily and echo-chambers

Homophily is the ubiquitous phenomenon that individuals more closely connected to one another within a social network tend to be more similar to one another than would be expected by chance across numerous dimensions, from demographics to attitudes (**mcphersonBirdsFeatherHomophily2001**). Whether homophily in virtual social networks is responsible for increases in polarisation is debated. Proponents point to increases in polarisation (e.g. in politics: !TODO['Pew Research Center, 2014']), to empirical studies demonstrating homophily in virtual social networks (**cardosoTopicalHomophilyOnline2017**; **colleoniEchoChamberPublic2014**), and to studies examining selective exposure online (**kobayashiSELECTIVEEXPOSUREPOLITI** and to echo chambers: egregious examples of highly homophilous networks with pathological polarisation (**sunsteinRepublicCom2002**; **sunsteinRepublicDividedDemocracy2** The empirical components of the argument are contested, with evidence that virtual social networks are less homogenous than offline social networks (**barberaHowSocialMedia2015**), and that selective exposure is a somewhat dubious finding which does not show up clearly online (**garrettEchoChambersOnline2009**; **garrettPoliticallyMotivatedReinforceme** **nelsonMythPartisanSelective2017**; **searsSelectiveExposureInformation1967**). Modelling work demonstrates, however, that where there is a bias in assimilation of information, homophily exacerbates polarisation (**dandekarBiasedAssimilationHomophily2013**) Where polarisation in turn increases homophily, for example through selective exposure or avoidance, a self-reinforcing spiral emerges wherein social connections become increasingly homogenous and attitudes increasingly extreme (**songDynamicSpiralsPut2017**).

Source selection and information weighting

Context-dependency of epistemic processes

Part II

Psychology of advice

Contents

2

Psychological mechanisms of advisor evaluation

Where feedback is unavailable, people may use their own sense of certainty as a yardstick for evaluating advice (**pescetelliRoleDecisionConfidence2018**; **pescetelliUseMetacogniti** advisors who agree when one is confident are perceived as more helpful; while those who disagree when one is confident are perceived as less helpful. Confidence serves as a proxy for objective feedback, and functions well in this role insofar as the judge has high metacognitive resolution (i.e. higher confidence is indicative of a greater probability of being correct).

Use of advice

Normative models of advice taking state that averaging estimates minimises errors. As discussed at length in Section III, the assumptions underlying the normative model do not always hold in the real world. The performance of the normative model can be characterised according to differences between the advisor and the decision-maker on ability and bias (Soll and Larrick 2009) !TODO[how does PAR relate to what's going on here?]. Recall that the normative model states that

advice should contribute to the final decision in proportion to the ability of the advisor compared to the decision-maker.

$$e_i^F = \frac{\omega_i e_i^I + \omega_j e_j^I}{2} \tag{2.1}$$

Where agent i is the decision-maker and agent j is the advisor. This weighting can be simplified to be expressed only in terms of the decision-maker's weighting of the advisor because the two are constrained to sum to 1 by virtue of being relative to one another:

$$\omega_i + \omega_j = 1 \tag{2.2}$$

$$\omega_j = 1 - \omega_i \tag{2.3}$$

$$\therefore e_i^F = \frac{(1-\omega_j)e_i^I + \omega_j a_i}{2} \tag{2.4}$$

Where a_i is the advice received (i.e. agent j's initial estimate - $a_i \equiv e_j^I$). In the normative model, the weighting is equivalent to the ratio of variance of the errors made by each agent:

$$\omega_j = \frac{\sigma_j^2}{\sigma_j^2 + \sigma_i^2} \tag{2.5}$$

The normative model thus represents weighting by relative ability. Precise knowledge of the ability of others relative to oneself is rarely available in the real world, however, and, as discussed in Section III, other assumptions concerning the trustworthiness or interpretability of advice may be violated.

The normative model can be adapted to provide a more psychologically-realistic account of advice usage by substituting the three factor model of trust into the equations in place of the ability variable. We start with the statement within the three factor model that trust (ω) is proportional to ability (a), benevolence (b), and integrity (g).

$$\text{trust} \propto \text{ability} + \text{benvolence} + \text{integrity} \tag{2.6}$$

We can thus replace the measure of accuracy in the normative model with the measure of trust in order to calculate the relative weighting:

$$\omega_j = \frac{\text{trust}_j}{\text{trust}_j + \text{trust}_i} \tag{2.7}$$

At this point we may question whether the variable trust_i is a meaningful property or simply an artefact of mathematical symbol manipulation. Mathematically it provides a fixed point against which trustworthiness of advisors can be measured, allowing for scaling weightings meaningfully across different advisors in different decisions. In real world terms, while it is generally unlikely that benevolence_i and integrity_i will be anything less than maximal, perceptions of one's own ability (ability_i) are likely to allow for others to exceed it. I make no strong claims on the relationship between trust and its component variables other than proportionality, and within this conception it is meaningful to consider weighting as a property of trust in another's judgement relative to one's own, adjusted in some manner for the perception of that other's benevolence and integrity. If the concept of self-trust still appears untenable, note that trust_i can be replaced with a constant without compromising the equations.

We thus return to the normative model of advice-taking, but with an alternative derivation of the weighting between advice and initial estimate:

$$e_i^F = \frac{(1-\omega_j)e_i^I + \omega_j a_i}{2} \tag{2.8}$$

2.0.1 Critique of the aggregation model

This conception of advice-taking as a weighted aggregation process between an initial estimate and advice underpins both the modelling and the experiments

presented in this thesis. It is thus worth taking a little space to highlight areas in which this model is known to depart from reality.

Generality beyond judge-advisor systems

Firstly, the model is an idealised situation approximated by the experimental method: a decision-maker makes an explicit initial estimate, then receives advice, then makes an explicit final decision. Ilan Yaniv and Choshen-Hillel (2012) showed that preventing decision-makers from making initial decisions resulted in very different advice weighting, suggesting that this may be a model of a specific scenario rather than of advice integration per se. The model presented here could in principle explain an integration process where an initial estimate can only be made after the advice is known, but empirically performs poorly. At best, it could be argued that pre-exposure to the advice either anchors the initial estimate (thus moving e_i^I systematically closer to a_i), or that having to trust advice because one cannot make one's own decision inflates the weighting of the advisor.

Multiple advisory estimates

Secondly, the model does not perform well when multiple advisors are consulted. The normative model, and the psychological derivative, predicts that a decision-maker's estimate ought to be weighted in conjunction with the other estimates. In other words, as the number of advisory estimates increases, the weight of the initial estimate should decrease. Hütter and Ache (2016) presented evidence that this does not happen: the weight of the initial estimate stays relatively constant while the weights of the advisor estimates are reduced. This implies that if a decision-maker were to average evenly their initial estimate with an advisor estimate ($\omega_i = .5$; $\omega_j = .5$), adding an extra advisor estimate would result in the weights of the advice being halved while the weight of the initial decision remained constant ($\omega_i = .5$; $\omega_{j \neq i} = .25$), rather than the more transparently optimal policy of weighting all estimates evenly ($\omega_i = \omega j \neq i = 1/3$).

Individual trial data

The model is supported by patterns in averages. Analysis of individual trials shows that the aggregate patterns of advice-taking appear to be roughly distributed between an averaging strategy and a picking strategy (Soll and Mannes 2011; Soll and Larrick 2009). The model, derived from these patterns, approximates the contribution of an individual trial to the overall average rather than the actual advice-taking strategy on any given trial.

2.0.2 Justification for use of the aggregation model

The criticisms above are important, but they do not invalidate the model for use in the present project. Here we seek to establish how differences in advice taking manifest according to properties of advisors. These differences are well characterised by the model, especially in the judge-advisor system used for the experiments. All models are inexact descriptions of reality, and inclusion of a more complex model capable of handling the cases outlined above would require greatly increased complexity for relatively little gain in explanatory power. For studying the questions at hand, the psychological model is an appropriate and useful approximation of human behaviour.

Updating advisor weights

The weights assigned to the advisors (relative to the decision-maker themself) are subject to change as the result of experience. This experience can be exogenous or endogenous to the decision-making task. In the exogenous case, advisors may be labelled in a particular way (Önkal, Gönül, et al. 2017; Tost et al. 2012; Schultze, Mojzisch, et al. 2017) or have some summary of their performance displayed (Gino, Brooks, et al. 2012; Ilan Yaniv and Kleinberger 2000). Endogenous experience refers to the information that advice on a given trial carries about the trustworthiness of an advisor. Exogenous experience is relatively straightforward, but endogenous experience requires some explanation.

Endogenous experience of advice means that the weighting of an advisor is in part dependent upon the past advice offered by that advisor. As each piece of advice is evaluated, the overall weighting of the advisor is updated accordingly. For clarity, two simplifiying assumptions are made in the explanation below. Firstly, while it is probable that properties of the advice are used to inform the dimensions of ability, integrity, and benevolence simultaneously, the examples below will deal with ability in isolation. Another project could explore in detail how experience of advice on any given trial updates an advisor's position in 3-dimensional trust space in a Bayesian manner according to the relative certainties about each dimension. This would capture the task of assigning blame for erroneous advice (e.g. was it unintentionally poor - a failure of ability - or deliberately misleading - a failure of benevolence?). Such an undertaking is beyond the scope of this project.

Secondly, it is assumed that advice is judged on its own merit as an estimate rather than on its usefulness as advice. The former means that advice is assessed in terms of the optimality of the decision recommended by the advice itself. The latter assesses advice based on the optimality of the decision based on advice relative to the optimality of the decision which *would have been made had the advice not been received*. There is some evidence that people alter their advice-giving behaviour in anticipation of discounting on the part of the decision-maker !TODO[CITE; for a case in human-machine teaming see @azariaStrategicAdviceProvision2016], somewhat akin to starting negotiations with a higher demand than one is hoping to settle for. There is no evidence as yet as to whether decision-makers anticipate and adjust for this adjustment on the part of the advisor. For the questions considered here, conclusions obtained under these simplifying assumptions are likely to hold even when the additional complexity is restored. The effects in the real world of interactivity between trust dimensions and game theoretic adjustments in the giving and interpretation of advice are likely to be small in comparison to general effects of advisor updating.

Evaluation of advice

A single piece of advice can be evaluated using its own properties and the properties of the advisor giving the advice. Furthermore, that evaluation can serve to update the properties of the advisor. A piece of advice's own properties will include its plausibility (e.g. participants in estimation tasks discount advice which is distant from their own initial estimates more heavily (I. Yaniv 2004)), while the properties of the advisor will include the advisor's trustworthiness (see above). The updating of trust following experience of advice is likely to be largely in the domain of ability, although other domains may be affected where the advice is particularly egregious.

Updating advisor evaluations

While a single piece of advice must be taken on its own terms, people can construct relatively accurate estimates of advisors' advice when provided with feedback on the decisions they use the advice to make (**pescetelliUseMetacognitiveSignals2017**; **sahCheapTalkCredibility2013**; Ilan Yaniv and Kleinberger 2000). This likely happens as an analogue of reinforcement learning, where feedback allows an error signal to be used to update the estimate of the advisor's ability ($\hat{s}^a$) rather than one's own beliefs about the world, according to some learning rate (λ).

this is wrong; need to check RL models for an analogue (2.9)

$$\hat{s}_{t+1} = \hat{s}_t + |e_t^a - v| \cdot \lambda \tag{2.10}$$

2.0.3 Criticism of the advisor evalutation model

Ecological validity

While many experiments have established the existence reinforcement learning in humans and other animals, it is unclear whether reinforcement learning operates in the social domain in which advising takes place. It is not obvious that there are many situations in the course of everyday relationships which can be characterised by the rapid advice-feedback cycle required to learn about advisor ability in the

manner modelled above. **feldmanhallViewingAdaptiveSocial2019** argued in a review that a wide variety of social phenomena could be explained via reinforcement learning processes. Additionally, **heyesKnowingOurselvesTogether2020** have argued that social learning is wholly explicable in terms of general reinforcement learning processes paired with attentional biases to social stimuli. Reinforcement learning in the social domain operates on the basis of rapid feedback, just as in the non-social domain. Below, the advisor evaluation model is extended to cases where objective feedback is not available by substituting the decision-maker's confidence for objective feedback. While not foolproof, the method allows better-than-average approximation of the quality of advisors provided several plausible assumptions are met !TODO[Discuss assumptions somewhere: independence of errors, better-than-chance accuracy, trying to help, rough metacognitive calibration/resolution].

Advisor evaluation without feedback

Where feedback is not available, participants in experiments continue to demonstrate an ability to respond rationally to differences in advisor quality (**pescetelliUseMetacognitiveSignal**
This is evidently not done through access to the correct real-world values, because feedback providing those values is unavailable, and, were participants aware of those values themselves, it stands to reason they would have provided those values (and thus not require advice!). Pescetelli and Yeung (**pescetelliUseMetacognitiveSignals2017**) suggest the mechanism for this ability to discriminate between advisors in the absence of feedback is performing updates based on confidence-weighted agreement.

Agreement

Consider first the non-weighted agreement case, where the advisor's estimate at time t (e^a_t) and the decision-maker's estimates (e^d) are binary ($\in 0, 1$). The estimate of the advisor's ability ($\hat{s}^a$) is updated positively if the advisor and decision-maker agree, and negatively otherwise, according to the learning rate λ.

$$\hat{s}^a_{t+1} = \hat{s}^a_t + (-2\ |e^d_t - e^a_t| + 1) \cdot \lambda \tag{2.11}$$

Confidence-weighted agreement

The updating of advice contingent on agreement may be weighted by confidence in the initial decision (c^d), such that agreement and disagreement are considered more informative about the quality of the advice when the decision with which they agree or disagree is more certain.

$$\hat{s}^a_{t+1} = \hat{s}^a_t + (-2\ |e^d_t - e^a_t| + 1) \cdot c^d_t \cdot \lambda \tag{2.12}$$

Continuous estimate case

2.1 Measuring advice-taking

Judge-advisor system

Table of experiments

This is now outdated. Update or remove?

Advisors	Choice	Task	Feedback	Result
AiC vs AiU	Yes	Perceptual, binary (MATLAB)	No	Suggestive
AiC vs AiU	Yes	Perceptual, binary	No	Inconclusive
In/accurate	Yes	Perceptual, binary	No	Accuracy selected more often
Low/High agreement	Yes	Perceptual, binary	No	Agreement selected more often

Advisors	Choice	Task	Feedback	Result
Accurate vs Agreement	No	Estimation, continuous	Yes and No	Accurate preferred given feedback, agreement preferred without feedback

3

Behavioural experiment method

The behavioural experiments reported in this thesis share a common structure. This structure is detailed here to reduce repetition elsewhere in the thesis. Individual experiments have truncated methods sections in which the specific deviations from the general method are noted.

3.1 General method

The experiments take place using a judge-advisor system. Participants give an **initial estimate** for a decision-making task, receive **advice**, and then provide a **final decision**. The advice is always computer-generated, although the specifics of the generating procedure vary between experiments.

3.1.1 Participants

Recruitment

Human participants were recruited from the online experiment participation platform Prolific (`https://prolific.co`). Participants were prevented from taking the study if they had participated in one of the other studies in the thesis, or if they had an overall approval rating on Prolific of less than 95/100.

Payment

Participants were paid approximately GBP10-15/hour pro rata. Experiments took the average participant between 10 and 30 minutes to complete.

Some participants encountered technical problems, prompting them to contact me via the Prolific platform. These participants were thanked, and additional information about the problem sought if necessary. These participants were paid on an ad hoc basis depending upon the time taken before the errors emerged and the detail of the error reports.

Later studies introduced attention checks which terminated the study as a consequence for failure. It is not clear whether this technique constitutes best practice on Prolific because automatic termination means participants may return the study rather than having their participation explicitly rejected (and thus affecting their Prolific participation rating). Participants who returned the studies were not paid. Participants who attempted to complete the study with an invalid code after having their participation terminated for failing attention checks were also not paid, and their completion attempt was rejected on the Prolific platform. There is an ongoing ethical debate concerning non-payment of participants who fail attention checks in online studies. The studies in this thesis used a mixture of paying for and not paying for participation attempts with failed attention checks. In online studies, where low-effort participation is a serious and enduring concern, platforms such as Prolific make clear to participants and researchers that payment is only expected for responses which are given with satisfactory effort. Participants are thus fully aware of and consenting to the process of screening results for adequate effort prior to payment. It is important to note the difference between low effort responses, for which payment may ethically be withheld, and atypical responses, which may represent genuine engagement with the task. It is unethical, in my view, to withhold payment from participants for atypical responses within an experiment (including low or high response times, accuracy, etc).[1] Participants should only be denied

[1]Data may be excluded from *analysis* for these reasons, but the participants should still be paid.

payment for failing to provide adequate responses to explicit attention checks.

Demographics

Demographic information on participants, such as age and gender, was not collected. While there is a robust case for collecting these data and conducting sex-disaggregated analyses (**criadoperezInvisibleWomenExposing2019**), initial concerns over General Data Protection Regulation resulted in a cautious approach to the collection of data concerning protected characteristics of participants. !TODO[move to discussion]Gender differences, whether due to socialisation, biological factors, or their interactions, may well alter advice-taking and expressed confidence in decisions. I suspect, although I can offer no evidence, that gender differences in the results presented in this thesis will at most show overlapping distributions. I do not think it highly plausible that different strategies are wholly the preserve of any particular gender, or that egocentric discounting is markedly stronger in any particular gender.

Participants were at least 18 years of age, confirmed by the requirements for possessing an account on the Prolific platform and by explicit confirmation when giving informed consent.

3.1.2 Ethics

Ethical approval for the studies in the thesis was granted by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (References: R55382/RE001; R55382/RE002).

3.1.3 Procedure

Participants visited the Uniform Resource Locator for the study by following a link from Prolific using their own device (Figure 3.1). Early studies only supported computers, but later studies included support for tablets and smartphones. After viewing an information sheet describing the study and giving their consent to participate, participants began the study proper. The study introduced the software

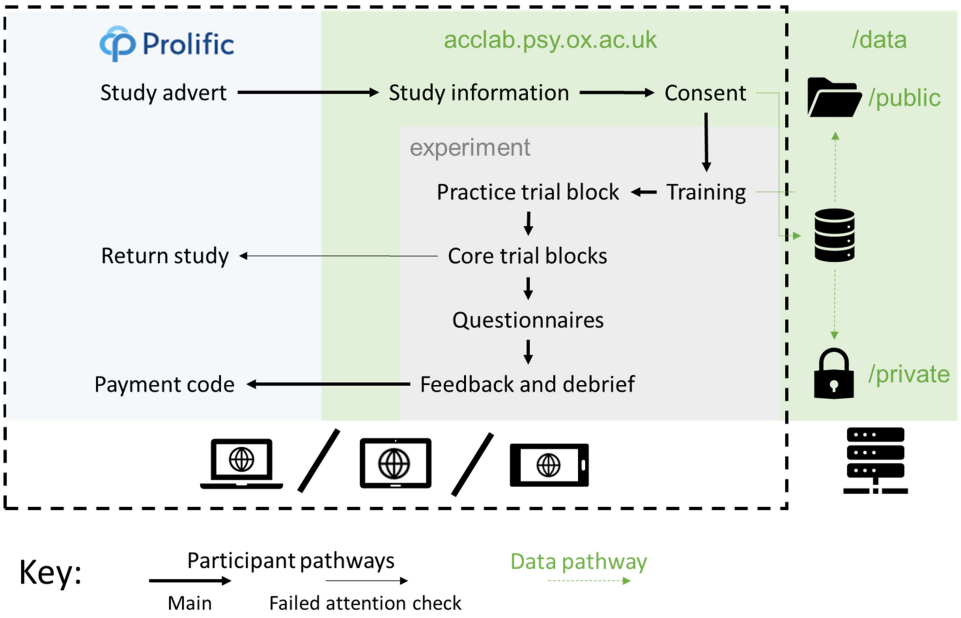


Figure 3.1: Participant pathway through the studies.
Participants used their own devices to complete the study, which was presented on a website written in HTML, CSS, and JavaScript. The data were saved on the server using PHP.

to the participant interactively, demonstrating the decision-making task and how responses could be made. Next, participants were given a block of practice trials to familiarise them with the decision-making task. Participants were then introduced to advice, and given a block of practice trials in which they received advice. The core experimental blocks followed the practice with advice. Finally, debrief questions were presented and feedback provided concerning the participant's performance, including a stable link to the feedback and a payment code. The participant entered the payment code into the Prolific platform and their participantion was at an end.

On each trial, participants were faced with a decision-making task for which they offered an initial estimate. They then received advice (on some trials they were able to choose which of two advisors would provide this advice). They then made a final decision. On feedback trials, they received feedback on their final decision. The schematic for this trial structure is shown for the Dots Task in Figure 3.2.

!TODO[Check feedback duration and style in image caption]

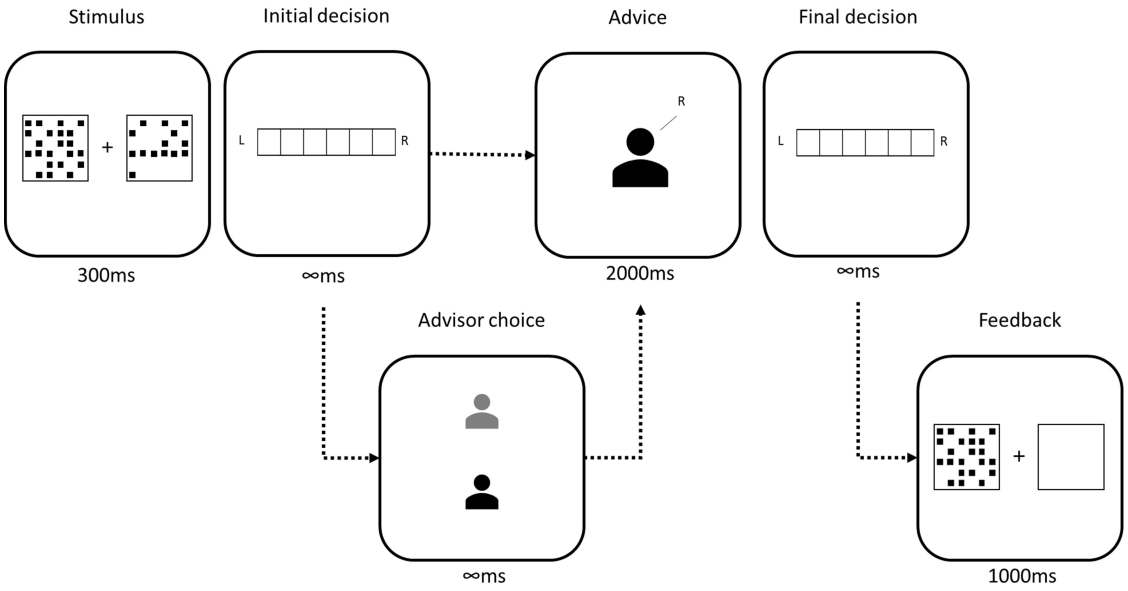


Figure 3.2: Trial structure of the Dots Task.
In the initial estimate phase, participants saw two boxes of dots presented simultaneously for 300ms. Participants then reported whether there were more dots on in the left or the right box, and how confident they were in this decision. Participants then received advice, sometimes being offered the choice of which advisor would provide the advice. The advice was displayed for 2000ms before participants could submit a final decision, again reporting which box they believe contained more dots and their confidence in their decision. On feedback trials, feedback was presented by redisplaying the correct box while showing the other box as empty.

Perceptual decision (Dots Task)

Stimuli in the Dots Task consisted of two boxes arranged to the left and right of a fixation cross (Figure 3.3). These boxes were briefly and simultaneously filled with an array of non-overlapping dots, and the participant was instructed to indentify the box with the most dots. The number of dots was exactly determined by the difficulty of the trial: the box with the !TODO[check this description] least dots had 200 - the difficulty, while the box with the most had 200 + the difficulty. The dots did not move during the presentation of the stimulus. There was thus an objectively correct answer to the question which, given enough time, could be precisely determined from the stimulus.

The Dots Task stimuli can be customised to make the discrimination easier or more difficult. This means that the stimuli can be adjusted to maintain a

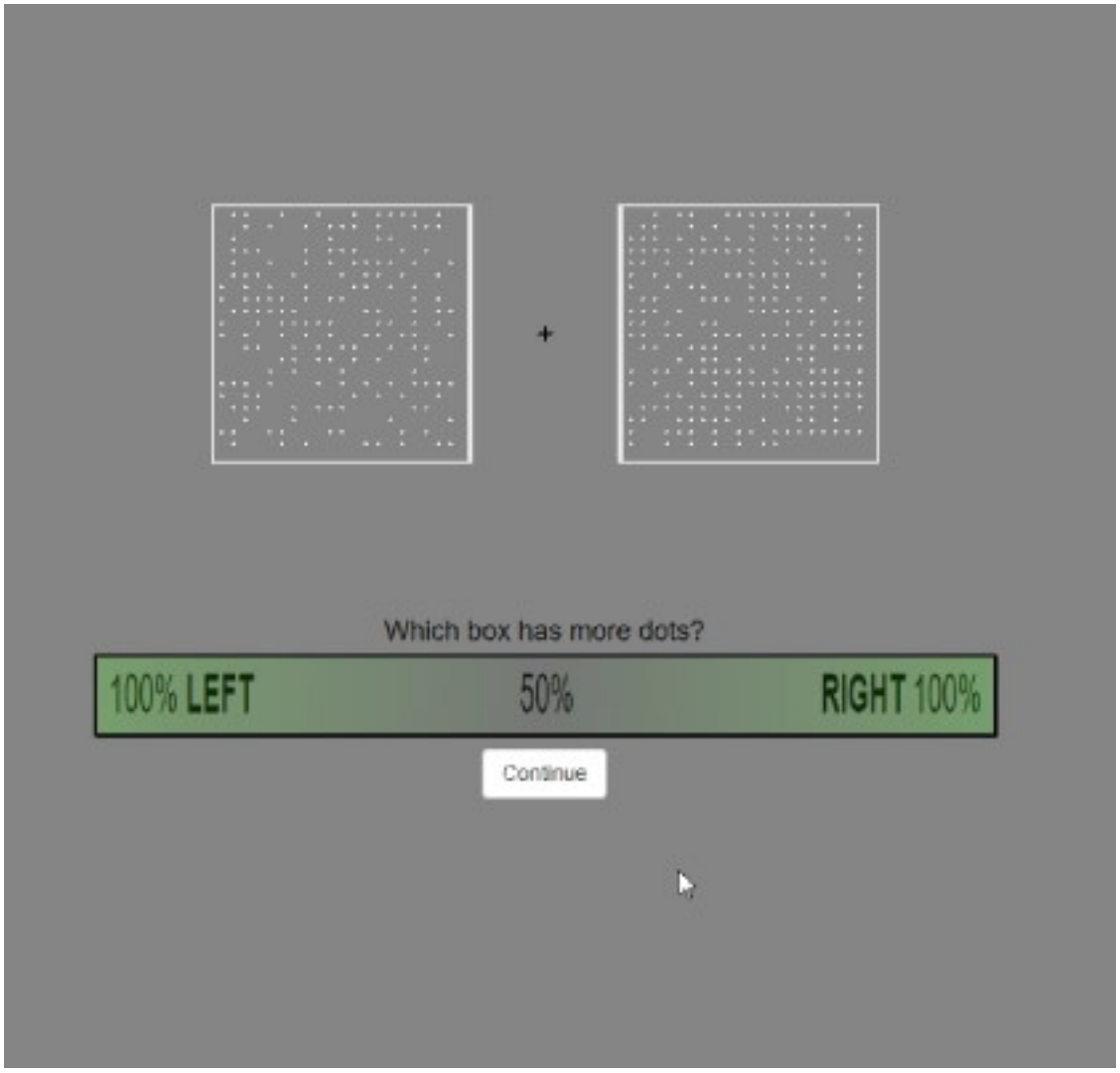


Figure 3.3: Dots Task stimulus.

specific accuracy for each individual participant, allowing confidence to be examined in the absence of confounds with the probability of being correct. Stimuli were continually adjusted throughout the experiment to maintain an initial estimate accuracy of around 72% using a 2-down-1-up staircase procedure. There were a substantial number !TODO[how many?] of trials in the practice block so that participants could eliminate practice effects and thus experience a more stable objective difficulty during the core trial blocks. After each block participants were told what percentage of the final decisions they had provided were correct and allowed to take a short, self-paced break.

Specific limitations The Dots task uses a perceptual decision with a high number of trials. This structure makes it plausible that participants respond to the advice primarily through simple reinforcement learning rather than through specific social processes thought to be at work in advice-taking and advisor evaluation.

Whether or not the Dots task taps into social processes, both the task itself and the experimental structure are very different from most advice-taking and advisor evaluation in the real world. Perceptual decisions are rarely the subject of advice, and thus the central task is an unusual one for joint decision-making. The amount of exposure to advisors also greatly exceeds that which would be obtained over a far longer period of time in most real world situations. This much greater level of exposure risks investing effects with artificial importance: while it is a strength of experimental designs to magnify the effects they aim to study we must not let such magnification blind us to the real relevance of these effects within the complex and dynamic context of real life.

Estimation (Dates Task)

Rationale

Continuous

Binary

Development !TODO[In some amount of detail TBD describe the process of coming up with questions (trawling general knowledge sites, eventually settling on a few timelines inc. Wikipedia, Oxford thingy; questions 1850-1950; rounds of getting participant answers + conf intervals; questions 1900-2000; answers + conf intervals)]

Specific limitations !TODO[Some discussion of limitations of this method: no control over participant error rates, difficult task, inconsistent presentation for an individual (e.g. solid knowledge of an obscure date)]

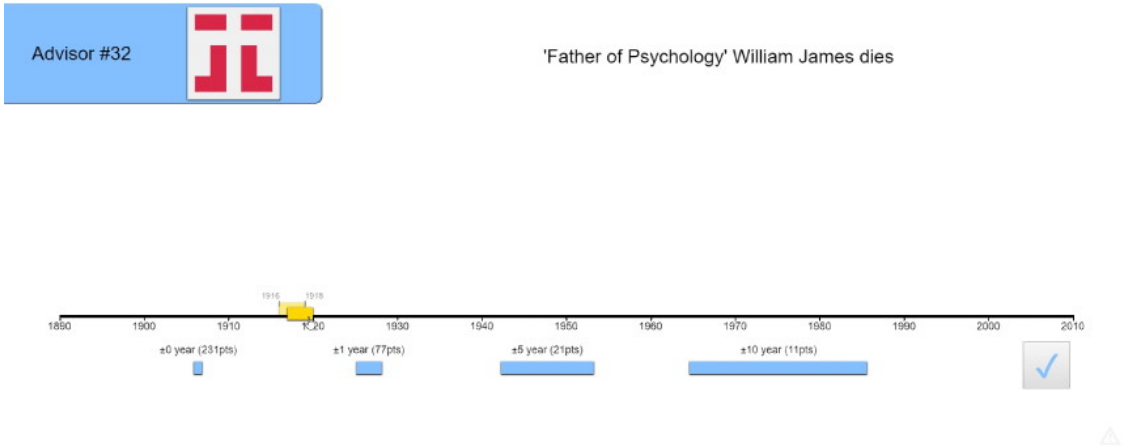


Figure 3.4: Dates Task with continuous responses.

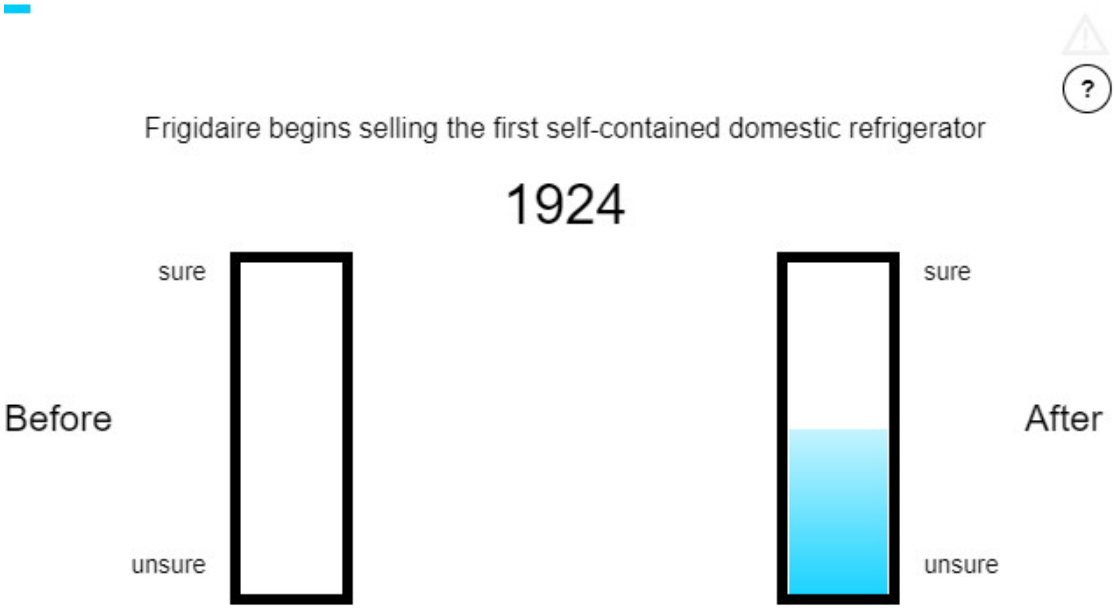


Figure 3.5: Dates Task with binary responses.

General limitations

There are several limitations common to both task designs. The most obvious limitation is that the advisors are not organically-interacting humans. There are other ecological validity limitations in the presentation of advice, the structure of the experiment, and the absence of other cues.

!TODO[Other limitations that aren't wider ecological validity ones?]

The use of artificial advisors means that advice can be carefully specified, and the experiments can be run easily, cheaply, and quickly. Transparently artificial advisors may, however, limit the generalisability of the experimental results in two ways. Firstly, if different integration processes exist for social and non-social information, it is plausible that, for at least some participants, the advice information is perceived as non-social information. While social and non-social information processing would not invalidate any findings (because such a factor would be unlikely to be systematically related to manipulations of interest), they may harm the ability of experimental results to inform us about the processes by which social information is integrated. Secondly, artificial advisors may not trigger a number of human-centred processes such as equality bias (Mahmoodi et al. 2015), meaning that effects revealed in these experiments may be much more difficult to observe in real human advice exchanges.

The advice presented to participants in the experiments is specific and impersonal. During real-life advice-taking, advice is often provided within a discussion, with estimates accompanied by reasons and points responded interactively. Although studies have indicated that advice-taking behaviour remains similar where discussion is allowed (Liberman et al. 2012; Minson et al. 2011), these experiments placed discussion within the context of advice exchanges over a decision made by both dyad members individually, and not with distinct roles for the advisor and judge as used in these experiments.

Further ecological validity limitations arise from the structure of the experiment. The task presents a series of trials sequentially, with a rapid procession through each. This structure is intended to condense a real-world relationship with an

Table 3.1: Confidence contingent agreement advisor advice profiles

	Initial decision confidence	Probability of agreement (%)	
		Bias Sharing	Anti Bias
Participant correct	High (top 30%)	80	60
	Medium (middle 40%)	70	70
	Low (bottom 30%)	60	80
Participant incorrect	Any	30	30
Total agreement	Participant correct	70	70
	Participant incorrect	30	30

advisor, built up over repeated interactions over time, into as narrow a time window as possible. It is likely that this temporal compression does not fundamentally alter the processes of advisor evaluation and trust formation, but we have little positive evidence to support this supposition.

The final major difference between real-life advice and the advice offered in these experiments is in the richness of the relationship between advisor and judge. In a real-life relationship there would be numerous other factors at play which may alter or overwhelm any advice-taking and advisor evaluation processes revealed by these experiments. How well a person gets along with another could matter beyond simply increasing the perceived benevolence of the other, for example.

3.1.4 Advisor advice profiles

!TODO[Adjust this to talk more precisely about what agreement means in different contexts, and be a more general introduction to the idea of agreement, accuracy, and their relationship.]

The advisers are virtual agents whose probability of agreeing with the participant's decision varies as a function of the participant's confidence and correctness in the initial decision phase. Table 3.1 illustrates how this relationship functions, and shows that the overall correctness and agreement rates of the advisers is equivalent overall. Importantly, on largest minority of trials, the middle 40%, the advisers are exactly equivalent, meaning these trials can be compared directly without confounds arising from agreement rate and initial confidence.

3.1.5 Analysis

Dependent variables

Pick rate Pick rate provides a measure of source selection behaviour. In most experiments there are some trials that offer participants a choice of which advisor they would like to hear from. There are always two choices, and a choice must always be made. The two choices are consistent within the experiment. Pick rate is the proportion of choice trials in which a specified advisor was chosen.

A participant's pick rate is an aggregate over a number of trials, and expresses the observed probability of picking the specified advisor. The mentally represented preference for that advisor is not measured directly (if such a thing even exists), and cannot be determined from the observed pick rate without knowing the mapping function for each individual participant. Mapping functions (such as softmax !TODO[cite something on this]) produce stochastic choice behaviour from a preference marked on a continuous scale, and are typically sigmoid. The relationship between preference and pick rate is non-linear and idiosyncratic, but it is likely monotonic for all participants: the stronger the preference the higher the pick rate. Despite this monotonic relationship, it is important not to infer that one participant's preference for an advisor is stronger than another's on the basis of the former picking that advisor more frequently: it may be instead that, for the former participant, smaller differences in preference lead to more consistent picking behaviour.

Weight on Advice Weight on Advice, and its compliment Advice Taking, are commonly used to quantify the relative contributions of advice and initial estimates in making final decisions. It is obtained by dividing the amount an initial estimate was updated by the amount the advisor recommended adjusting the initial estimate. It thus expresses the amount the estimate changed as a proportion of the advised change.

Formally, Weight on Advice is given by $(e^F - e^I)/(a - e^I)$ where e^I and e^F are the initial estimate and final decision, respectively, and a is the advice.

Where the final answer moves away from advice (i.e. the final decision is further from the advice than the initial estimate), the value of Weight on Advice is negative, and where the adjustment towards advice exceeds the advice itself (i.e. the advice falls between the final decision and the initial estimate) the value of Weight on Advice is greater than 1. These values are typically truncated to 0 and 1, respectively.

In cases where the advice is exactly equal to the initial estimate, the denominator is equal to zero and the value for Weight on Advice is thus undefined. Trials in which the advice is exactly equal to the initial estimate are consequently discarded when calculating Weight on Advice.

Influence

Capped influence Influence, the dependant variable in some analyses, is calculated as the extent to which the judge's initial decision is revised in the direction of the advisor's advice. The initial (C_1) and final (C_2) decisions are made on a scale stretching from -55 to +55 with zero excluded, where values <0 indicate a 'left' decision and values >0 indicate a 'right' decision, and greater magnitudes indicate increased confidence. Influence (I) is given for agreement trials by the shift towards the advice:

$$I|\text{agree} = f(C_1)\begin{cases} C_2 - C_1 & C_1 > 0 \\ -C_2 + C_1 & C_1 < 0 \end{cases} \tag{3.1}$$

And by the inverse of this for disagreement trials:

$$I|\text{disagree} = -I|\text{agree} \tag{3.2}$$

The confidence scale excludes 0, and thus the final decision can always be more extreme when moving against the direction of the initial answer than when moving further in the direction of the initial answer. A capped measure of influence was used to minimise biases arising from the natural asymmetry of the scale. This

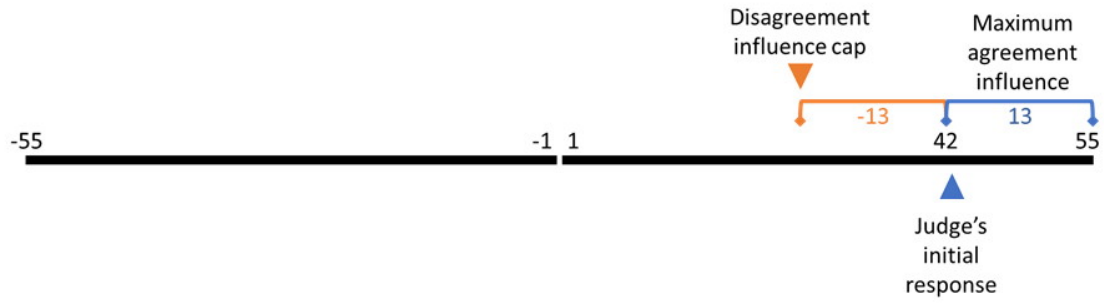


Figure 3.6: Capping influence to avoid scale bias.
In this example the judge's initial response is 42, meaning that their final decision could be up to 13 points more confident or up to 97 points less confident. Any final decision which is more than 13 points less confident is therefore capped at 13 points less confident.

measure was calculated by truncating absolute influence values which were greater than the maximum influence which could have obtained had the final decision been a maximal response in the direction of the initial answer (Figure 3.6).

The capped influence measure I_{capped} is obtained by:

$$I_{\text{capped}} = f(C_1) \begin{cases} \min(I, 2C_1 - S_{\max}) & C_1 > 0 \\ \max(I, 2C_1 + S_{\max}) & C_1 < 0 \end{cases} \tag{3.3}$$

!TODO[Check this equation and explain its terms]

Statistics

Statistical analyses are conducted using a both Frequentist and Bayesian statistics.

Frequentist statistics Frequentist statistics, often in psychology referred to simply as 'statistics', are a family of statistical tests which minimise the long-term error rates rates of the conclusions they invite. To perform one of these tests, the user first specifies a null hypothesis which the test will seek to reject. Next, the user selects a rate of false-positives that is acceptable, termed an alpha (α). In psychology α is almost universally .05, meaning that 5% of results for tests where there is no true effect will be positives. The principle statistic of interest from these tests is the p-value: the chance that a sample equivalent to the user's sample would be as extreme or more extreme than the observed sample *assuming*

that the null hypothesis is true. The p-value quantifies how expected the sample observations are where the null hypothesis is the model: its complement quantifies how surprising the data are given the null hypothesis is true. If the p-value is lower than the α the null hypothesis may be rejected on the reasoning that such an unlikely sample as the current one is more consistent with some other hypothesis. Where the null hypothesis is rejected the sample is labelled as being *significantly different* from that expected from the null model, and this general feature of the difference is known as statistical significance.

It is common practice when using frequentist statistics to set up experiments such that two, and only two hypotheses can possibly explain the data. One of these is deemed the null hypothesis and tested as described above. The other hypothesis is termed the 'alternate' hypothesis, and, if the null hypothesis is rejected, is accepted in place of the null hypothesis because there were only two possible explanations and the null hypothesis has been ruled out.

Frequentist statistics are useful, better understood by psychologists than alternatives, and capable of delivering genuine insights when used correctly. A great many caveats surround their use and interpretation, however. Frequentist statistics cannot express positive evidence for the null hypothesis (because the null hypothesis is an assumption of the tests): p-values are uniformly distributed among samples drawn from the null hypothesis (**murdochPValuesAreRandom2008**) and so no argument can be made from any particular value that it constitutes more or less evidence for the null. Likewise, p-values below the chosen α cannot strictly be interpreted as expressing more or less evidence for the alternate hypothesis (or against the null hypothesis). Frequentist statistics control long-term error rates rather than capturing relative likelihoods of theories, and thus the only legitimate inferences from a frequentist test are to accept or reject the null hypothesis. The uniform distribution of p-values under the null also means that repeated sampling from the null distribution without adjusting the α will increase the rate at which false-positive conclusions are drawn. This repeated sampling can happen in many ways such as running related tests on the same data, using slightly different data

(e.g. by adjusting exclusions), and running tests at multiple time points in data collection (especially if a significant result terminates data collection).

In this thesis a range of frequentist statistical tests are used, most frequently t-tests and analyses of variance (ANOVA). The α is always set at .05 unless otherwise stated. Null hypotheses are always the expected distribution if the effect being tested is nil. Where the level of influence exerted by two different advisors is studied, for example, the null hypothesis would be that there were no systematic differences in influence exerted by those advisors.

Bayesian statistics Bayesian statistics consist in a statistical approach that attempts to capture the (posterior) likelihood of a hypothesis being true on the basis of its (prior) plausibility and the strength of the evidence. These tests are frequently adapted to quantify the relative likelihood of two competing hypotheses given their relative prior plausibilities and how consistent the observed evidence is with each theory. To perform these Bayesian statistical tests the user specifies the hypotheses to be compared, their relative likelihoods, and the evidence sampled.

The principle statistic in Bayesian tests is the Bayes Factor (BF). BF quantifies the posterior likelihood of one hypothesis over another. In the notation used here, BF always quantifies a more complex model over a simpler one: in the case of a Bayesian t-test it therefore quantifies the alternate hypothesis (in which there is an effect) over the null hypothesis (in which there is not an effect). The BF takes a value between 0 and infinity. Values below 1 indicate a greater likelihood for the simpler model, and that greater likelihood is given by 1/BF. Where BFs below 1 are reported in this manuscript the notation 1/BF is used to allow an intuitive reading of the strength of evidence in favour of the simpler model.

Bayesian tests produce a continuous measure of relative likelihood which simply describes the data and prior beliefs. In order to draw categorical inferences, thresholds are placed on this continuous outcome. Here these thresholds are $1/3 < \text{BF} < 3$, meaning that a BF of less than 1/3 constitutes evidence in favour of the simpler model while a BF greater than 3 constitutes evidence in favour

of the more complex model. These values are those suggested by !TODO[CITE whoever has that neat little table. Jeffreys?] as representing !TODO[which word does the paper use?] evidence in a given direction. Where the BF lies between these thresholds it is labelled as 'uninformative'. An uninformative result supports neither the simpler nor the more complex model, and indicates that the data are insufficient to distinguish the hypotheses.

The Bayesian tests used here rely on the priors specified by the BayesFactor R package !TODO[CITE BayesFactor]. These priors govern the expected distributions of observed differences between samples where there is or is not a genuine effect creating systematic differences. The use of the same, weakly-informative priors for all tests means the approach used here is an 'objective Bayesian' approach. This objective Bayesian approach can be contrasted with a 'subjective Bayesian' approach in which the goal is to specify the exact amount of belief one should have in one hypothesis over another. Neither the objective nor subjective approach is clearly superior. The subjective approach is used here because it is simpler. There is some risk that results will be a poor fit to reality because the priors are inappropriate, but this risk is fairly low and somewhat mitigated by the additional inclusion of frequentist statistics.

Integrating statistical results In most cases, Bayesian and frequentist statistics produce the same conclusion !TODO[pretty sure there's a paper which states/demonstrates this]. Where this is not the case, results should be interpreted very cautiously: a significant frequentist test with an uninformative or null-favouring Bayesian test can indicate that the result may be a false-positive, while clear Bayesian support for the alternate hypothesis in the absence of a significant frequentist test can indicate that the priors in the Bayesian test are inappropriate.

Where null conclusions are to be drawn, i.e. the null hypothesis is to be retained, only Bayesian statistics can be considered informative. In these cases Bayesian statistics will be interpreted, with the caveat that the safeguard of using two independent approaches to draw statistical conclusions has lapsed.

Software Data analysis was performed using R (**rcoreteamLanguageEnvironmentStatistical2**
For a full list of packages and software environment information, see !TODO[figure out where to include this stuff. Appendix? Also link to a containerized version of this.]

3.2 Open science approach

3.2.1 Open science

Nullius in verba ("take nobody's word for it") is written in stone above the entrance to the Royal Society's library. This fundamental principle of science, that it proceeds on evidence rather than assertion, has frequently been forgotten in practice. Concerns about sloppy, self-deluding, or outright fradulent science have existed since at least the time of Bacon. The modern open science movement in psychology dates from the early 2010s. Simmons et al. demonstrated how easily false positive results could emerge from unconstrained researcher degrees of freedom in analysis (**simmonsFalsePositivePsychologyUndisclosed2011**), Nosek and colleagues published a roadmap for improving the structure and function of academic research and publishing (**nosekScientificUtopiaOpening2012**; **nosekScientificUtopiaII2012**), and the Open Science Collaboration began (**collaborationEstima**
In the years following, a deluge of papers, movements, and practical changes have emerged. The meaning of open science varies within each sub-discipline, and this section outlines how the experiments comprising this thesis have been conducted in a reproducible and transparent manner.

3.2.2 Badges

Following the Center for Open Science (`https://cos.io`), this thesis uses a series of badges to indicate adherence to particular aspects of open science. Three badges, *preregistration*, *open materials*, and *open data*, are adopted directly from the Centre and used according to the Centre's rules (`https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/`). Studies which qualify for a badge will have the badge

displayed immediately below their title. Each badge will contain a link to online resources which provide the content for which the badge is awarded.

Preregistration

Preregistration of a study means that information about the study has been solidified prior to the analysis of the data. This means that hypotheses cannot be changed to represent unanticipated or overly-specific findings as a priori predicted (**kerrHARKingHypothesizingResults1998**). In practice in this thesis, preregistration means describing in detail the design and analysis plan for an experiment and depositing the description with a reputable organisation prior to data being collected. The links which accompany the preregistration badge will point to the preregistration document. These measures help to prevent presenting a highly selected and biased interpretation of the data as the result of a natural analytical process.

The preregistration badge also appears within results sections to designate those statistical investigations which were included in the preregistration. Some analyses are exploratory. These exploratory analyses are not included in the preregistration, because they are inspired by the data themselves. They are reported after the preregistered analyses, or are clearly designated as exploratory in the text.

Open materials

A foundational principle of science is that findings can be reproduced by other people. Open materials facilitate reproduction by making it easier to rerun an experiment. Open materials also increase the likelihood that errors can be identified. In the case of the behavioural experiments reported here, the open materials include computer code necessary to run the experiment. The links accompanying the open materials badge points to this code.

Open data

Theories are the output of science as a whole, but data are the output of any individual study. Sharing data directly allows other scientists to check and extend

the data analysis conducted, to reuse the data in meta-analyses, and to repurpose the data for other investigations. This increases the robustness of the results, and increases the efficiency of science as a whole. The links accompanying the open data badge point to online storage where the data can be obtained for a study, along with appropriate metadata.

3.2.3 Thesis workflow

This thesis is written in RMarkdown, with the data fetched and analysed at the time the document is produced using the publically available pipeline - the entire document can be reproduced locally using the source code in an appropriate environment. A Docker environment copying the environment used to produce this document is available at !TODO[the containerisation thing]

4

Psychology of advice-taking

4.1 !TODO[Cite Niccolo's best paper on his thesis advice-use experiments]

!TODO[Describe Niccolo's stuff, especially why accuracy vs agreement might be a good target for replication with a new experimental design.]

4.2 Extending results to the Dates task

4.2.1 Introduction

!TODO[Why the dates task? Generalising to a new kind of task. Quicker, more realistic in some ways, more fun]

!TODO[Why agreeing vs accurate advisors? It's a distillation of the core insight from Niccolo's work.]

Open scholarship practices


`https://osf.io/fgmdw`


!TODO[OSFify data for these studies]

https://github.com/oxacclab/ExploringSocialMetacognition/blob/f90b6f9266a901211a4ddb7b5ee1de1c74e8df57/ACv2/index.html

Unanalysed data This branch of experiments was the core of the date task. As such, most experimentation and piloting happened on this branch, meaning that there were many versions of the study where data were collected for which analysis is not included here (v0-0-8, v0-0-9, v0-0-10, v0-0-11, v0-0-12, v0-0-13, v0-0-14, v0-0-15, v0-0-16, v0-0-17, v0-0-18, v0-0-20, v0-0-21, v0-0-22, v0-1-22, v1-0-0, v1-1-0). !TODO[Explain the non-analysed versions.] Overall, data were collected from 262 participants across all of these versions.

The study reported here is a preregistered replication of a pilot study (v0-0-21) which produced the same results. Data, analysis script, and analysis for the pilot study are also available at !TODO[OSFify pilot data/analysis stuff].

Method

This study used the continuous version of the Dates Task (§Estimation (Dates Task) - Continuous).

Results

Exclusions Participants' data could be excluded from analysis where they fail attention checks, fail to complete the entire experiment, or have more than 2 outlying trials. Outlying trials are calculated after excluding participants who failed to complete the experiment, and are defined as trials for which the total trial time was greater than 3 standard deviations away from the mean of all trials from all participants.

A browser compatibility issue in this study meant that any participants completing the study using the Safari family of browsers had to be excluded because the advice was not presented appropriately.

Task performance

Table 4.1: Participant exclusions for Dates task advice influence experiment

Reason	Participants excluded
Attention check	NA
Unfinished	2
Too many outlying trials	1
Missing offbrand trial data	1
Total excluded	**NA**
Total remaining	**29**

Table 4.2: Bayesian ANOVA for Dates task advice influence experiment

M1	M2	BF.M1.M2
Advisor	Intercept only	1/1.13
Feedback	Intercept only	1/1.37
Advisor + Feedback + Advisor:Feedback	Advisor + Feedback	3.40

Mainpulation checks

Hypothesis test

```
## Warning: Problem with `mutate()` input `bf`.
## i data coerced from tibble to data frame
## i Input `bf` is `map(...)`.
## Warning: data coerced from tibble to data frame
## Warning: Problem with `mutate()` input `bf`.
## i data coerced from tibble to data frame
## i Input `bf` is `map(...)`.
## Warning: data coerced from tibble to data frame
## Warning: Data is unbalanced (unequal N per group). Make sure you specified a
## well-considered value for the type argument to ezANOVA().
```

Table 4.3: Frequentist ANOVA for Dates task advice influence experiment

Effect	$F(1, 27)$	p		η^2
Feedback	2.02	.167		.048
Advisor	5.10	.032	*	.057
Feedback:Advisor	9.72	.004	*	.103

Degrees of freedom: 1, 27

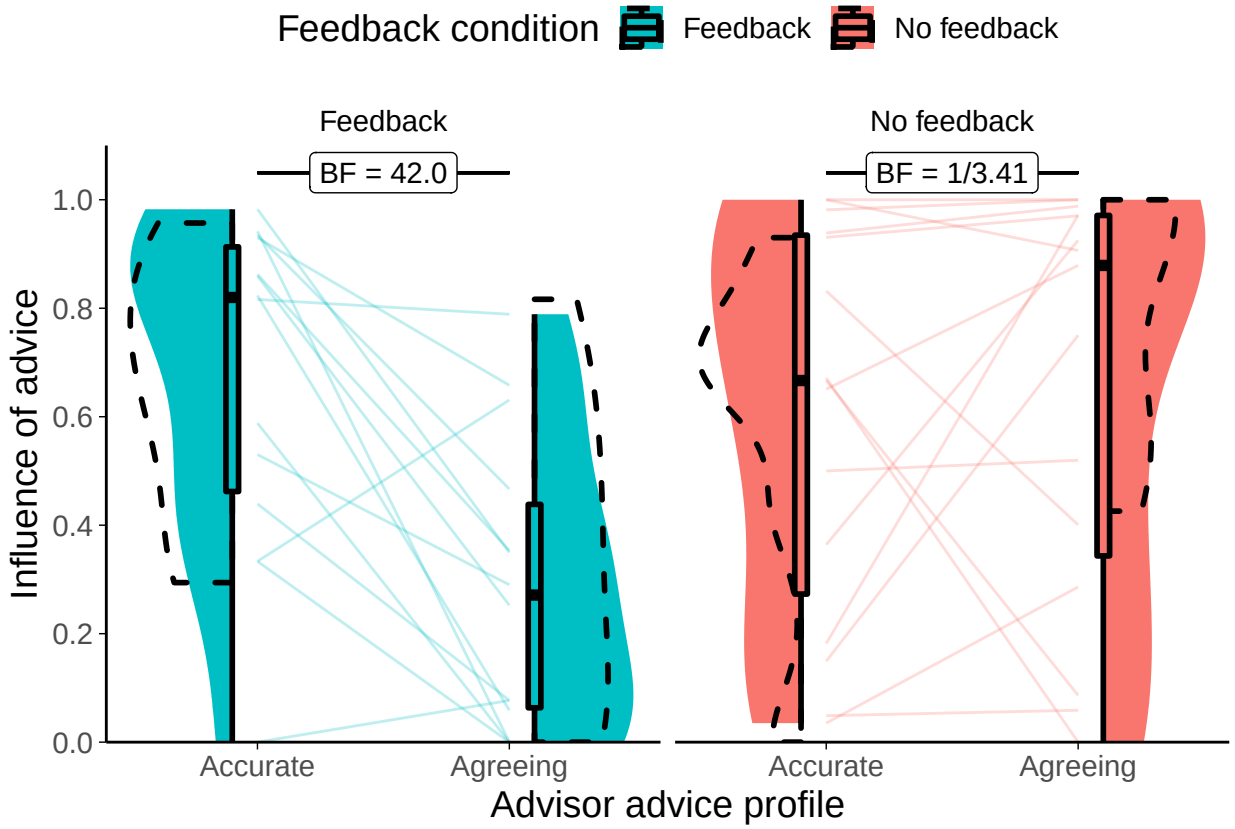


Figure 4.1: Dates task advisor influence for high accuracy/agreement advisors. Shows the influence of the advice of the advisors. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The dashed outline shows the distribution of participant means in the original study of which this is a replication.

There were systematic differences in the influence of advice on the key trials where the advice itself was balanced between advisors. Frequentist (Table 4.3) and Bayesian (Table 4.2) ANOVA analyses both indicated an interaction between the Advice profile of an advisor and the Feedback condition of the participant. The frequentist ANOVA indicated a main effect of Advice profile, although there was no reliable evidence either way for this effect in the Bayesian test.

Within each condition, T-tests revealed that the Accurate advisor was more influential than the Agreeing advisor in the Feedback condition ($t(13) = 4.27$, $p < .001$, $d = 1.34$, BF = 42.0; $M_{Accurate|Feedback} = 0.67$ [0.50, 0.84], $M_{Agreeing|Feedback} = 0.29$ [0.13, 0.44]). The advisors were equivalently influential in the No feedback

condition ($t(14)$ = -0.50, p = .626, d = 0.14, BF = 1/3.41; $M_{Accurate|\text{ň}Feedback}$ = 0.60 [0.40, 0.80], $M_{Agreeing|\text{ň}Feedback}$ = 0.65 [0.44, 0.86]).

Questionnaires

4.2.2 Discussion

The experiments show that people can learn to disregard advice which provides support but no information, provided that feedback is available. Where feedback is unavailable, people do not appear to distinguish between useful and supportive advice. These results are not wholly consistent with an account of advice-taking in which people use agreement to evaluate advisors in the absence of feedback. Under such an account, we would expect participants in the No feedback condition to have shown a greater susceptibility to advice from the Agreeing advisor, but this did not happen. The equivalence of the influence of the advisors in the No feedback condition may be a consequence of relatively high levels of advice influence overall, producing a ceiling effect (the numerical advantage for the Agreeing advisor is as predicted by the theory). The relatively high levels of advice influence are a feature of the Dates task, a consequence of the difficulty of the questions for most participants. Another possible explanation for the equivalence of influence between the advisors is that participants may not have had enough exposure to the advisors to properly learn about the value of their advice. The limitation on exposure is another feature of the Dates task: while the Dots task provides participants with many tens of trials in which to update their assessment of advisors, the Dates task provides a level of exposure more similar to normal social interaction (although not necessarily very similar).

While not wholly consistent with the agreement-as-proxy account, the results were broadly consistent with previous work: participants systematically took more advice from the more accurate advisor. Is that a fair summary? Seems only true in the feedback condition. Does that mean it really is consistent? The distinction between agreeing and accurate advice revealed by Niccolo and Nick - what distinction did they actually make? generalises to a substantially different task which uses a

different kind of decision and different level of exposure to advisors' advice. In the experiments which follow (§Psychology of source selection), both the Dots and Dates tasks are used to investigate differences in source selection behaviour.

The design of this experiment deliberately violated an assumption which may be generally true in real life, that the advice is sufficiently independent as to convey at least some information regarding the correct answer. Had we told participants that the agreeing advisor would agree with them no matter what they the participants said, the participants may have disregarded the advice. Nevertheless, the results show that, even where they would have performed objectively better by preferring accurate over agreeing advice, participants were not able to detect the more accurate advice without objective feedback.

4.3 Confidence-contingent advice

 !TODO[OSFify data for these studies]

 `https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACv2/index.html`

ACv2/withConfidence_coreAnalysis_v0.0.1

!TODO[This experiment is kinda crappy and failed its manipulation. Can we run a version which does its job properly and provide actual evidence for/against the confidence modulation?]

4.3.1 Method

4.3.2 Results

4.3.3 Discussion

Participants appeared to pay more attention to the advice than the advisor. In other words, participants distinguished between individual pieces of advice but did not translate these distinctions into distinctions between advisors. This study

thus provided no evidence in favour of the confidence-weighting adjustement to the agreement model.

5

Psychology of source selection

The model of advisor evaluation described earlier (§2.0.3) requires empirical support. The model of advisor evaluation with feedback is well supported by data which indicate that, given objective feedback, people can use the feedback to learn about the trustworthiness of advisors. The extent to which advice is taken (§3.1.5) is commonly used as a measure of a participant's trust in an advisor, on the argument that the participant seeks to maximise task performance and task performance is maximised by taking more advice from more trustworthy advisors. !TODO[Mini lit review - advisor accuracy training with feedback (Yaniv?), Niccolo and Nick's feedback stuff]. Overall, this means that people prefer more accurate advice over less accurate advice.

When objective feedback is unavailable, people can still demonstrate a greater dependence upon advice from more as opposed to less accurate advisors. !TODO[Mini lit review, Niccolo's stuff, maybe prior stuff - check Niccolo's thesis for references]. This is a consequence of agreement: where the base probability of being correct is greater than chance, the independent estimates of people who are more accurate will agree more often (leading to 100% agreement on the correct answer for two independent decision-makers of perfect accuracy). In the absence of feedback, therefore, agreement can be used as a proxy for accuracy, as formalised in the model.

The role of agreement is demonstrated clearly in experiments where the objective accuracy of advisors is balanced, but the agreement rates of the advisors is varied. !TODO[cite Niccolo CITE]. Pescetelli and Yeung demonstrated that advice is more influential from advisors who tend to agree with a participant more frequently when objective feedback is not provided. This is despite the fact that advice is more influential when it disagrees with the participant's initial estimate. [1] These data suggest that people may be using agreement as a proxy for accuracy, although they may simply prefer agreement over disagreement when there is no accuracy cost to be paid. I report the results of an experiment in which an agreeing advisor was compared with an accurate advisor under conditions of feedback or no feedback. Results indicated that, as predicted by the models, participants preferred the accurate advisor when feedback was provided and the agreeing advisor when feedback was withheld.

Pescetelli and Yeung developed a more sophisticated model of advisor evaluation in which the increase in trust gained when an advisor agreed with the decision-maker was contingent upon the confidence of the decision-maker's initial estimate (§2.0.3). Intuitively, if I am highly certain that I am correct on a given question, an advisor who disagrees with me is likely to be incorrect, whereas one who agrees with me is likely to be correct. Provided confidence is indicative of the objective probability of being correct, as confidence in the initial decision increases it more closely approximates objective feedback for the purposes of evaluating advice. !TODO[detailed account of Niccolo's evidence for the confidence-weighted model]. I report the results of experiments designed to extend Pescetelli and Yeung's results to the domain of advisor influence, using two different decision-making tasks.

5.1 Accuracy

!TODO[check Niccolo covered this] {#ac-acc}

[1] This is partly due to the nature of the judge-advisor system: there is always room for disagreement to be more extreme than agreement, because agreement is lower-bounded by the participant's initial estimate.

Pescetelli and Yeung !TODO[cite new paper] demonstrated that more accurate advisors are more influential (regardless of the presence of feedback) in a lab-based perceptual decision-making task. We attempted to extend this finding to the domain of advisor selection in two online tasks: a 'Dots Task' requiring similar perceptual decision-making to the task used by Pescetelli and Yeung, and an estimation-based 'Dates Task'.

The ability to distinguish between accurate advisors in these experiments is important because they relate directly to the phenomenon we are attempting to explain: rational advice-seeking behaviour in the absence of feedback.

5.1.1 Dots Task

Open scholarship practices

 `https://osf.io/u5hgj`

 !TODO[OSFify data for these studies]

 `https://github.com/oxacclab/ExploringSocialMetacognition/blob/9932543c62b00bd96ef7ddb3439e6c2d5bdb99ce/AdvisorChoice/index.html`

Unanalysed data Several early versions of this experiment were run where bugs in the experiment code made the results unreliable. The earliest versions contained a bug where advisors instructed to agree with a participant instead provided advice identifying the correct answer. Other versions had a bug in the staircasing code used to titrate the difficulty of the task was converging on too high a value (74% initial decision accuracy as opposed to 71%). Once the staircasing bug was fixed, two more experiments were run, one with 60 practice trials in which participants did not quite reach the desired accuracy before the beginning of the main experiment, and one with 120 practice trials which constitutes the data analysed below. Overall, 0 participants' data was collected in these excluded versions and not included in analysis. While not useful for the hypothesis of this experiment, the excluded data

Table 5.1: Advisor advice profiles for Dots task Accuracy experiment

	Probability of agreement			
Advisor	Participant correct	Participant incorrect	Overall	Overall accuracy
High accuracy	.800	.200	.626	.800
Low accuracy	.600	.400	.542	.600

Table 5.2: Participant exclusions for Dots task Accuracy experiment

Reason	Participants excluded
Accuracy too low	0
Accuracy too high	0
Missing confidence categories	3
Skewed confidence categories	6
Too many participants	0
Total excluded	**9**
Total remaining	**50**

can be used for analysing responses to advice, provided care is taken with the very early versions to ensure advice is interpreted correctly.

Method

!TODO[clarify any methodological differences from the main methods chapter]

Advice profiles The two advisor profiles (Table 5.1) used in the experiment were High accuracy and Low accuracy. The High accuracy advisor was correct 80% of the time while the Low accuracy advisor was correct 20% of the time. The advisor profiles were not balanced for overall agreement rates.

Results

Exclusions Participants' data could be excluded from analysis where they have an average accuracy below 0.6 or above 0.85, do not have trials in all confidence categories, have fewer than 12 trials in each confidence category, or have completed the experiment after the preregistered amount of data has already been collected. Overall, 9 participants were excluded, with the details shown in Table 5.2.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial and final decisions can be seen in Figure 5.1.

```
## Picking joint bandwidth of 268
```

All participants had similar patterns: initial and final responses were approximately normally distributed, with final responses having a higher variance (because they include the variance for the initial response). Most participants show some trials on which initial or final responses took substantially longer than usual.

!TODO[Perhaps this plot would be better showing individual participants' distributions and box-plots/3SD markers, especially if we want to exclude trials on the basis of taking too long (we don't currently). Perhaps tying final response time to final response start would be better, too, because then initial and final decisions can be more sensibly compared.]

Accuracy Accuracy of initial decisions was controlled by a staircasing procedure which aimed to pin accuracy to 71%. The accuracy of final decisions was free to vary according to the ability of the participant to take advantage of the advice on offer. As Figure 5.2 shows, participants' accuracy scores for initial decisions were close to the target values (partly because participants whose accuracy scores diverged considerably were excluded). Participants tended to improve the accuracy of their responses following advice from High accuracy advisors, while the evidence was unclear as to whether there was any difference in response accuracy

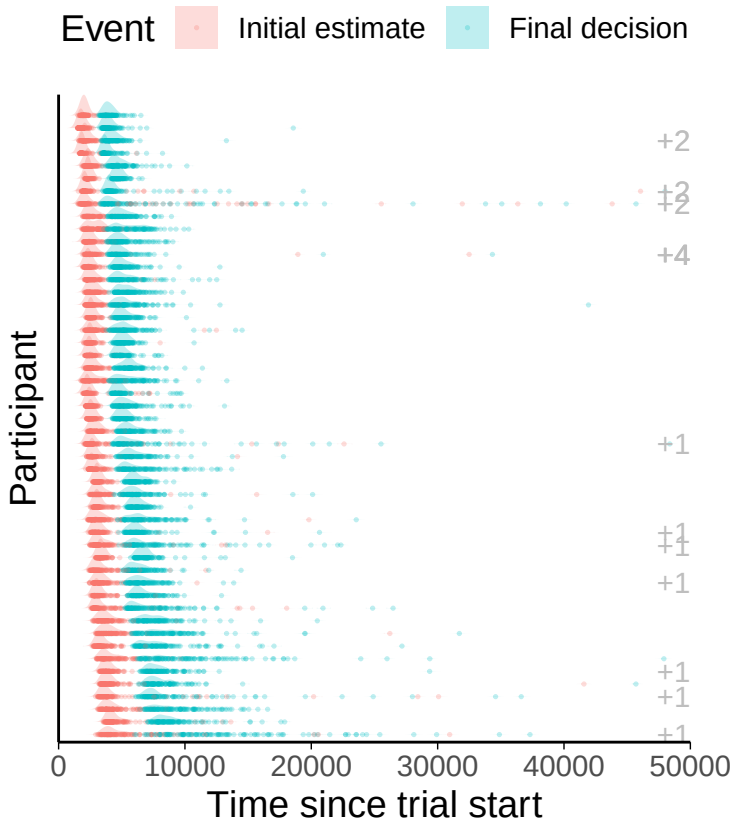


Figure 5.1: Response times for the Dots task with in/accurate advisors.
Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

with Low accuracy advice. The distribution of initial estimate accuracy for trials with the Low accuracy advisor is slightly unusual, with a bimodal structure and a median somewhat higher than the target value. There is no obvious reason why this should be the case.

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions.

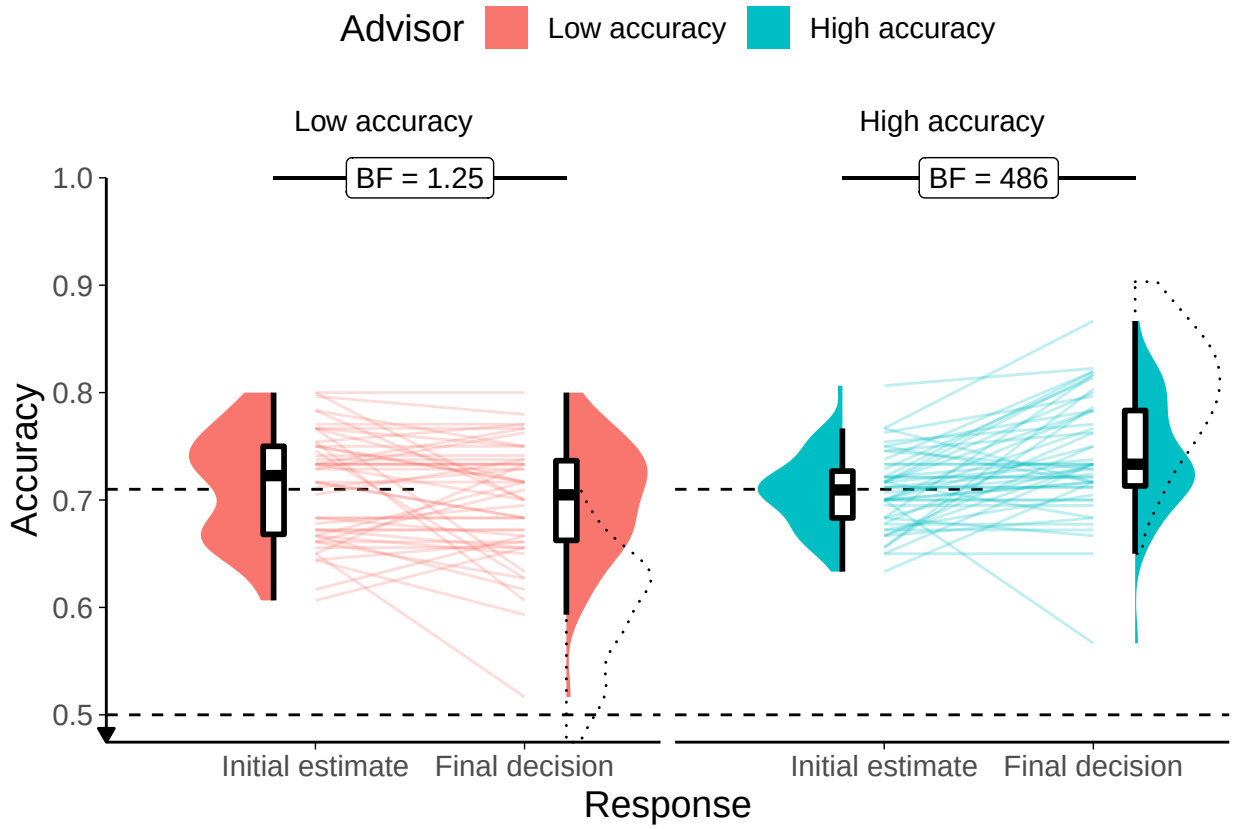


Figure 5.2: Response accuracy for the Dots task with in/accurate advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

Metacognitive ability Where performance on the underlying task is held constant, as here, metacognitive sensitivity can be measured in a bias-free way by plotting Receiver Operating Characteristic (ROC) curves for metacognitive responses (**flemingHowMeasureMetacognition2014**). [ˆThe constant underlying task performance is only true for initial estimates in the paradigm used here, and thus the ROC curves for final decisions should be interpreted with caution because they cannot be proven to be unaffected by metacognitive bias.] ROC curves are obtained by calculating at each of a number of different points on a confidence scale, the probability that the confidence is at least that high for correct versus incorrect answers. The area under the ROC curve gives a measure of the ability of confidence ratings to distinguish correct and incorrect responses. An area under the ROC curve

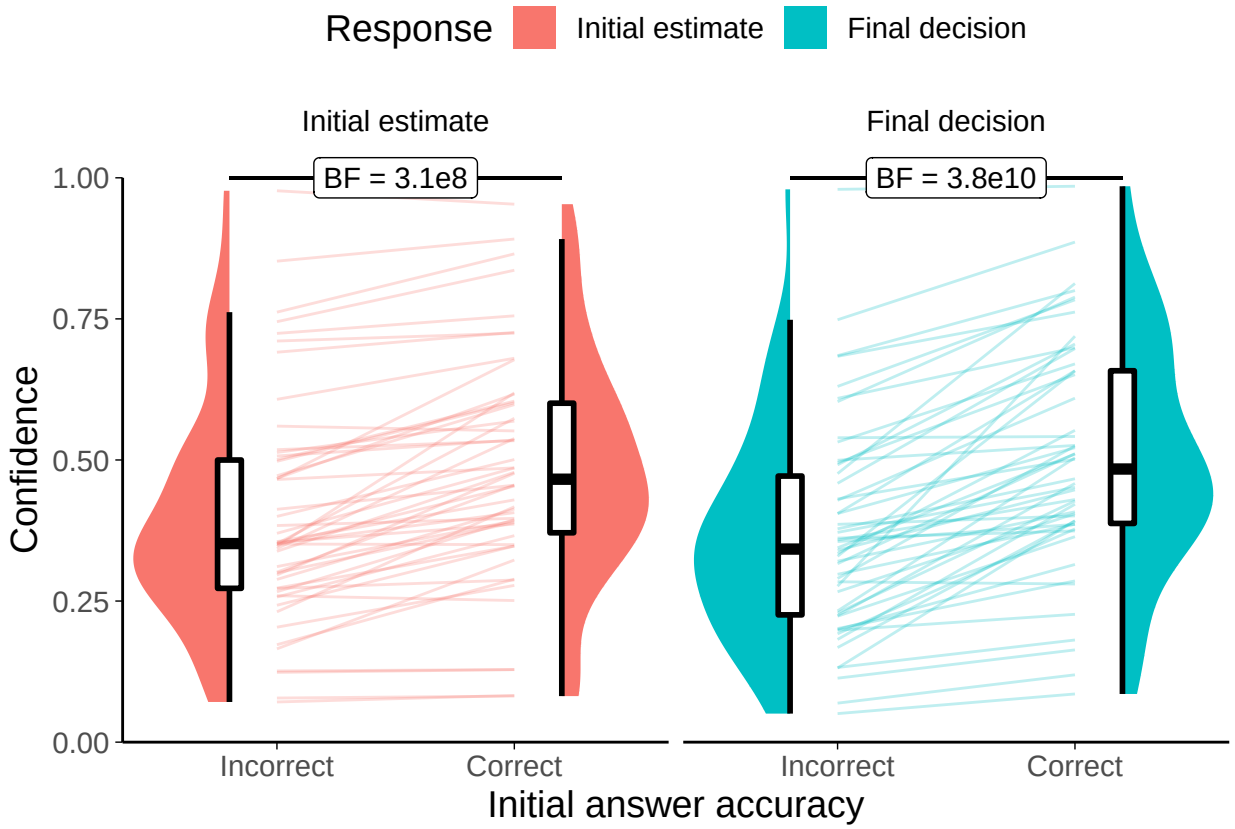


Figure 5.3: Confidence for the Dots task with in/accurate advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions.

of .5 indicates chance performance, and a value of 1 indicates perfect discrimination.

As shown by Figure 5.4, almost all participants showed above-chance metacognitive sensitivity for initial estimates and final decisions. Participants generally showed higher metacognitive sensitivity for final decisions, although this may be an artefact of a change in metacognitive bias. Participants' metacognitive sensitivity was not particularly high !TODO[What are typical values we might expect in the dots task and similar tasks? Is there a useful mapping between meta-d' and Type II ROC to compare with e.g. Roualt's stuff?]. There was no evidence of participants' metacognitive sensitivity being correlated with their task performance (Figure 5.5). This is expected when task performance is tightly controlled, because under these conditions variation in task performance reflects variation in ability within a participant rather than between participants.

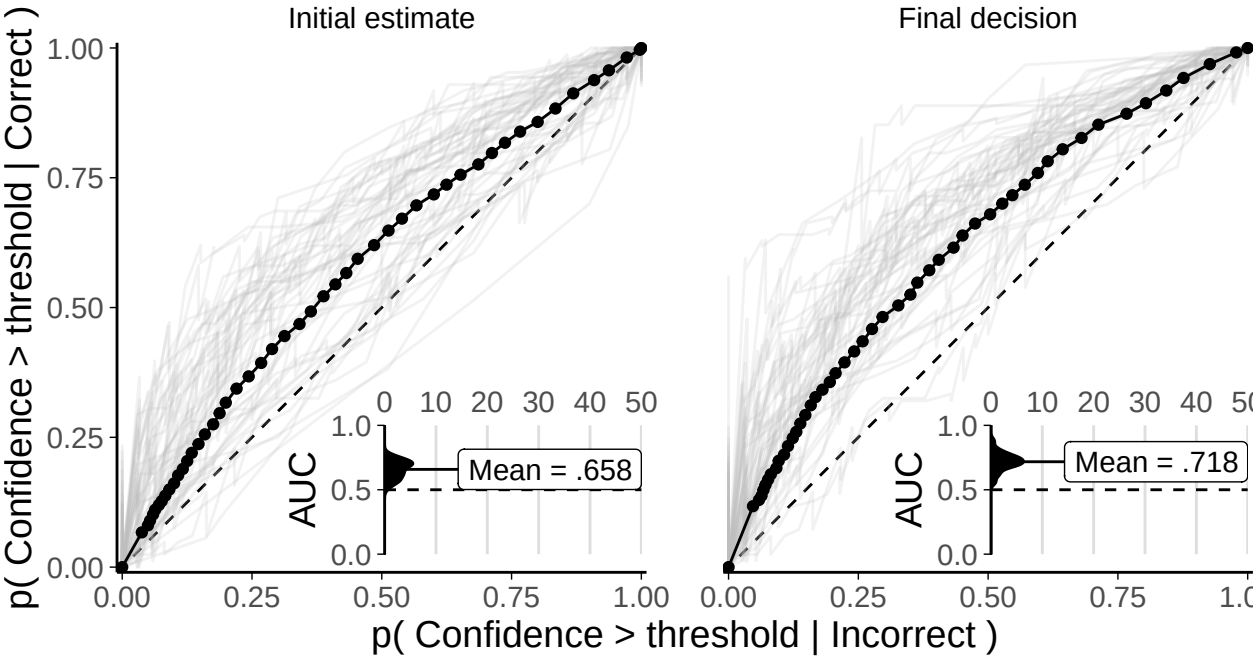


Figure 5.4: ROC curves for the Dots task with in/accurate advisors.
Faint lines show individual participant data, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers.

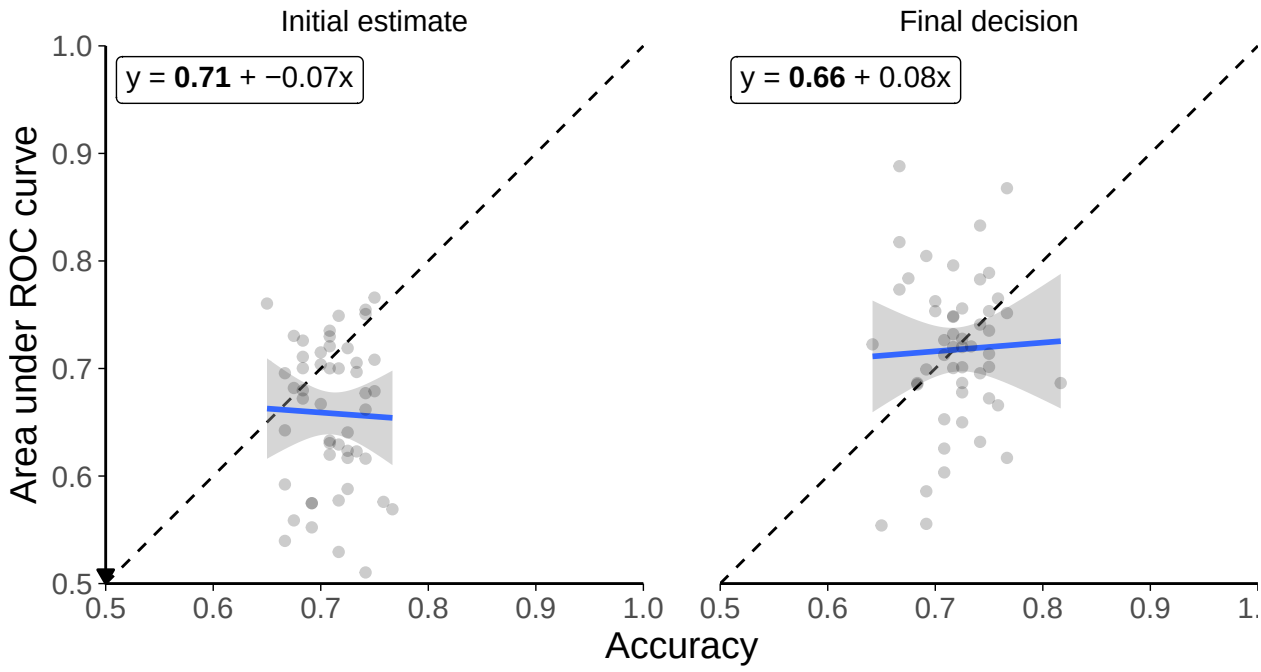


Figure 5.5: AUROC-accuracy correlation for the Dots task with in/accurate advisors. Points show individual participant data for their area under the receiver operator characteristic (ROC) curve and their accuracy on initial estimates and final decisions. The blue lines and equation text show best-fit regression, and the shaded area gives its standard error. The equations give the regression equation plotted in blue, with bold coefficients being significant at p = .05.

Confidence change The extent and manner of confidence changes is an important indicator of the extent to which participants treat advice as informative (Figure 5.6). As expected from participants who are paying attention to the task and attempting to maximize the calibration and accuracy of their final decisions, when participants receive agreeing advice they tend to increase their confidence in their initial response. Likewise, where participants receive disagreeing advice, they tend to reduce their confidence in their answer. Somewhat surprisingly, when participants change their decision (a relatively uncommon event) they quite often make their final decision with a confidence equivalent to their initial estimate, producing a distinctive off-diagonal pattern along the $y = -x$ line.

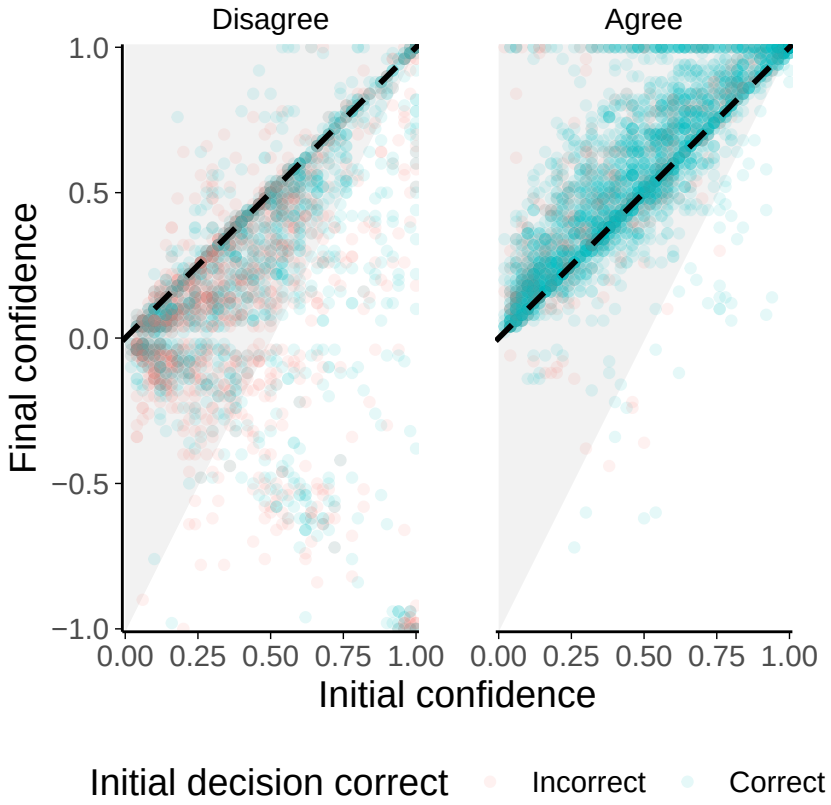


Figure 5.6: Confidence change on the Dots task with in/accurate advisors.
Each point shows the initial and final confidence on a single trial. Final confidence is coded relative to initial confidence so that increasing confidence in the opposite decision (i.e. confidence on trials where the participant changed their mind) is increasingly negative. Points above the dashed y = x line represent increased confidence, while those below it give decreased confidence. Points close to the y = x line indicate relatively little change, while points further away indicate relatively greater change. The shaded grey area shows the zone outside which influence is capped (by moving vertically towards the grey zone boundary) when using the capped influence measure. Agreement and disagreement trials are plotted seperately, with trials coloured according to whether the initial decision was correct.

On disagreement trials there is a tendency for some participants to change their minds while preserving their confidence. According to an intuitive model of confidence updating following advice, advice to the contrary of one's opinion should reduce confidence in one's initial estimate, and, if this confidence reduction is sufficiently strong, reverse the categorical decision. This model suggests that a given piece of advice moves an estimate a given distance along a continuous response dimension, and thus that the more confidently made the initial estimate

is, the less confidently made the final decision will be (and the less likely a change of mind will occur at all). This model would predict a pattern of responses on change-of-mind trials which follows a $y = x - 1$ line where the intercept indicates a change of mind. Instead, responses appear to lie more clearly on a $y = -x$ line, where higher initial confidence predicts *higher* final confidence.

The off-diagonal ($y = -x$) line is a puzzle, but it is likely an effect of aggregating data from multiple individuals rather than a general tendency to answer by jumping from one confidence bar to another and preserving confidence while altering the categorical decision. This is shown in Figure 5.7. Participants exhibit a range of relationships between their initial and final confidences when they change their minds. Some participants have lower final confidence the higher their initial confidence (positive Beta value for the slope), most have show a fairly flat relationship (very few are significantly different from zero), and a few demonstrate the positive relationship between initial and final confidence that produces the off-diagonal pattern. Of those participants who do show the off-diagonal pattern, only a few show it relatively clearly, and even these participants generally confine their responses to small parts of the scale. The overall pattern of a clear off-diagonal is thus made up of the responses of a few participants who show hints of that pattern and a good many participants whose responses cluster on that off-diagonal while internally having a flat or negative relationship between initial and final responses.

The flat response patterns, which make up the majority of participants' graphs, indicate giving very similar final confidence scores after a change of mind regardless of the initial confidence. This is an intuitive strategy if the category boundary between left and right responses is seen as important. Participants may have some level of confidence in one categorical answer, and may be persuaded to abandon that answer following advice, but may not have any meaningful variation in their confidence following that change of mind. This makes more sense when we note that the majority of these participants with flat response patterns have final confidence scores very close to zero, i.e. their final decisions are made very tentatively following their change of mind.

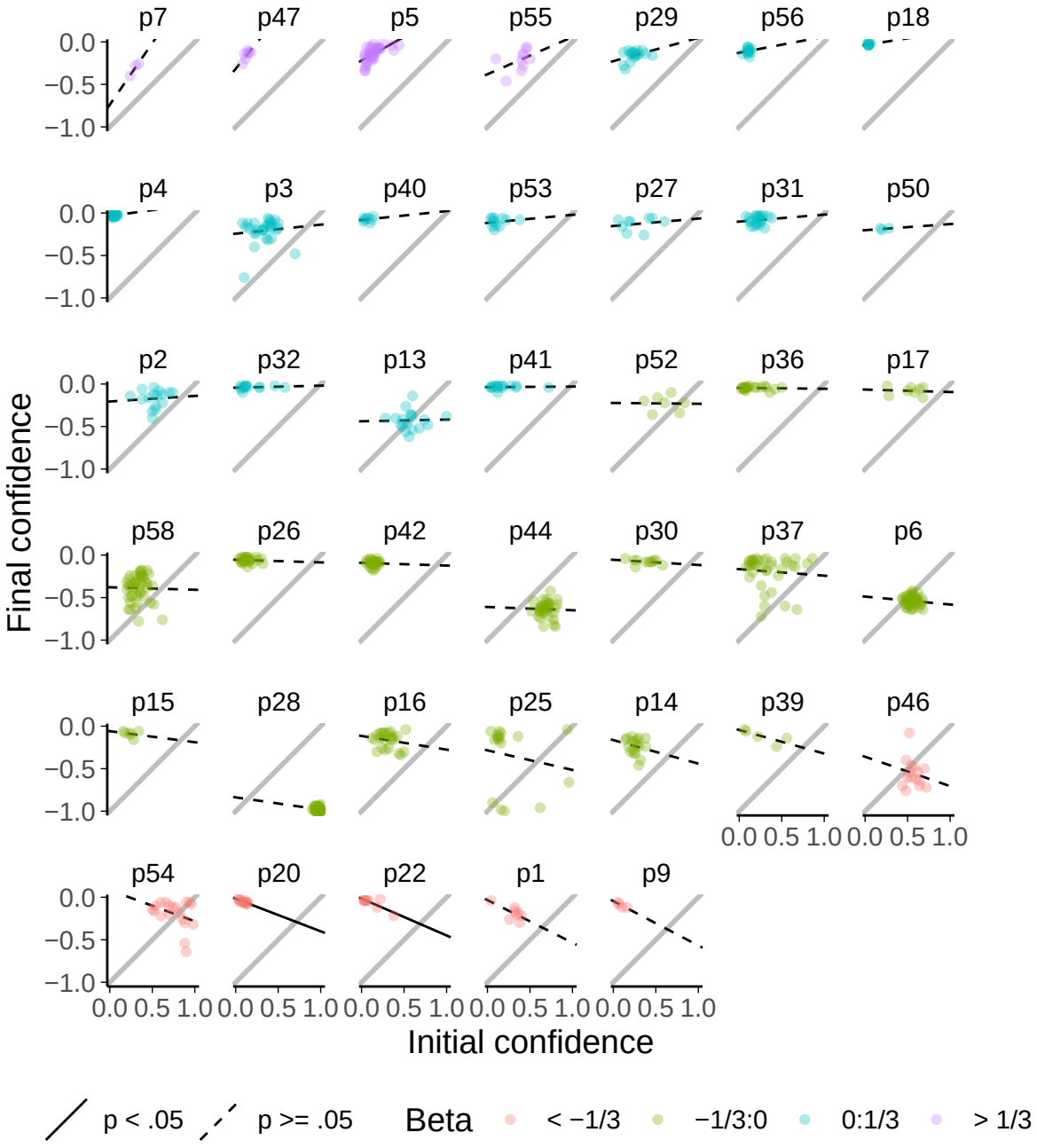


Figure 5.7: Change-of-mind confidence updating on the Dots task.
Each facet shows data from a single participant for trials where they changed their mind on the categorical decision between the initial estimate and the final decision. Participants who never changed their mind are not included. Each point shows the initial and final confidence on a single trial. All final confidence scores are negative because they are coded relative to initial confidence; increasing confidence in the final decision is increasingly negative.
Lines show the best fit for a linear prediction of final from initial confidence, with solid lines indicating that the slope is significantly different from zero at alpha = .05. Points are coloured according to the value of the slope parameter. The grey line is the y = x - 1 line that shows the expected fit line according to the intuitive model of confidence updating.

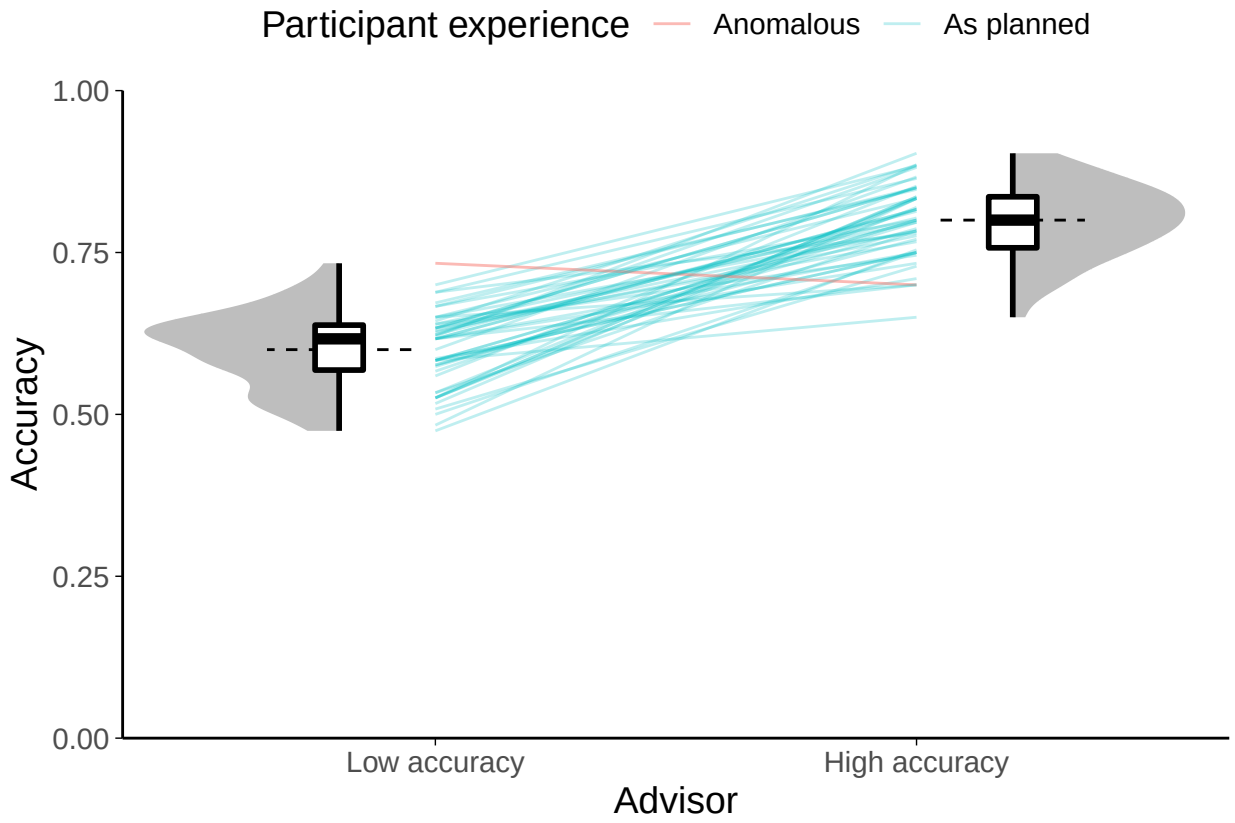


Figure 5.8: Advisor accuracy for Dots task with in/accurate advisors.
Coloured lines show the average accuracy of the advisors as experienced by an individual participant. The colour of the line indicates whether the more accurate advisor was more accurate as per the experiment design. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the accuracy level for the advisors specified in their design.

Experience with advisors The advice is generated probabilistically from the rules described previously in Table 5.1. It is thus important to get a sense of the actual advice experienced by the participants.

Advisor accuracy As shown in Figure 5.8, all but one participants experienced the High accuracy advisor as providing more accurate advice than the Low accuracy advisor, as intended in the experiment design. This indicates that the manipulation was effective for almost all participants individually, as well as for the sample on average.

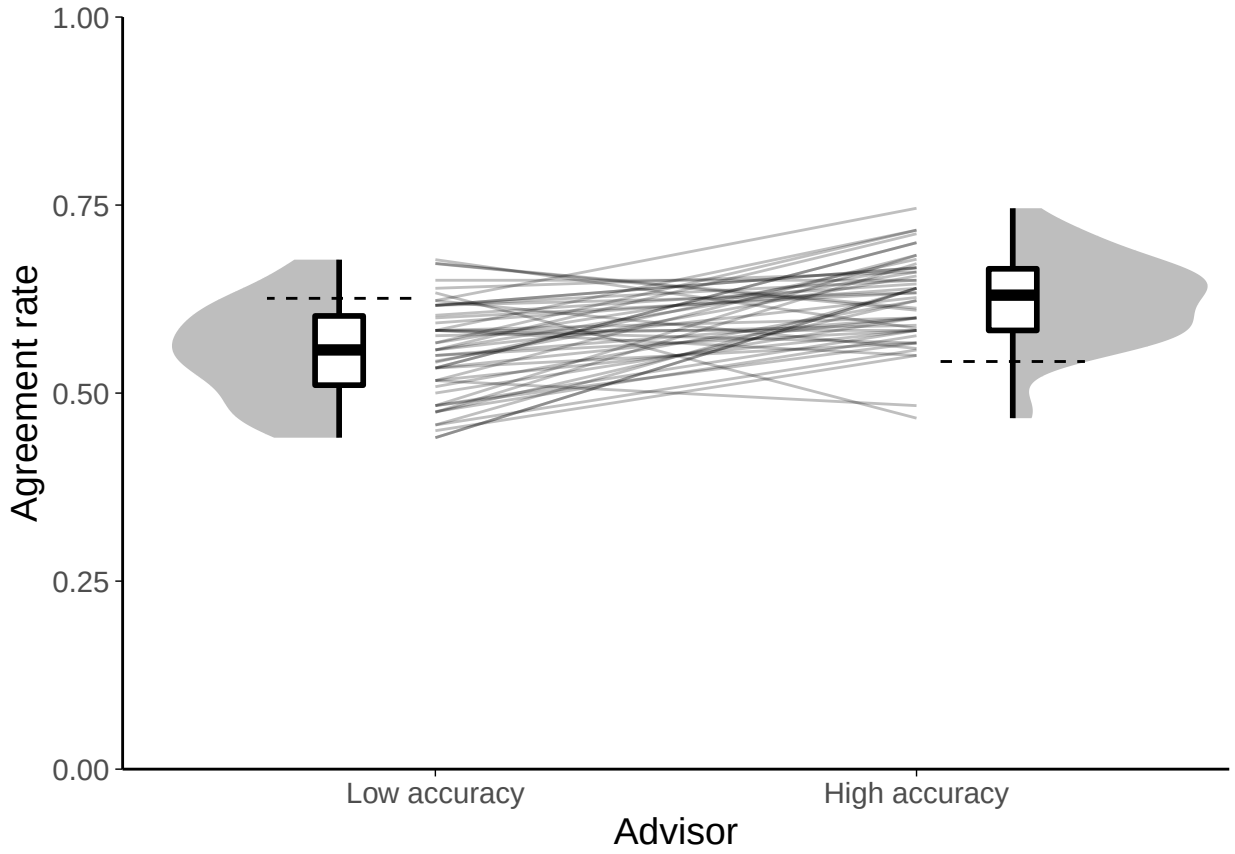


Figure 5.9: Advisor agreement for Dots task with in/accurate advisors.
Faint lines show the average agreement rate of the advisors as experienced by an individual participant. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the agreement level for the advisors specified in their design.

Advisor agreement Figure 5.9 shows the agreement rates experienced by each participant. Most participants experienced a higher agreement rate from the High accuracy advisor than from the Low accuracy advisor, but this was not unanimous. According to our theory, in the absence of feedback, as in this experiment, agreement rate should predict advisor choice better than advisor accuracy, and the heterogeneity between agreement rates and accuracy should allow this to be tested.

Should this break down agreement by initial in/correct as per the experiment design?

Advisor influence During the Familiarization stage of the experiment, participants were assigned advisors by fiat. Although participants were primarily learning

about the advisors at this time, we can nonetheless look at differences in the influence of the advisors without the confound of advisor choice. As shown in Figure 5.10, the level of influence was equivalent between advisors, with most participants being almost exactly equally influenced by both advisors. Do we want to perhaps look at the last half of the trials on these blocks for influence, so we can make the argument that they've had a chance to get to know the advisors a little? How long are the blocks?

Hypothesis test As shown in Figure 5.11, participants selected the High accuracy advisor at a rate greater than would be expected if their choosing were random ($t(49) = 3.09$, $p = .003$, $d = 0.44$, BF = 9.96; $M = 0.57$ [0.52, 0.61], $\mu = 0.5$). The modal choice remained at chance level (.5), but almost all participants manifesting a preference preferred the High accuracy advisor.

5.1.2 Dates Task

Open scholarship practices

`https://osf.io/5xpvq`

!TODO[OSFify data for these studies]

`https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACBin/acc.html`

Unanalysed data Early versions of this experiment (v0-0-1, v0-0-2) included a bug which prevented feedback from being shown during the familiarisation phase even to participants in the Feedback condition. The 13 participants whose data was collected in these versions is not included in analysis. These participants could theoretically be included in the No feedback condition regardless of their condition label in the data, but this is not done here.

Method

This study used the binary version of the Dates Task (§3.1.3).

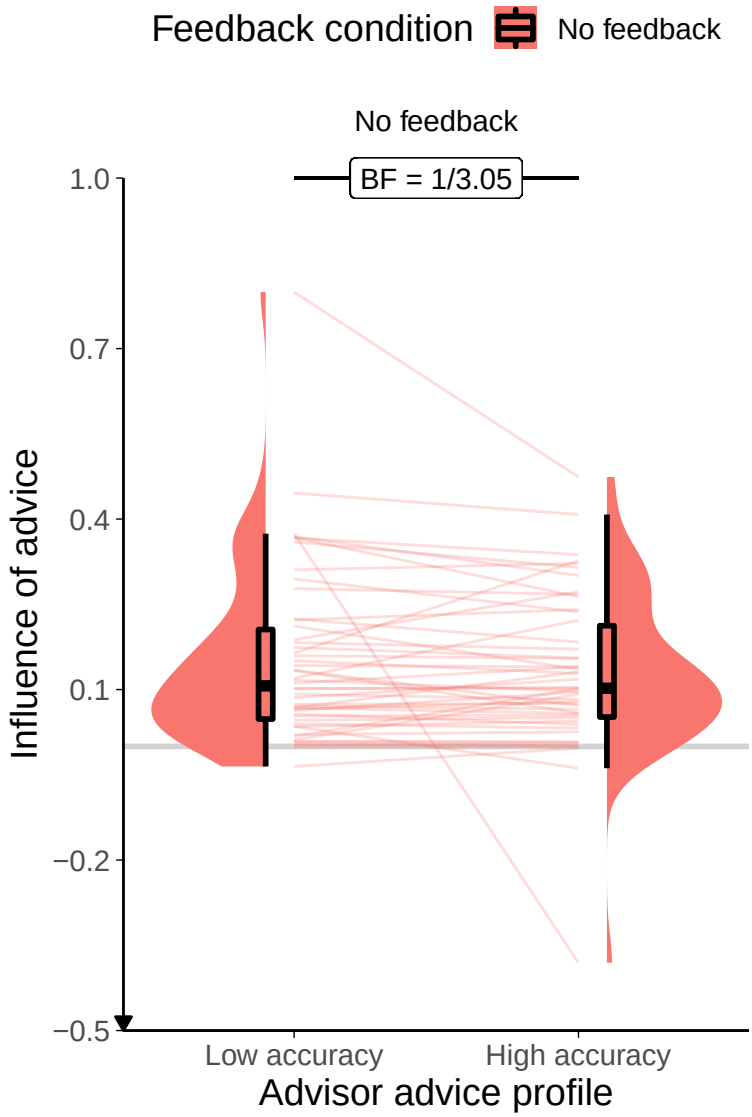


Figure 5.10: Dot task advisor influence for in/accurate advisors.
Participants' weight on the advice for advisors in the Familiarization stage of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

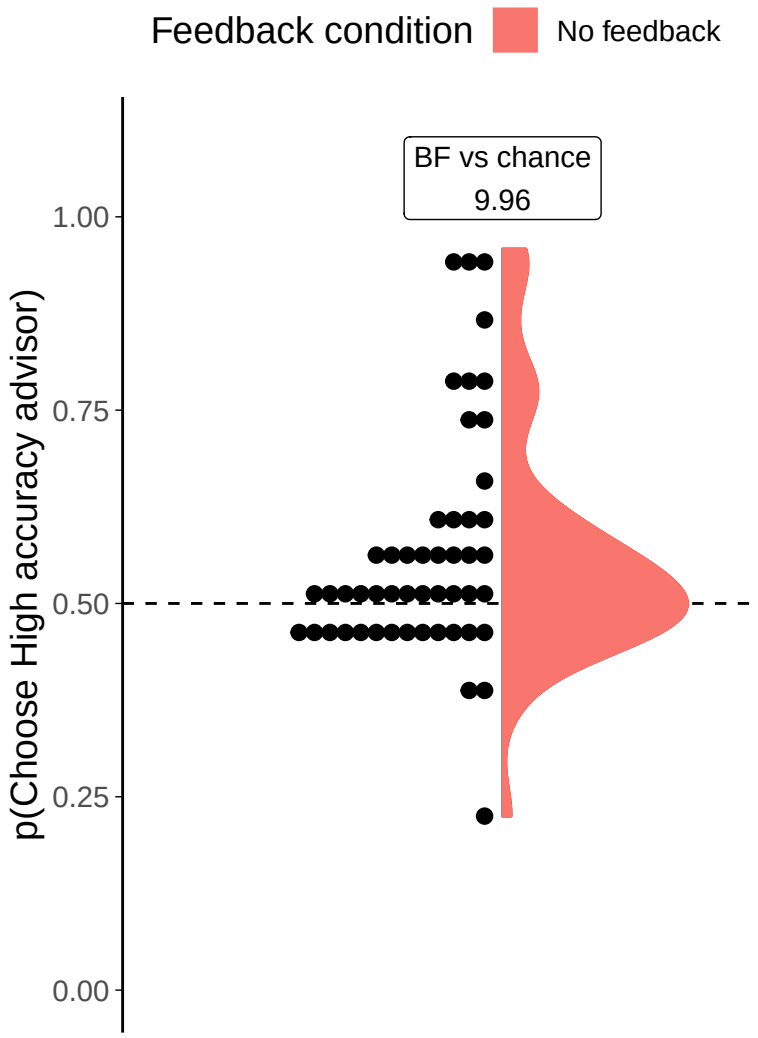


Figure 5.11: Dot task advisor choice for in/accurate advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

Table 5.3: Advisor advice profiles for Dates task Accuracy experiment

	Probability of agreement (%)		
Advisor	Participant correct	Participant incorrect	Overall accuracy
High accuracy	.800	.200	.800
Low accuracy	.590	.410	.590

Table 5.4: Participant exclusions for Dates task Accuracy experiment

Reason	Participants excluded
Too few trials	0
Insufficient advice taking	0
Too few choice trials	0
Total excluded	**0**
Total remaining	**62**

Advice profiles The High accuracy and Low accuracy advisor profiles issued binary advice (endorsing either the 'before' or 'after' column) probabilistically based on whether or not the participant had selected the correct column in their initial estimate. The High accuracy advisor was agreed with the participant's initial estimate on 80% of the trials where the participant was correct, but only 20% of the trials on which the participant was incorrect, meaning that the High accuracy advisor was correct 80% of the time. Using an analogous setup, the Low accuracy advisor was correct 59% of the time. The advisor profiles were not balanced for overall agreement rates because the agreement rate experienced by a participant depends upon the accuracy of that participant's initial estimates.

Results

Exclusions Individual trials were screened to remove those that took longer than 60s to complete. Participants were then excluded for having fewer than 11 trials remaining, fewer than 10 trials on which they had a choice of advisor, or for giving the same initial and final response on more than 90% of trials. Overall, 0 participants were excluded, with the details shown in Table 5.4.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour,

it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial and final decisions can be seen in Figure 5.12. The distribution of these response times helps characterise some differences between the Dots task and the Dates task. In the former, decisions for both initial estimates and final decisions are tightly clustered, with a clear structure and pattern to the responses for all participants (5.1). In the Dates task however, response times are not only longer, but they are also much more varied within participants. Some increase in variance is expected with an increase in mean, especially with fewer trials for each participant, but the extent of the differences clearly shows that the tasks provide participants with different experiences: the Dots task is tightly rhythmic and repetitive, while the Dates task is more heterogeneous.

```
## Picking joint bandwidth of 1030
```

Accuracy Unlike in the Dots version of the task, participant accuracy is not controlled. Participants managed to improve their performance from their initial estimates to their final decisions with both advisors (Figure 5.13). This is likely because the advisors themselves were more accurate than the participants, so following their advice was generally a good strategy, and the difficulty of the task meant that participants were very willing to be influenced by advice.

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions.

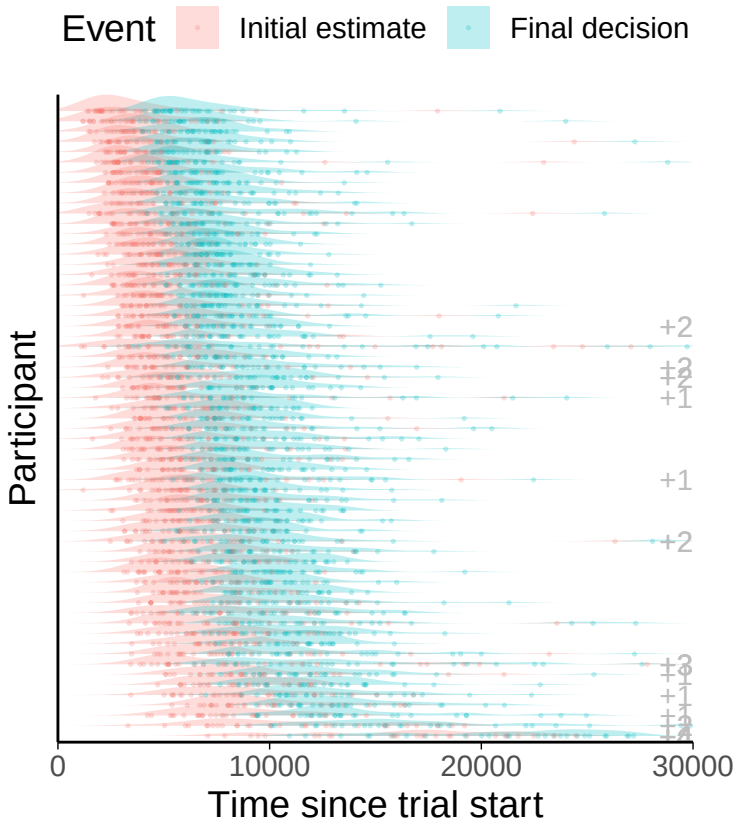


Figure 5.12: Response times for the Dates task with in/accurate advisors.
Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

Metacognitive ability The participants' metacognitive abilities were highly variable, with many participants displaying below-chance metacognitive ability (Figure 5.15). While this may appear concerning, recall that metacognitive sensitivity and bias vary substantially and cannot be reliably estimated using ROC curves where performance accuracy on the underlying task is highly variable, so it is not necessarily the case that these values give cause for alarm. The correlation between performance on the underlying task and metacognitive ability (Figure 5.16) shows that, as one might expect, participants with a greater ability to perform the Dates task have a greater insight into their performance on the Dates task. This in turn suggests that, despite the low number of trials on the task, we are able to obtain meaningful

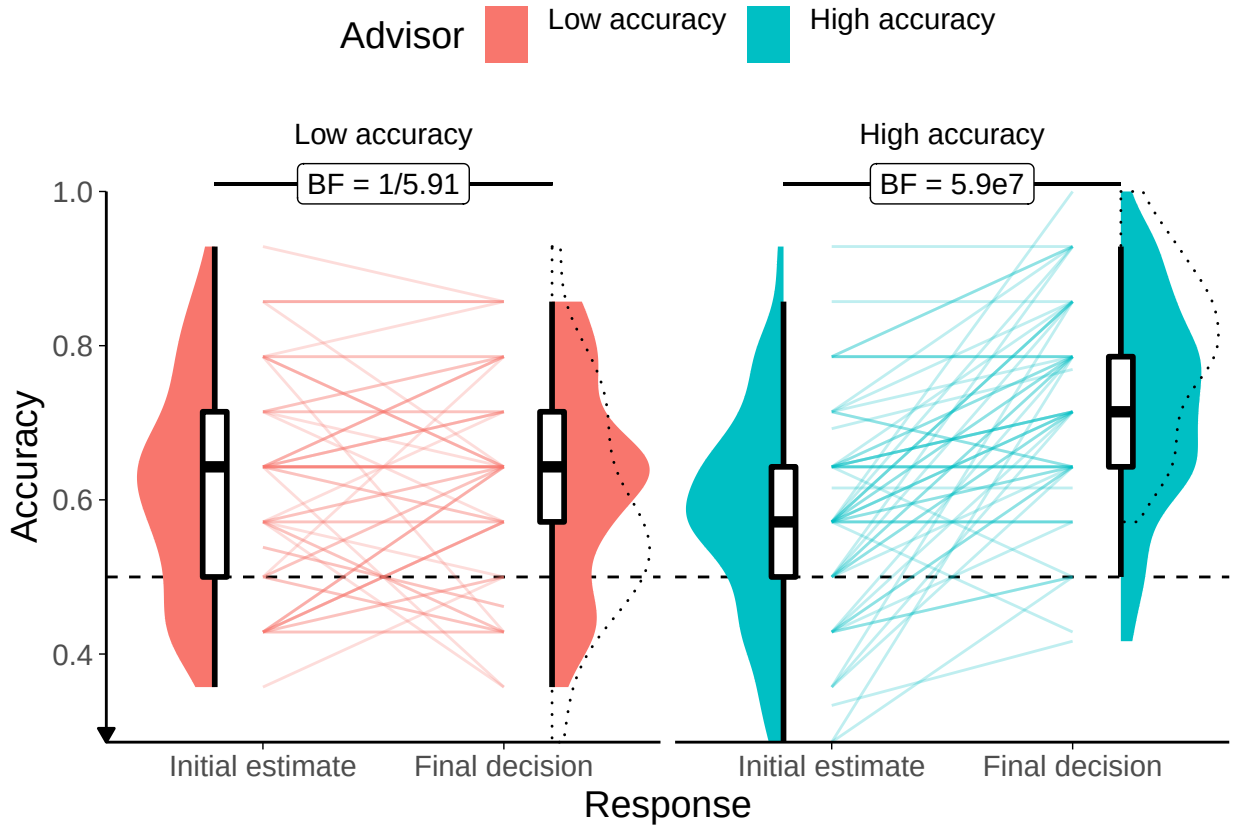


Figure 5.13: Response accuracy for the Dates task with in/accurate advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.
Because there were relatively few trials, the proportion of correct trials for a participant generally falls on one of a few specific values. This produces the lattice-like effect seen in the graph. Some participants had individual trials excluded for over-long response times, meaning that the denominator in the accuracy calculations is different, and thus producing accuracy values which are slightly offset from others'.

insights into participants' metacognitive abilities, albeit without being able to precisely estimate the metacognitive sensitivity or bias for an individual participant.

Confidence change The extent and manner of confidence changes is an important indicator of the extent to which participants treat advice as informative (Figure 5.17). As with the Dots task (Figure 5.6, when participants receive agreeing advice they tend to increase their confidence in their initial response, suggesting they are paying attention to the task and attempting to maximize the calibration and accuracy of their final decisions. Likewise, where participants receive disagreeing

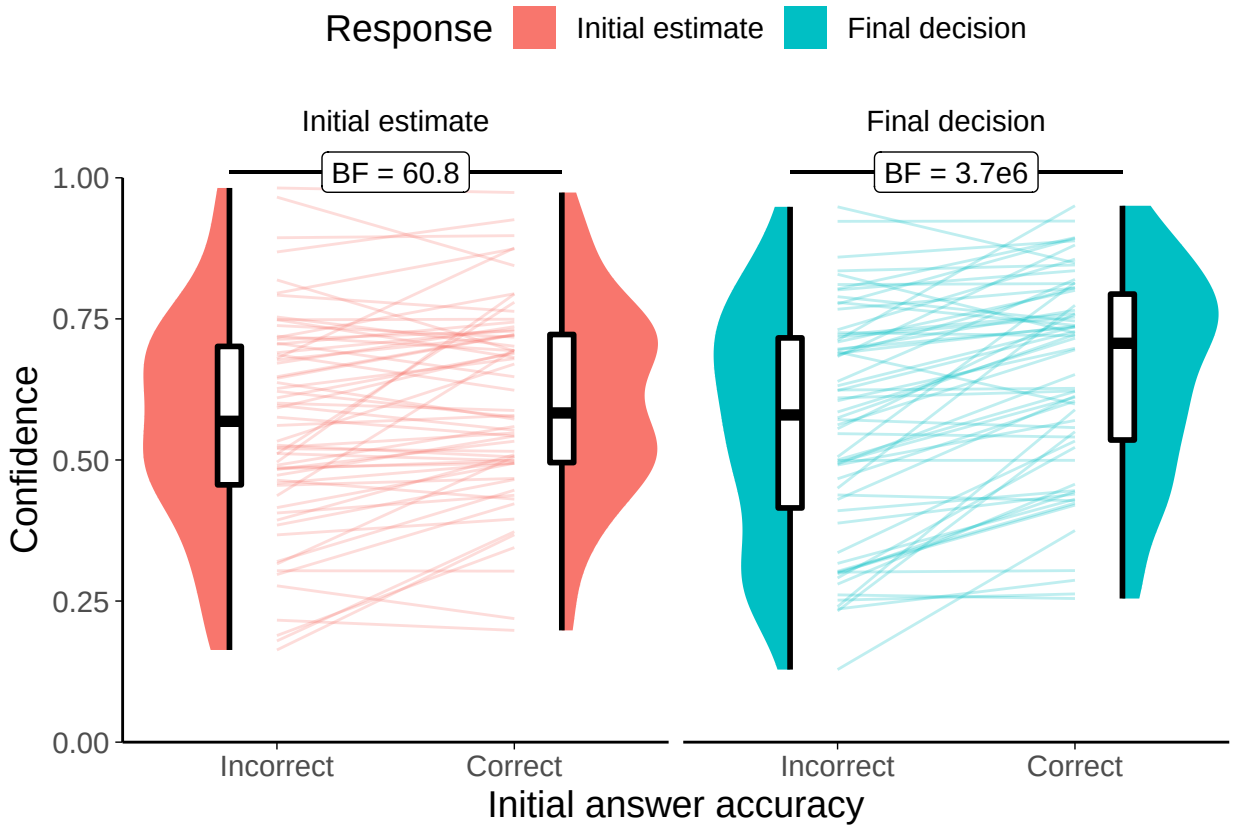


Figure 5.14: Confidence for the Dates task with in/accurate advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions.

advice, they tend to reduce their confidence in their answer. As before, when participants change their decision (a relatively uncommon event) they quite often make their final decision with a confidence equivalent to their initial estimate, producing the distinctive off-diagonal pattern along the $y = -x$ line. This pattern was explained as largely an artefact of aggregating data from multiple individual participants in the Dots task, but the same explanation is less tenable here.

The individual participant data suggest that some participants may actually be porting their confidence from the initial estimate directly into the final decision, despite the fact that they have changed their mind (Figure **??**). Whereas in the Dots task the off-diagonal pattern in the aggregate data looked to be an artefact of combining data from individual participants with relatively little variation within their confidence responses placed at various points along the off-diagonal (Figure

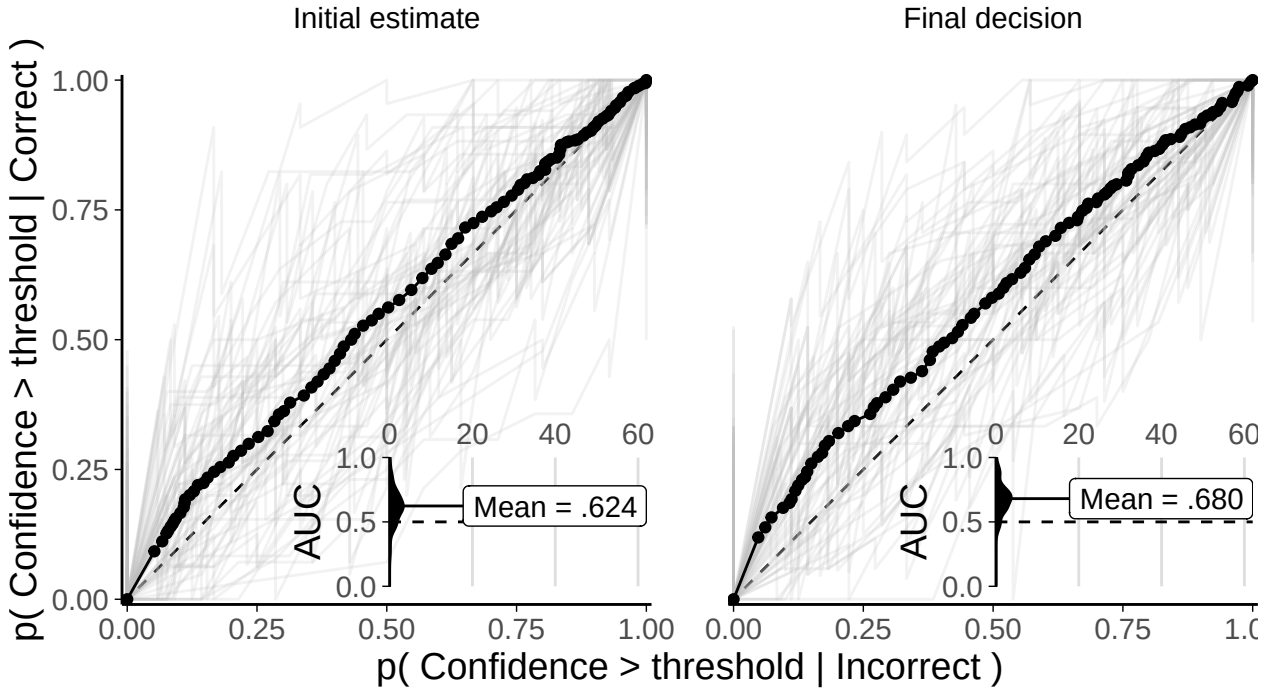


Figure 5.15: ROC curves for the Dates task with in/accurate advisors.
Faint lines show individual participant data, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers.

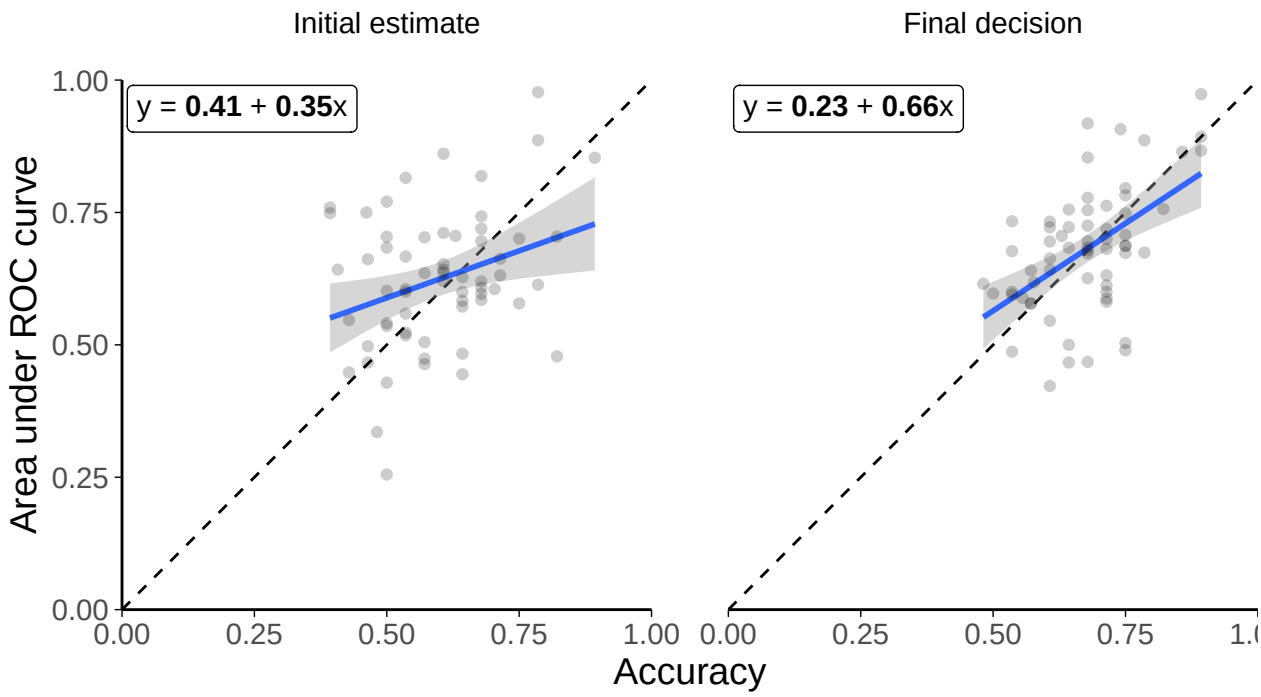


Figure 5.16: AUROC-accuracy correlation for the Dots task with in/accurate advisors. Points show individual participant data for their area under the receiver operator characteristic (ROC) curve and their accuracy on initial estimates and final decisions. The blue lines and equation text show best-fit regression, and the shaded area gives its standard error. The equations give the regression equation plotted in blue, with bold coefficients being significant at p = .05.

??), data from this task suggested that several participants were using a range of values for their final decisions and displaying a clear positive correlation between initial and final confidence.

This increased tendency for participants to response in this way on the Dates task may be a consequence of the layout of the response bars on the screen. In the Dots task, the bars were oriented horizontally, such that the extreme values for one were furthest away from extreme values for the other (Figure 3.3). In the Dates task, the bars were oriented vertically, meaning that the shortest distance from one point on one bar was to the same point on the other bar (Figure 3.5). Participants changing their minds may well have felt that the important feature

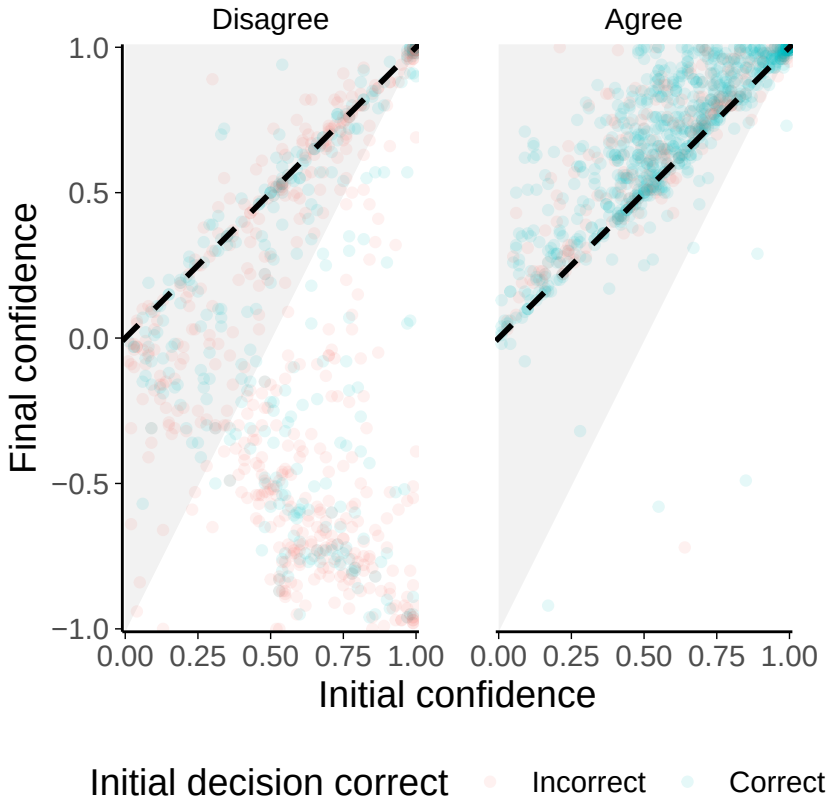


Figure 5.17: Confidence change on the Dates task with in/accurate advisors.
Each point shows the initial and final confidence on a single trial. Final confidence is coded relative to initial confidence so that increasing confidence in the opposite decision (i.e. confidence on trials where the participant changed their mind) is increasingly negative. Points above the dashed y = x line represent increased confidence, while those below it give decreased confidence. Points close to the y = x line indicate relatively little change, while points further away indicate relatively greater change. The shaded grey area shows the zone outside which influence is capped (by moving vertically towards the grey zone boundary) when using the capped influence measure. Agreement and disagreement trials are plotted seperately, with trials coloured according to whether the initial decision was correct.

was the change of response bar, and not concerned themselves with reporting their confidence. These participants would naturally have taken the shortest route from their initial estimate to the other response bar, which would be the same confidence as for the initial estimate, producing the positive correlation between initial and final confidence seen in these data.

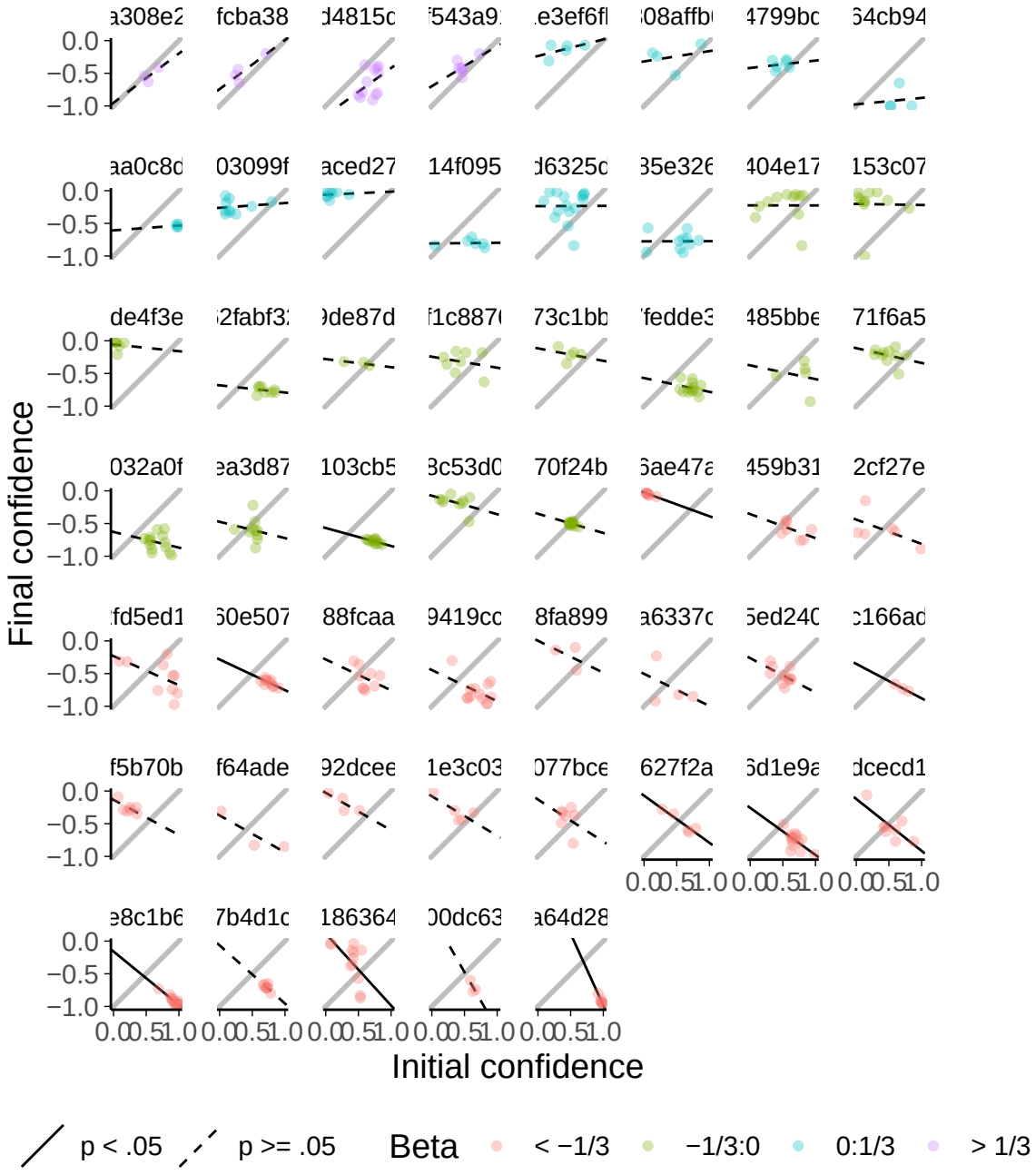


Figure 5.18: Change-of-mind confidence updating on the Dates task.
Each facet shows data from a single participant for trials where they changed their mind on the categorical decision between the initial estimate and the final decision. Participants who never changed their mind are not included. Each point shows the initial and final confidence on a single trial. All final confidence scores are negative because they are coded relative to initial confidence; increasing confidence in the final decision is increasingly negative.
Lines show the best fit for a linear prediction of final from initial confidence, with solid lines indicating that the slope is significantly different from zero at alpha = .05. Points are coloured according to the value of the slope parameter. The grey line is the y = x - 1 line that shows the expected fit line according to the intuitive model of confidence updating.

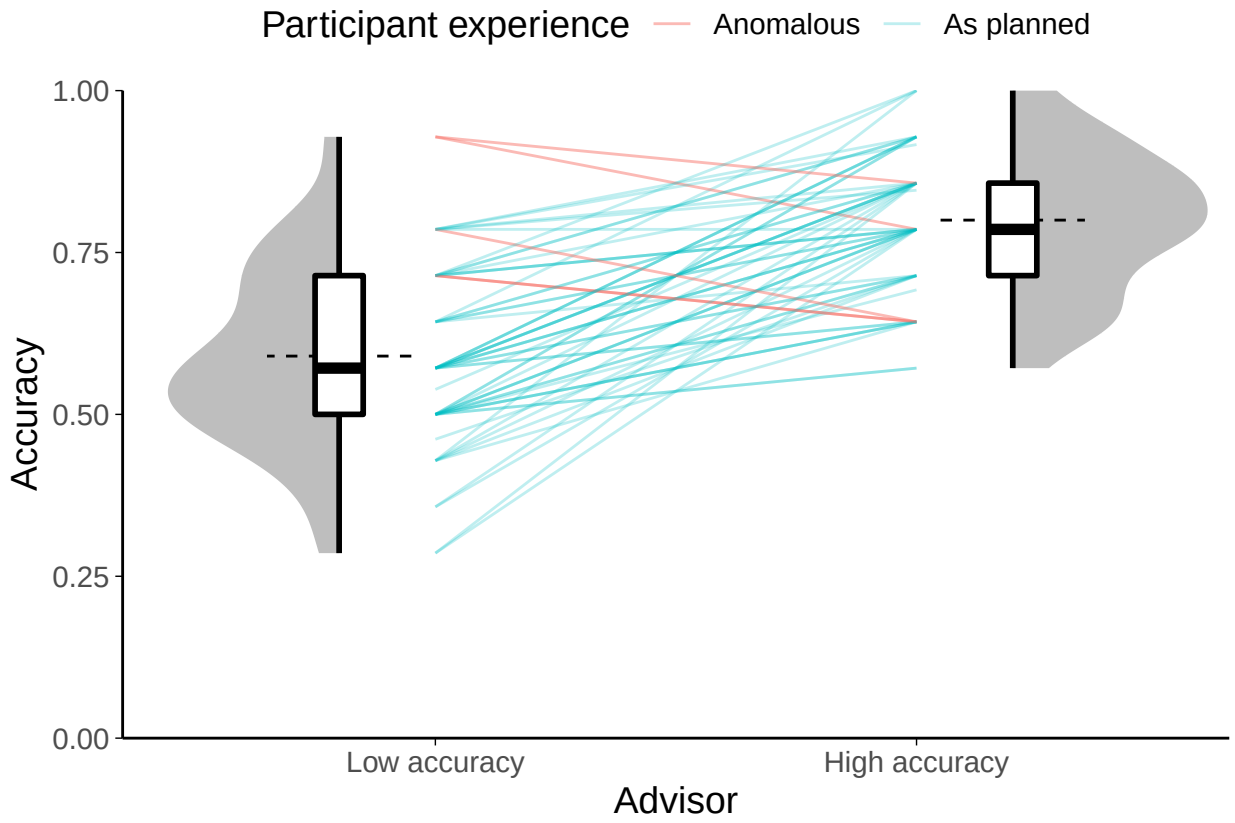


Figure 5.19: Advisor accuracy for Dates task with in/accurate advisors.
Coloured lines show the average accuracy of the advisors as experienced by an individual participant. The colour of the line indicates whether the more accurate advisor was more accurate as per the experiment design. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the accuracy level for the advisors specified in their design.

Experience with advisors The advice is generated probabilistically from the rules described previously in Table 5.1. It is thus important to get a sense of the actual advice experienced by the participants.

Advisor accuracy As shown in Figure 5.19, most participants experienced the High accuracy advisor as providing more accurate advice than the Low accuracy advisor, as intended in the experiment design. This indicates that the manipulation was effective for most participants individually, as well as for the sample on average.

Advisor agreement Figure 5.20 shows the agreement rates experienced by each participant. There was a mixture of participants who experienced a higher

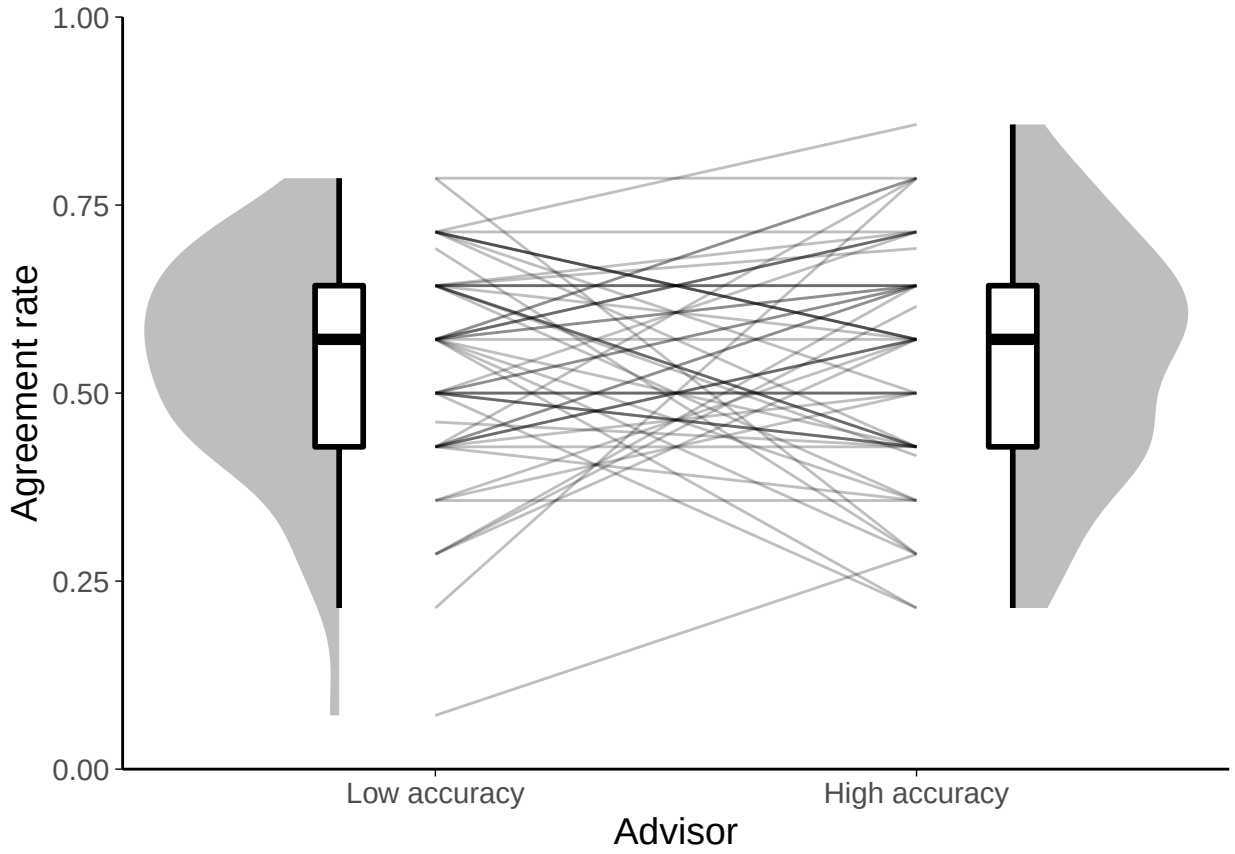


Figure 5.20: Advisor agreement for Dates task with in/accurate advisors.
Faint lines show the average agreement rate of the advisors as experienced by an individual participant. Box plots and violins show the distribution of the participant means.

agreement rate each advisor. According to our theory, in the absence of feedback, as in this experiment, agreement rate should predict advisor choice better than advisor accuracy, and the heterogeneity between agreement rates and accuracy should allow this to be tested.

Advisor influence Neither advisor appeared to be substantially more influential than the other, either with or without feedback (Figure 5.21). In both conditions, the High accuracy advisor was numerically more influential than the Low accuracy advisor. The accuracy measurements are calculated on the Familiarization phase trials in which participants are not offered a choice of advisor. It is during this phase that participants are learning about the value of the advice (especially in the Feedback condition), and thus any influence on later trials may be diluted

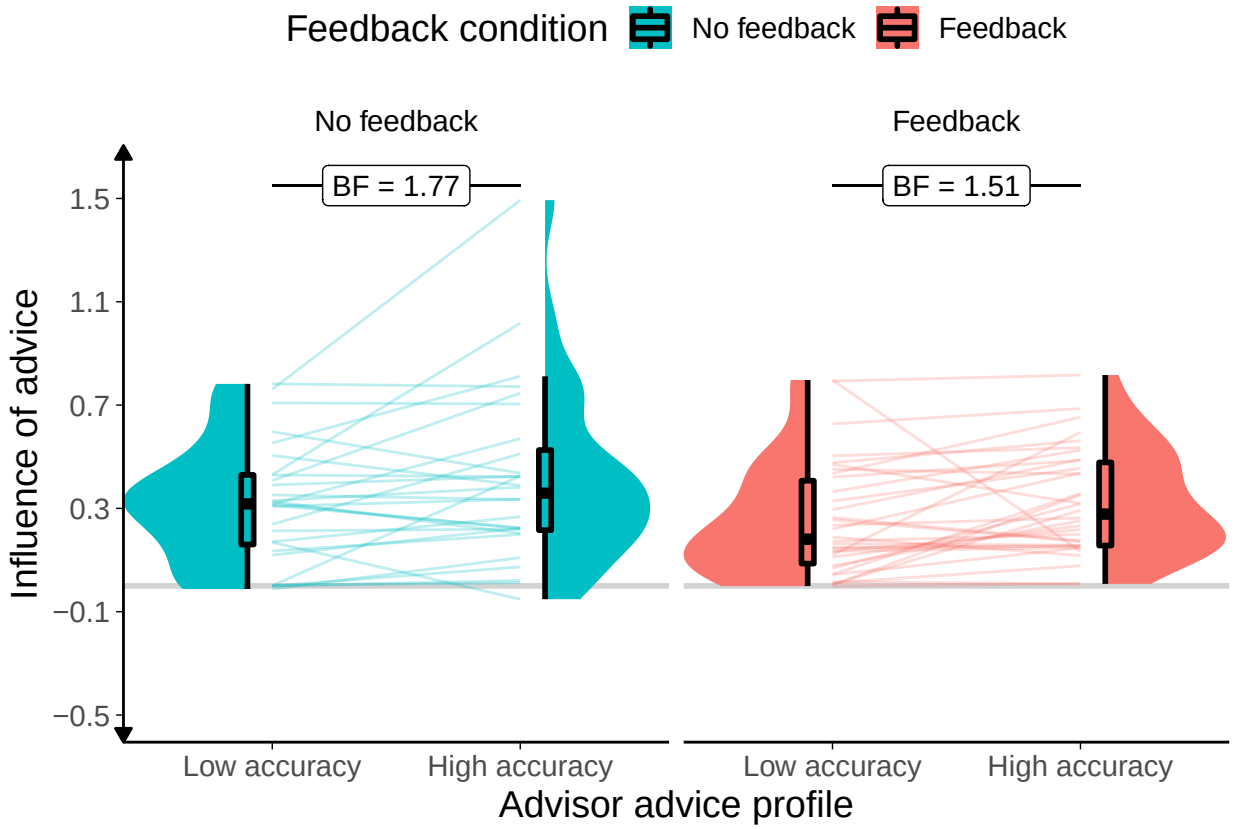


Figure 5.21: Date task advisor influence for in/accurate advisors.
Participants' weight on the advice for advisors in the Familiarization phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

by low influence on trials which occur before an advisor has had time to develop a reputation as reliable. This means that influence cannot be used as a reliable outcome measure for this experimental design, but it is nevertheless useful to explore to get a sense of how participants responded to the advice. An inspection of the individual participants' data shows that very few participants had large influence differences between advisors.

Hypothesis test In the No feedback condition the mean of the distribution of participant picking preferences between the advisors was equivalent to chance ($t(27)$ = -0.93, p = .363, d = 0.18, BF = 1/3.37; $M_{Nofeedback}$ = 0.45 [0.33, 0.57], μ = 0.5). This is a different result to that observed in the Dots task (§5.1.1),

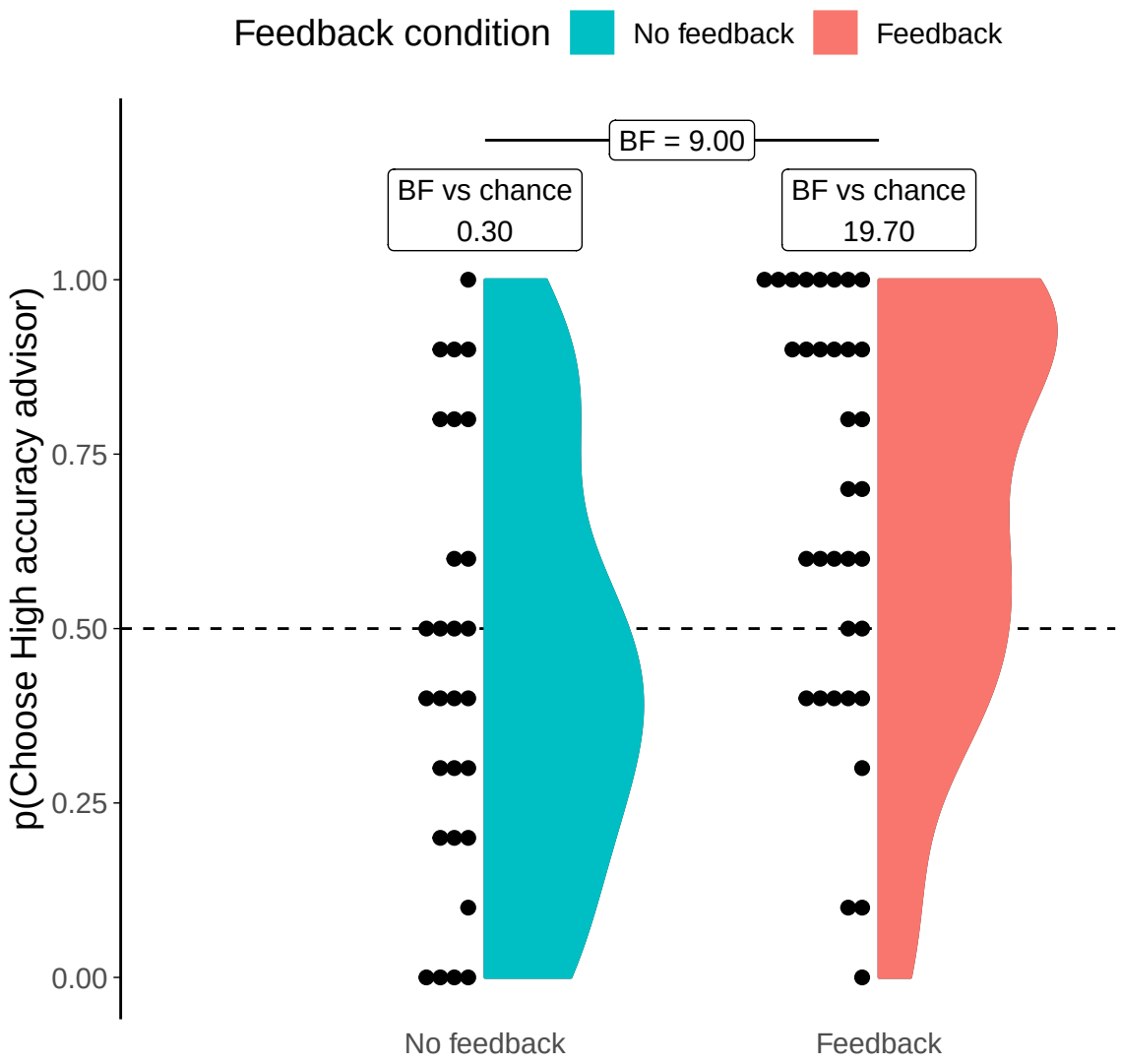


Figure 5.22: Dates task advisor choice for in/accurate advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line. Participants in the Feedback condition received feedback during the Familiarization phase, but not during the Choice phase.

which also had no feedback. Preferences were quite evenly distributed across the full range of directions and strengths, with a slight numerical advantage for the Low accuracy advisor (Figure 5.22).

In the Feedback condition the mean of the distribution of selection rates was clearly different from chance. The High accuracy advisor was preferred by more participants, and preferred more strongly ($t(33) = 3.41$, $p = .002$, $d = 0.58$, BF = 19.7; $M_{Feedback} = 0.67$ [0.57, 0.78], $\mu = 0.5$). The modal selection strategy was to select the High accuracy advisor at every opportunity. This indicates that participants could identify the more accurate advisor when feedback was provided and preferred to receive advice from that advisor.

5.1.3 Discussion

Where feedback is provided on advisors' performance, people seem to prefer high accuracy advisors to low accuracy advisors. Where feedback is not provided, people may need substantially more experience to learn that some advisors are more accurate than others, because this happens in the Dots task but not in the Dates task.

5.2 Agreement !TODO[check Niccolo covered this]

Pescetelli and Yeung !TODO[cite new paper] demonstrated that advisors who agree !TODO[somewhere we need to talk about what we mean by agreement, how Niccolo defined it, how we define it (varies between binary/continuous tasks), etc.] more frequently are more influential (regardless of the presence of feedback) in a lab-based perceptual decision-making task.

There were differences in how participants selected the advisors between the Dots task (which has no feedback) and the No feedback condition of the Dates task for High versus Low accuracy advisors. We may expect more pronounced effects in the absence of feedback when contrasting High versus Low agreement advisors, because we expect that agreement is the driving force behind the accuracy differences where feedback is not provided.

Table 5.5: Advisor advice profiles for Dots task Agreement experiment

Advisor	Probability of agreement			Overall accuracy
	Participant correct	Participant incorrect	Overall	
High agreement	.840	.610	.773	.709
Low agreement	.660	.170	.518	.709

5.2.1 Dots Task

Open scholarship practices

 NA

 !TODO[OSFify data for these studies]

 `https://github.com/oxacclab/ExploringSocialMetacognition/blob/9932543c62b00bd96ef7ddb3439e6c2d5bdb99ce/AdvisorChoice/index.html`

Unanalysed data There were no unanalysed data for this experiment.

Method

!TODO[clarify any methodological differences from the main methods chapter]

Advice profiles The two advisor profiles (Table 5.5) used in the experiment were High agreement and Low agreement The High agreement advisor gave advice that endorsed the same answer side as the participant's initial estimate 77.3% of the time while the Low agreement advisor agreed with the participant 51.8% of the time. The advisor profiles were balanced for overall accuracy rates.

Results

Exclusions Participants' data could be excluded from analysis where they have an average accuracy below 0.6 or above 0.85, do not have trials in all confidence categories, have fewer than 12 trials in each confidence category, or finish the experiment after 50 participants have already submitted data which passed the other exclusion tests. Overall, 18 participants were excluded, with the details shown in Table 5.6.

Table 5.6: Participant exclusions for Dots task Agreement experiment

Reason	Participants excluded
Accuracy too low	0
Accuracy too high	0
Missing confidence categories	7
Skewed confidence categories	12
Too many participants	0
Total excluded	**18**
Total remaining	**50**

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial and final decisions can be seen in Figure 5.23.

```
## Picking joint bandwidth of 233
```

All participants had similar patterns: initial and final responses were approximately normally distributed, with final responses having a higher variance due to incorporating the initial responses within them. Most participants show some trials on which initial or final responses took substantially longer than usual.

Accuracy Accuracy of initial decisions was controlled by a staircasing procedure which aimed to pin accuracy to 71%. The accuracy of final decisions was free to vary according to the ability of the participant to take advantage of the advice on offer. As Figure 5.24 shows, participants' accuracy scores for initial decisions were close to the target values. Participants did not tend to improve the accuracy of their responses following advice from High agreement advisors,

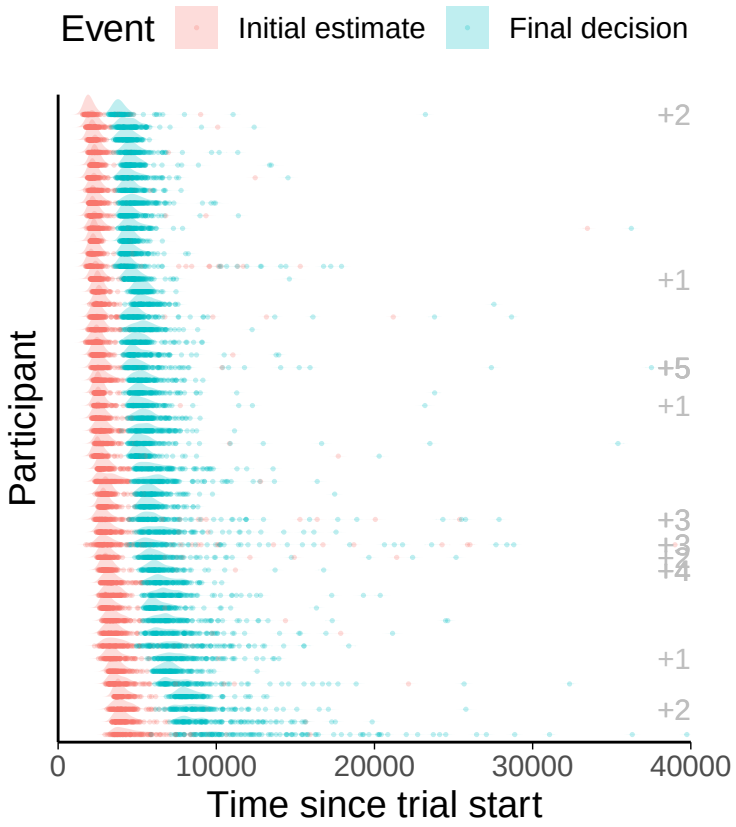


Figure 5.23: Response times for the Dots task with in/accurate advisors.
Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

while the evidence was unclear as to whether there was any difference in response accuracy with Low agreement advice. The distribution of initial estimate accuracy for trials with the High agreement advisor is slightly unusual, with an almost-bimodal structure and a median somewhat higher than the target value. There is no obvious reason why this should be the case.

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions.

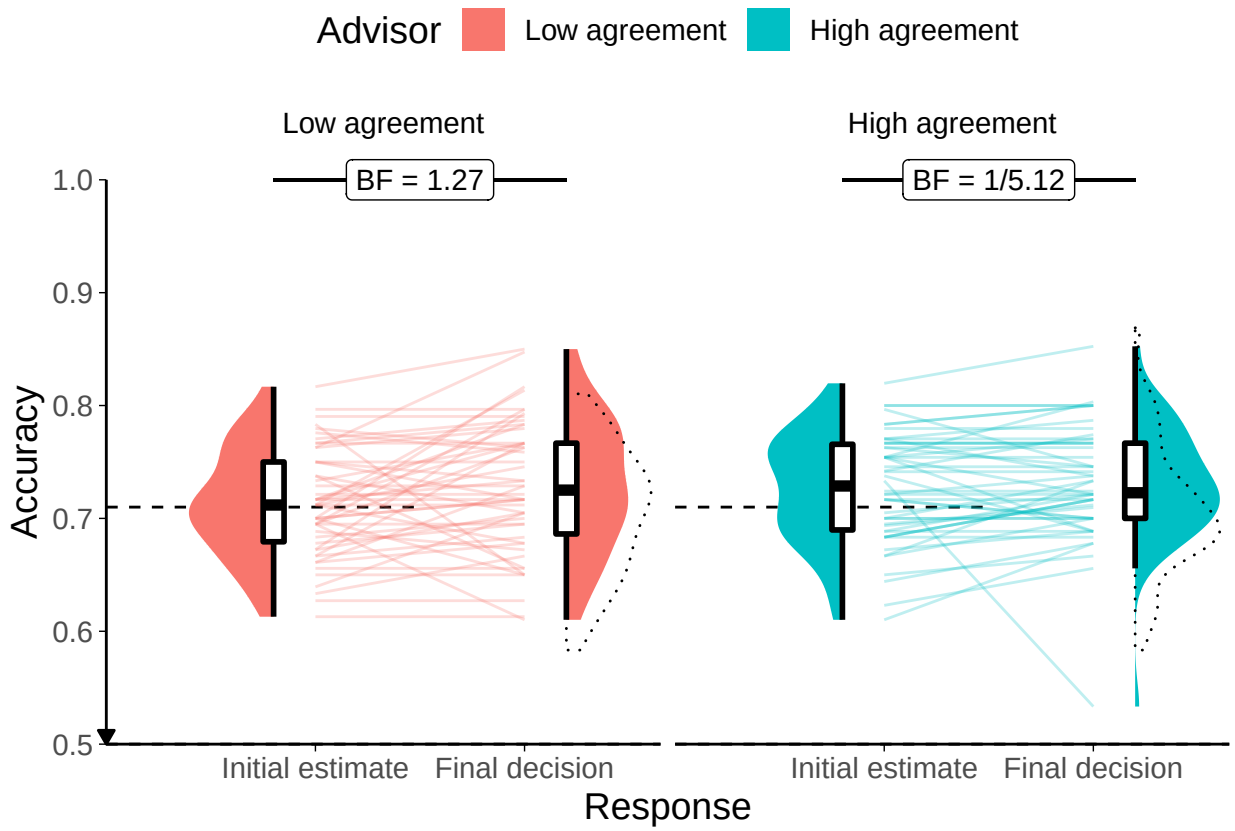


Figure 5.24: Response accuracy for the Dots task with in/accurate advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

Metacognitive ability As shown by Figure 5.26, almost all participants showed above-chance metacognitive sensitivity for both initial estimates and final decisions. Participants generally showed higher metacognitive sensitivity for final decisions, although this may be an artefact of a change in metacognitive bias. Participants' metacognitive sensitivity was not particularly high !TODO[What are typical values we might expect in the dots task and similar tasks? Is there a useful mapping between meta-d' and Type II ROC to compare with e.g. Roualt's stuff?]. There was no evidence of participants' metacognitive sensitivity being correlated with their task performance (Figure 5.27). This is expected when task performance is tightly controlled, because under these conditions variation in task performance reflects variation within a participant rather than between participants.

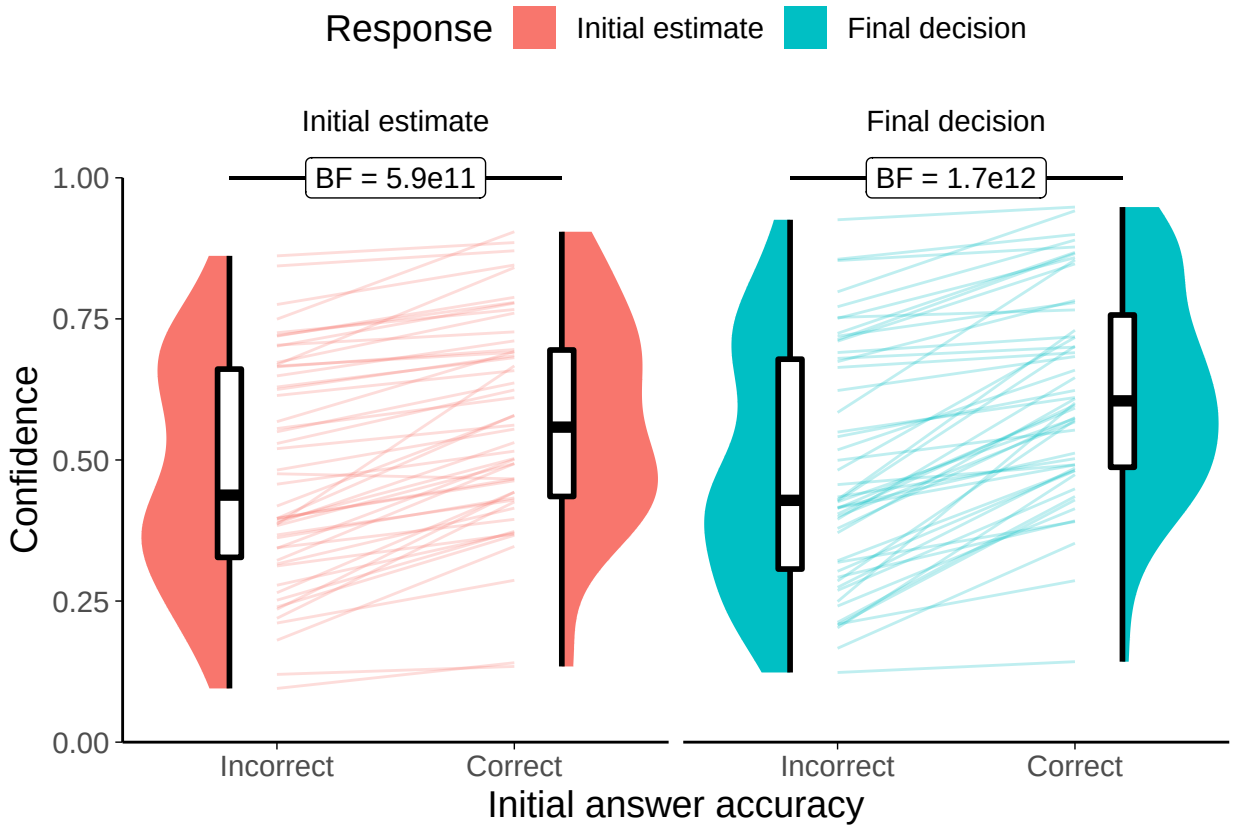


Figure 5.25: Confidence for the Dots task with in/accurate advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions.

Confidence change As with the High vs Low accuracy experiment (§5.1.1), there was an evident off-diagonal pattern in the disagreement trials, indicating a tendency for some participants to change their minds while preserving their confidence.

As before, inspection of individual participant plots for disagreement trials where participants changed their minds (Figure 5.29), this pattern is more likely due to aggregation than a behaviour of any particular individual. There were, however, more participants who showed off-diagonal-like patterns than with the Accuracy experiment.

Experience with advisors The advice is generated probabilistically from the rules described previously in Table 5.5. It is thus important to get a sense of the actual advice experienced by the participants.

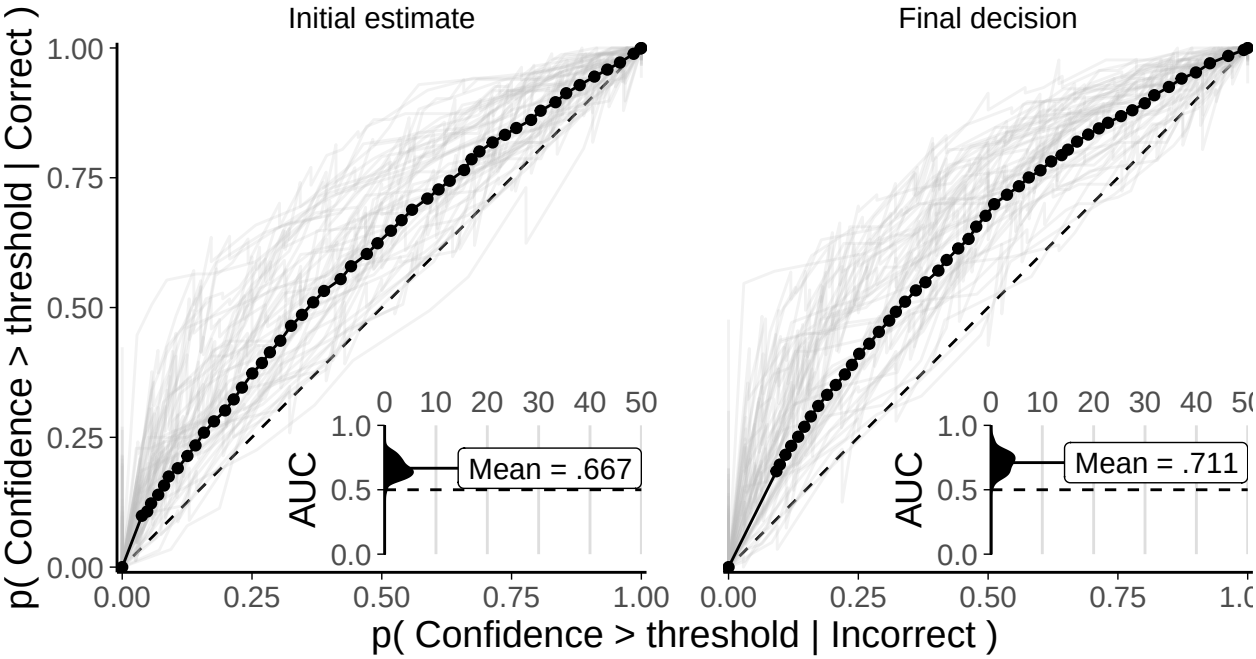


Figure 5.26: ROC curves for the Dots task with in/accurate advisors.
Faint lines show individual participant data, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers.

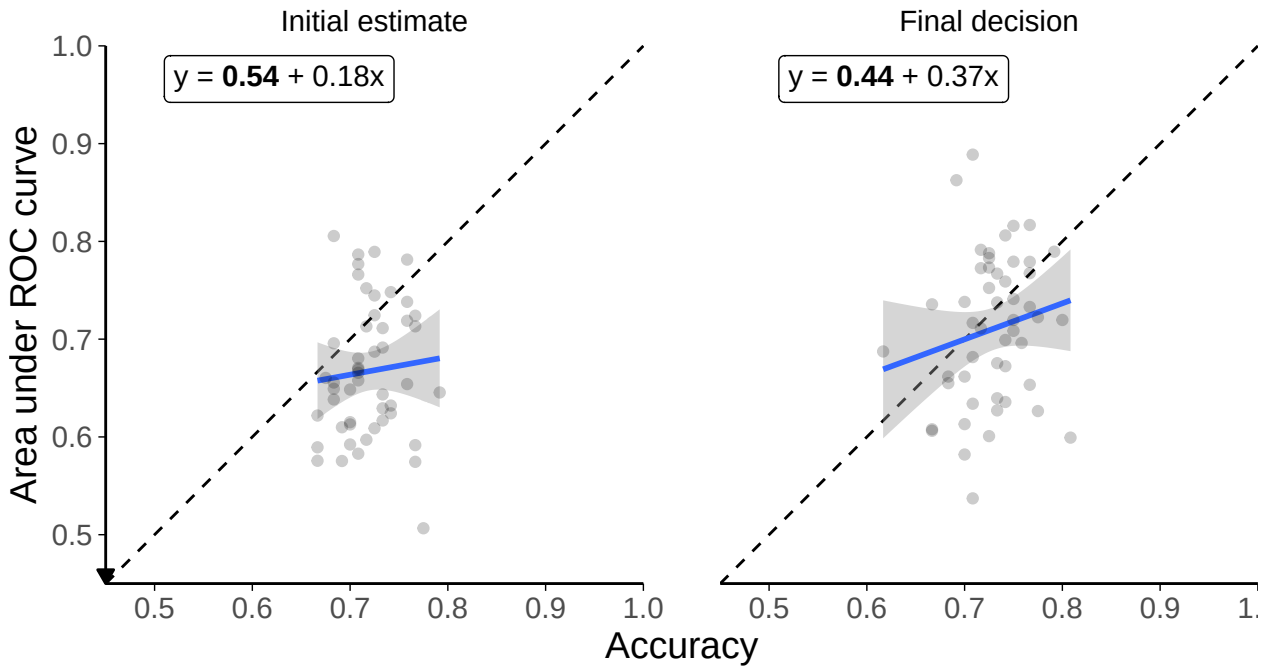


Figure 5.27: AUROC-accuracy correlation for the Dots task with in/accurate advisors. Points show individual participant data for their area under the receiver operator characteristic (ROC) curve and their accuracy on initial estimates and final decisions. The blue lines and equation text show best-fit regression, and the shaded area gives its standard error. The equations give the regression equation plotted in blue, with bold coefficients being significant at p = .05.

Advisor accuracy As shown in Figure 5.30, the advisors were similarly accurate on average as expected. Nevertheless, some participants experienced in practice 10-20% differences in advisor accuracy (although neither advisor was systematically more accurate across participants).

Advisor agreement Figure 5.31 shows the agreement rates experienced by each participant. All participants experienced a higher agreement rate from the High agreement advisor than from the Low agreement advisor.

Should this break down agreement by initial in/correct as per the experiment design?

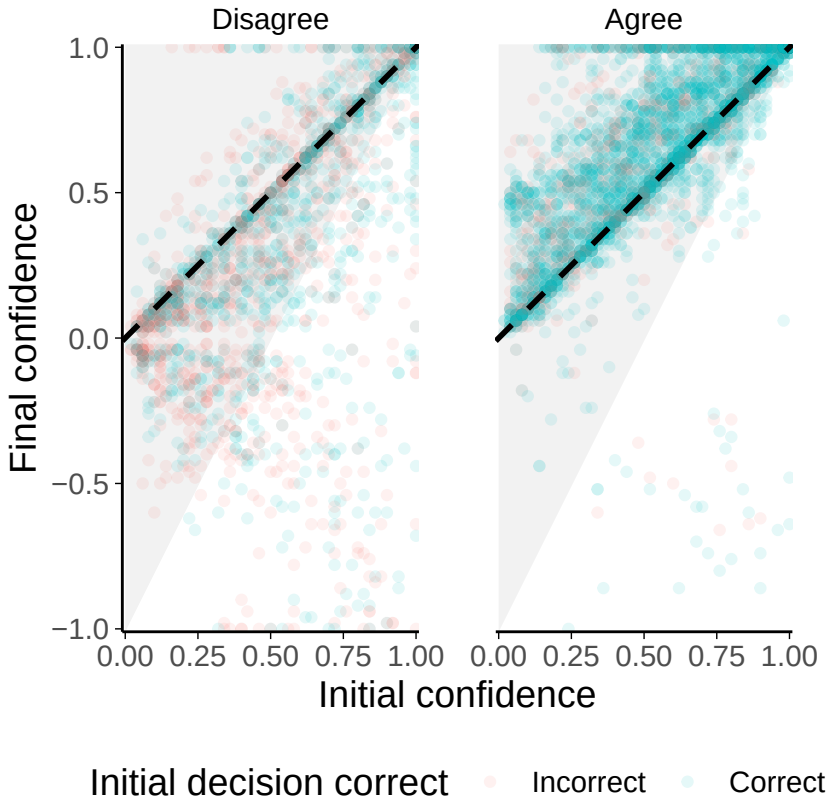


Figure 5.28: Confidence change on the Dots task with in/accurate advisors.
Each point shows the initial and final confidence on a single trial. Final confidence is coded relative to initial confidence so that increasing confidence in the opposite decision (i.e. confidence on trials where the participant changed their mind) is increasingly negative. Points above the dashed y = x line represent increased confidence, while those below it give decreased confidence. Points close to the y = x line indicate relatively little change, while points further away indicate relatively greater change. The shaded grey area shows the zone outside which influence is capped (by moving vertically towards the grey zone boundary) when using the capped influence measure. Agreement and disagreement trials are plotted seperately, with trials coloured according to whether the initial decision was correct.

Advisor influence There were no systematic differences in the influence of the two advisors during the Familiarization phase (Figure 5.32). Some participants were substantially more influenced by one or other advisor, but these extreme cases were also relatively evenly distributed between the advisors.

Hypothesis test {##ac-agr-r-h} The results of the advisor choice (Figure 5.33) were the same as those for the Dots task with High versus Low accuracy

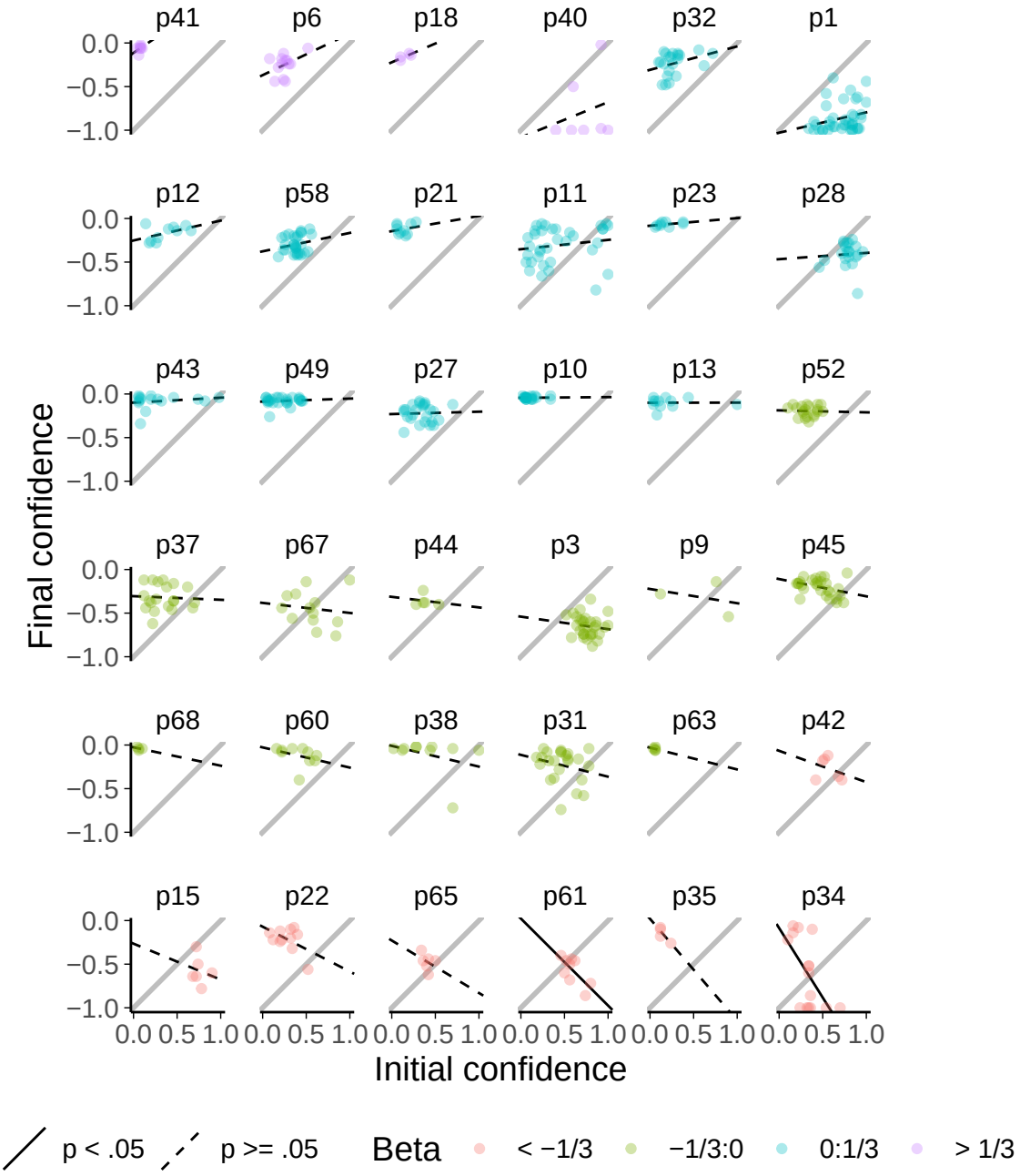


Figure 5.29: Change-of-mind confidence updating on the Dots task.
Each facet shows data from a single participant for trials where they changed their mind on the categorical decision between the initial estimate and the final decision. Participants who never changed their mind are not included. Each point shows the initial and final confidence on a single trial. All final confidence scores are negative because they are coded relative to initial confidence; increasing confidence in the final decision is increasingly negative.
Lines show the best fit for a linear prediction of final from initial confidence, with solid lines indicating that the slope is significantly different from zero at alpha = .05. Points are coloured according to the value of the slope parameter. The grey line is the y = x - 1 line that shows the expected fit line according to the intuitive model of confidence updating.

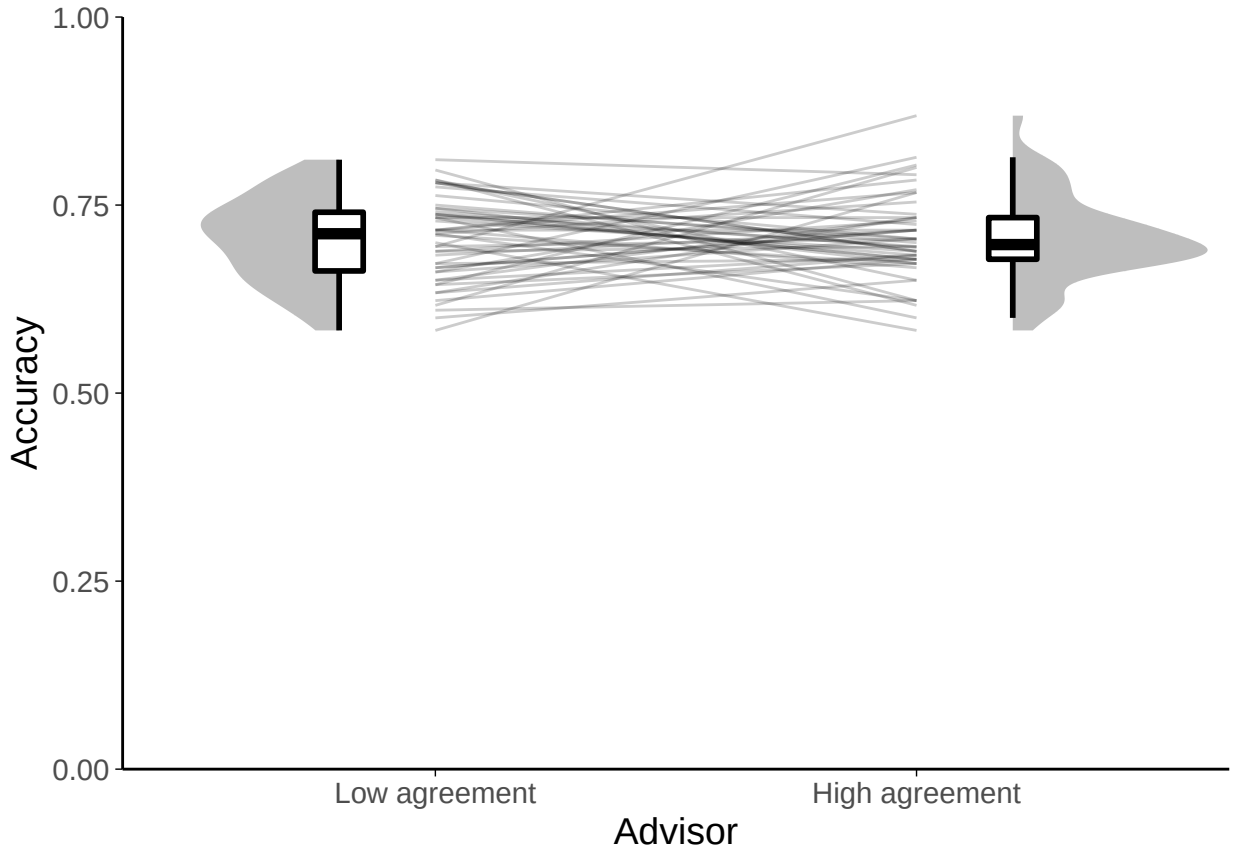


Figure 5.30: Advisor accuracy for Dots task with in/accurate advisors.
Coloured lines show the average accuracy of the advisors as experienced by an individual participant. Box plots and violins show the distribution of the participant means.

advisors (§??). The High agreement advisor was preferred at a rate greater than that expected by chance ($t(49) = 5.43$, $p < .001$, $d = 0.77$, BF = 9.8e3; $M = 0.61$ [0.57, 0.65], $\mu = 0.5$). The modal preference remained at chance, although, as in the accuracy experiment, almost all participants who manifested a preference preferred the High agreement advisor.

5.2.2 Dates Task

Open scholarship practices

`https://osf.io/8d7vg`

!TODO[OSFify data for these studies]

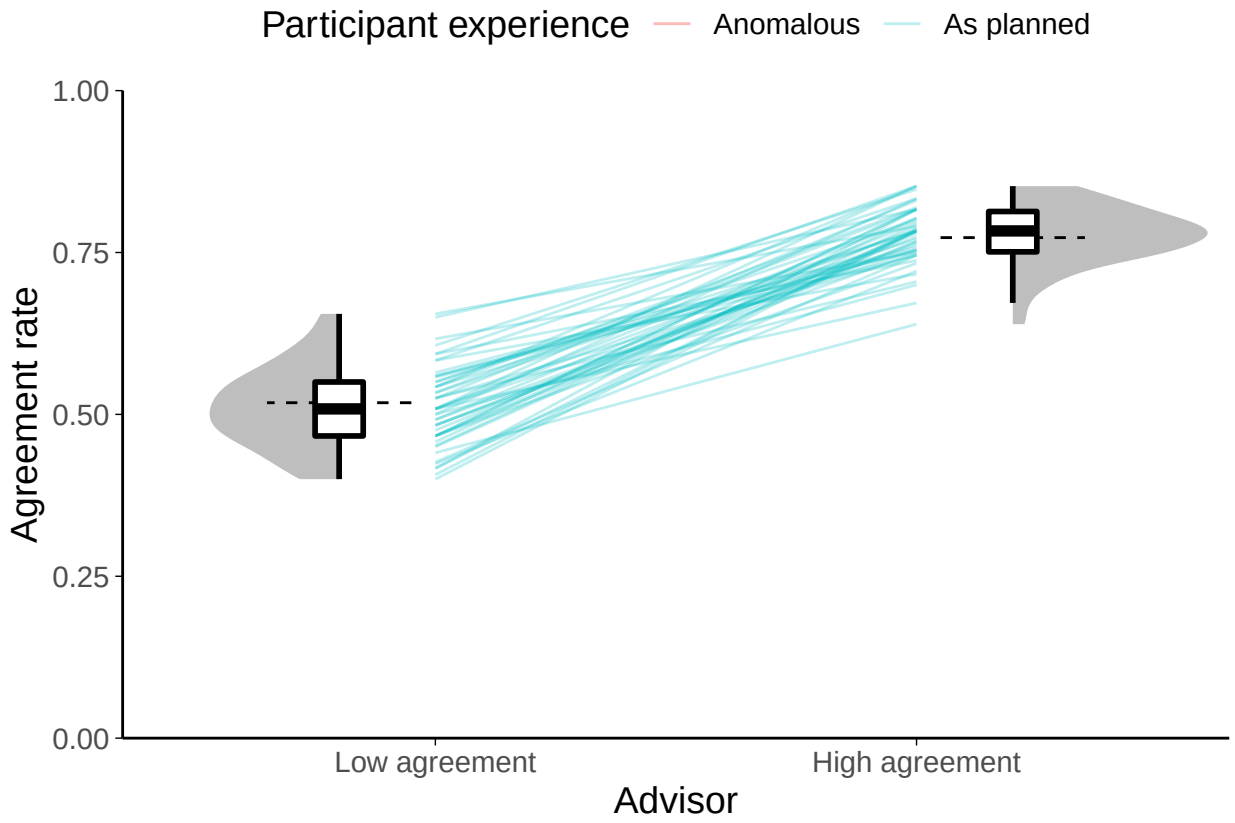


Figure 5.31: Advisor agreement for Dots task with in/accurate advisors.
Faint lines show the average agreement rate of the advisors as experienced by an individual participant. The colour of the line indicates whether the higher agreement advisor agreed with the participant more often as per the experiment design. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the agreement level for the advisors specified in their design.

`https://github.com/oxacclab/ExploringSocialMetacognition/blob/master/ACBin/acc.html`

Unanalysed data Early versions of this experiment (v0-0-1, v0-0-2) included a bug which prevented feedback from being shown during the familiarisation phase even to participants in the Feedback condition. The 16 participants whose data was collected in these versions is not included in analysis. These participants could theoretically be included in the No feedback condition regardless of their condition label in the data, but this is not done here.

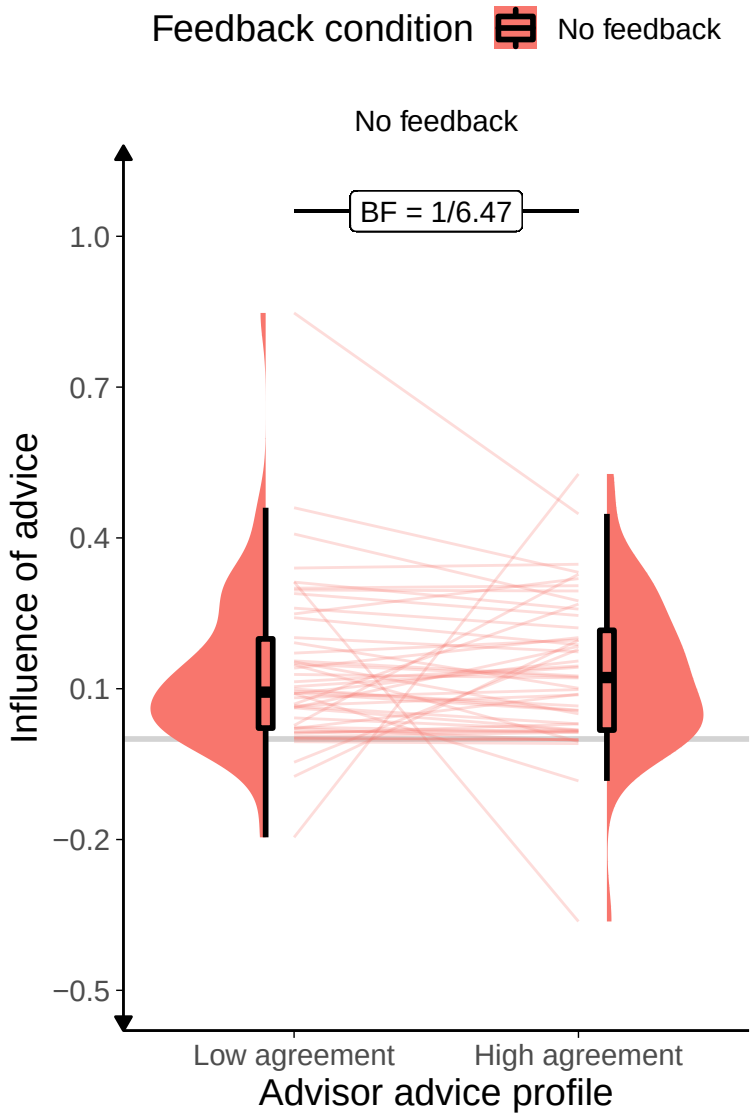


Figure 5.32: Dot task advisor influence for high/low agreement advisors. Participants' weight on the advice for advisors in the Familiarization phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

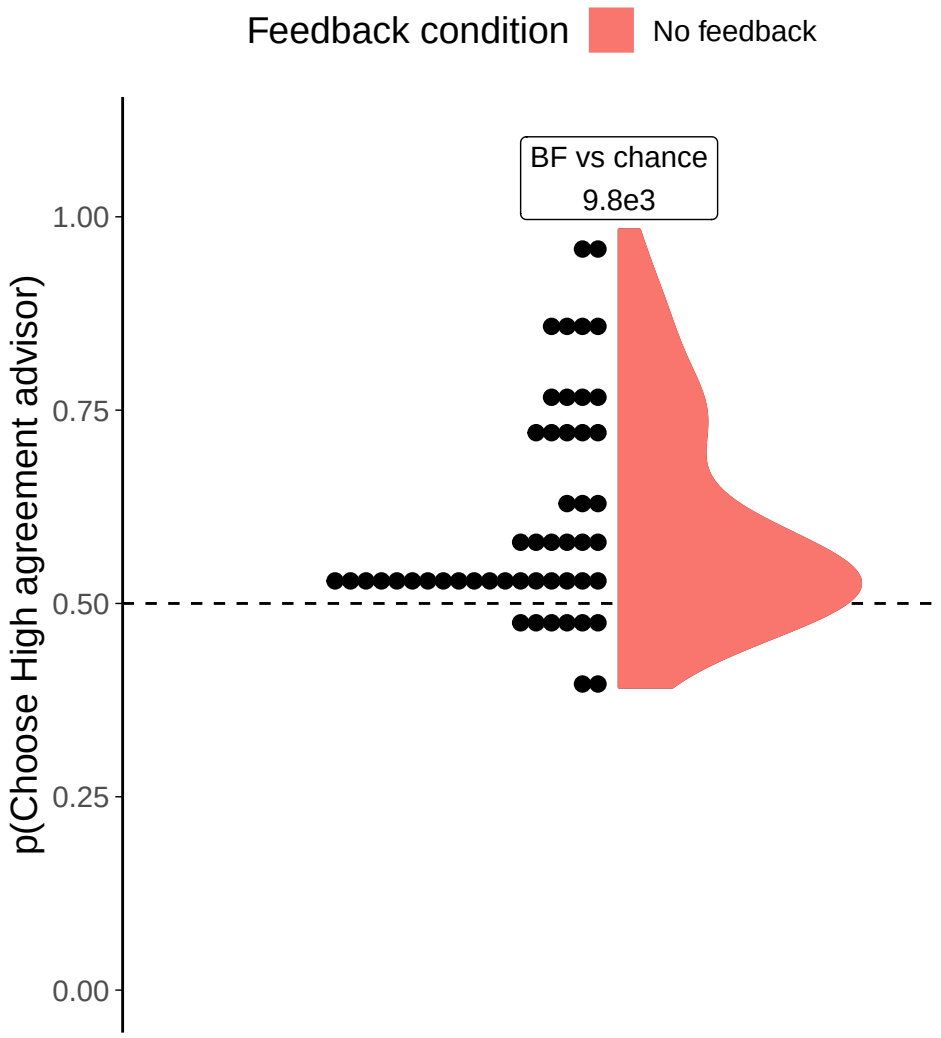


Figure 5.33: Dot task advisor choice for high/low agreement advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

Table 5.7: Advisor advice profiles for Dates task Agreement experiment

	Probability of agreement (%)	
Advisor	Participant correct	Participant incorrect
High agreement	.900	.650
Low agreement	.750	.350

Table 5.8: Participant exclusions for Dates task Agreement experiment

Reason	Participants excluded
Too few trials	0
Insufficient advice taking	0
Too few choice trials	0
Total excluded	**0**
Total remaining	**74**

Method

This study used the binary version of the Dates Task (§3.1.3).

Advice profiles The High agreement and Low agreement advisor profiles issued binary advice (endorsing either the 'before' or 'after' column) probabilistically based on which column the participant had selected in their initial estimate and whether that was the correct answer.

Results

Exclusions Individual trials were screened to remove those that took longer than 60s to complete. Participants were then excluded for having fewer than 11 trials remaining, fewer than 10 trials on which they had a choice of advisor, or for giving the same initial and final response on more than 90% of trials. Overall, 0 participants were excluded, with the details shown in Table 5.8.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants

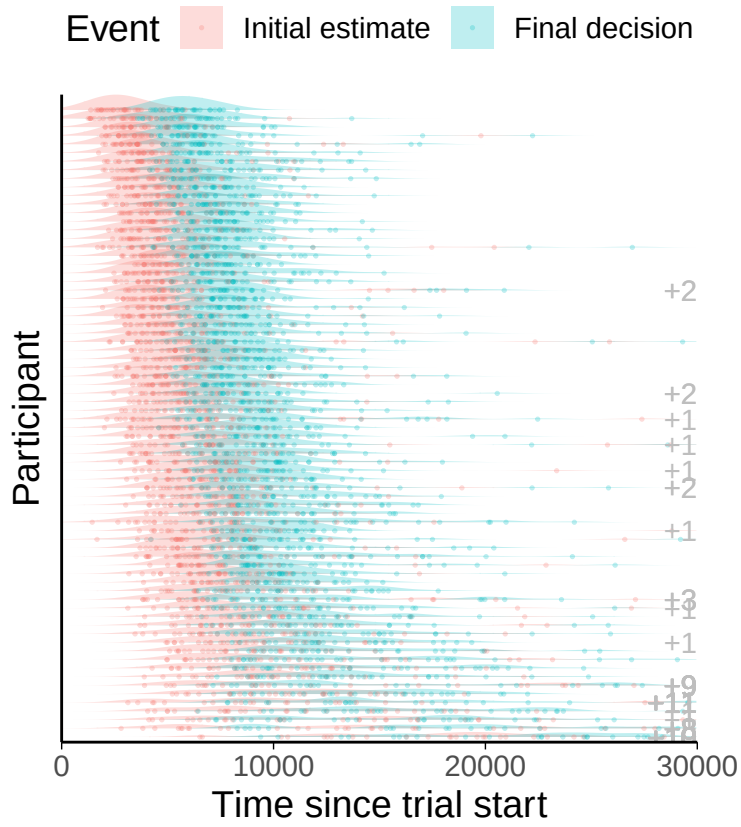


Figure 5.34: Response times for the Dates task with High/Low agreement advisors. Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial and final decisions can be seen in Figure 5.34. As with the Accuracy experiment, when the Dates task response times are compared to the Dots task response times (5.23), they are not only longer, but they are also much more varied within participants.

```
## Picking joint bandwidth of 1130
```

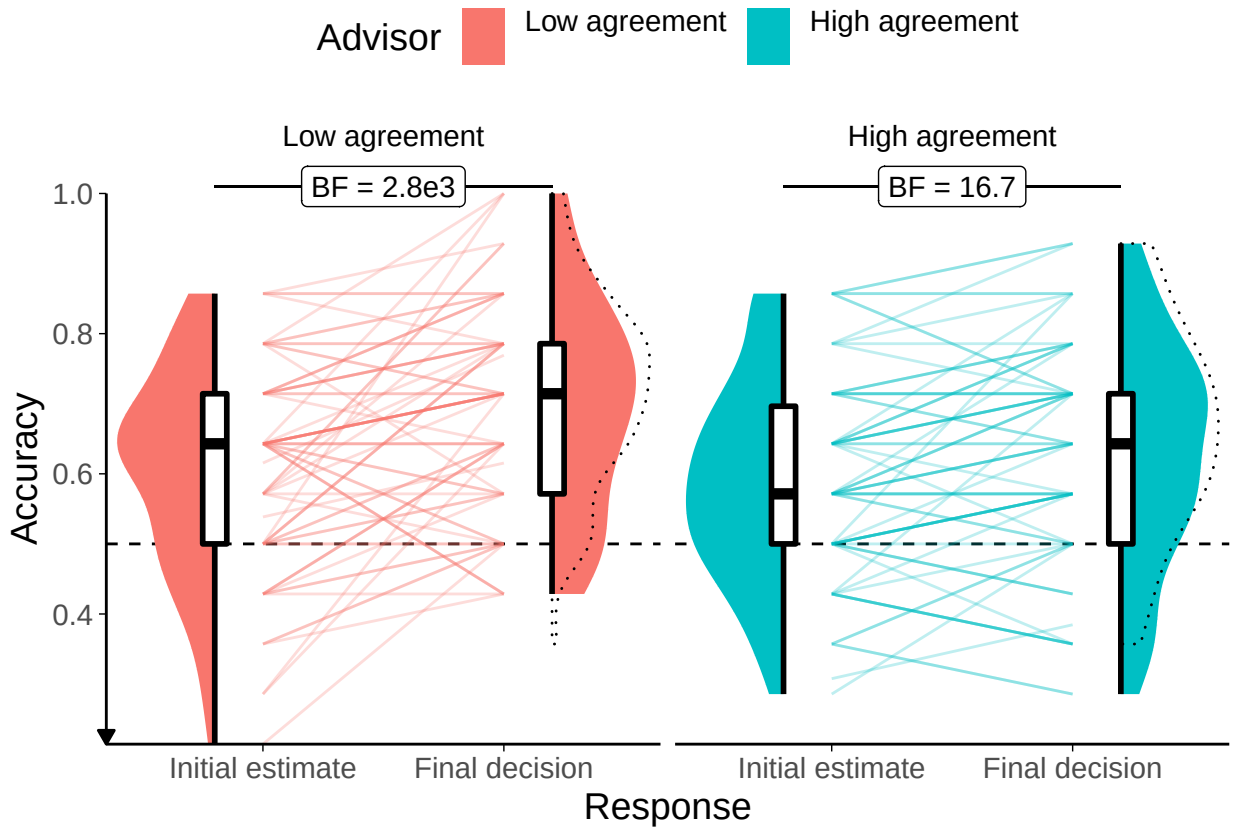


Figure 5.35: Response accuracy for the Dates task with High/Low agreement advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.
Because there were relatively few trials, the proportion of correct trials for a participant generally falls on one of a few specific values. This produces the lattice-like effect seen in the graph. Some participants had individual trials excluded for over-long response times, meaning that the denominator in the accuracy calculations is different, and thus producing accuracy values which are slightly offset from others'.

Accuracy Unlike in the Dots version of the task, participant accuracy is not controlled. Participants managed to improve their performance from their initial estimates to their final decisions with both advisors (Figure 5.35). This is likely because the advisors themselves were more accurate than the participants, so following their advice was generally a good strategy, and the difficulty of the task meant that participants were very willing to be influenced by advice.

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants

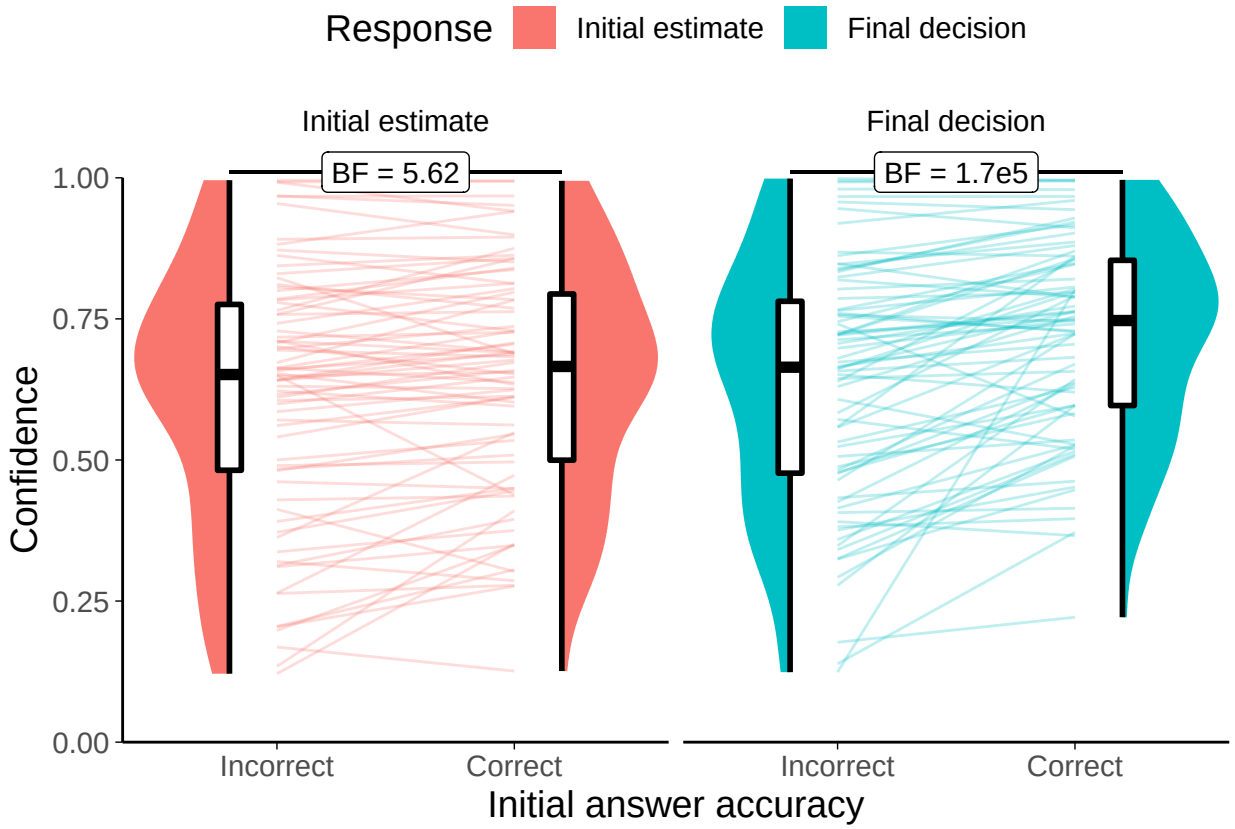


Figure 5.36: Confidence for the Dates task with High/Low agreement advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions.

were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions.

Metacognitive ability The participants' metacognitive abilities were highly variable, with several participants displaying below-chance metacognitive ability (Figure 5.37). While this may appear concerning, recall that metacognitive sensitivity and bias vary substantially and cannot be reliably estimated using ROC curves where performance accuracy on the underlying task is highly variable, so it is not necessarily the case that these values give cause for alarm. The correlation between performance on the underlying task and metacognitive ability (Figure 5.38) shows that, as one might expect, participants with a greater ability to perform the Dates task have a greater insight into their performance on the Dates task.

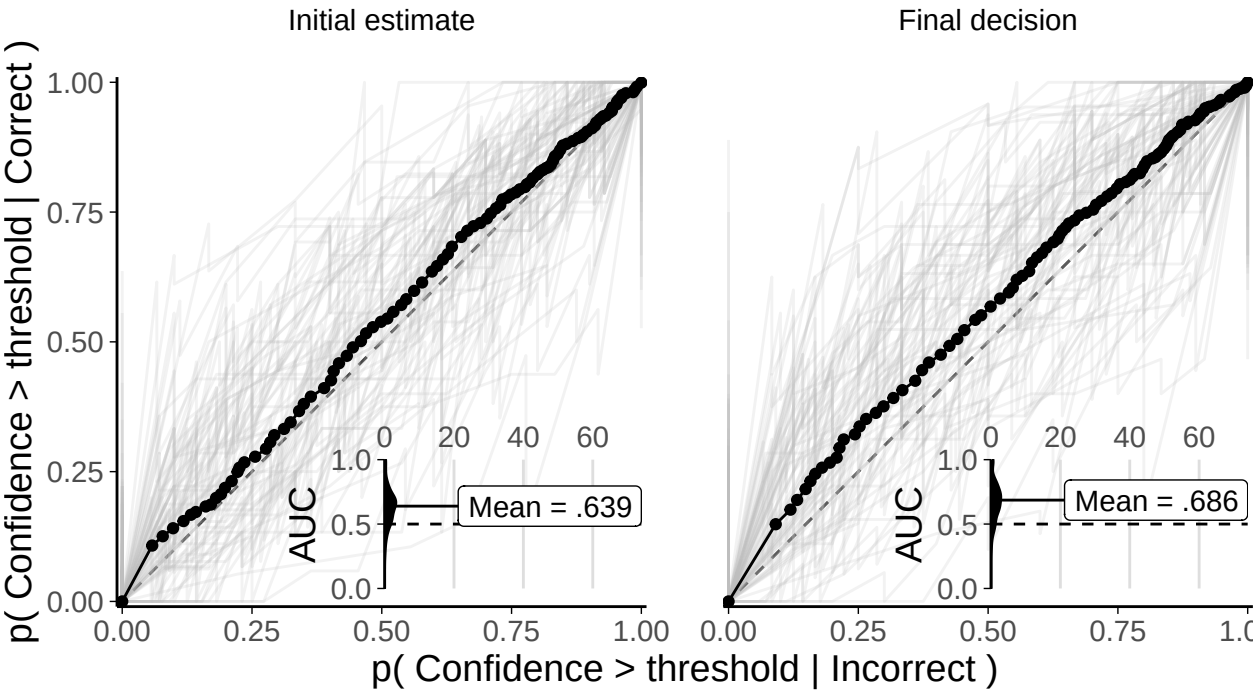


Figure 5.37: ROC curves for the Dates task with High/Low agreement advisors. Faint lines show individual participant data, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers.

This in turn suggests that, despite the low number of trials on the task, we are able to obtain meaningful insights into participants' metacognitive abilities, albeit without being able to precisely estimate the metacognitive sensitivity or bias for an individual participant.

Experience with advisors The advice is generated probabilistically from the rules described previously in Table 5.5. It is thus important to get a sense of the

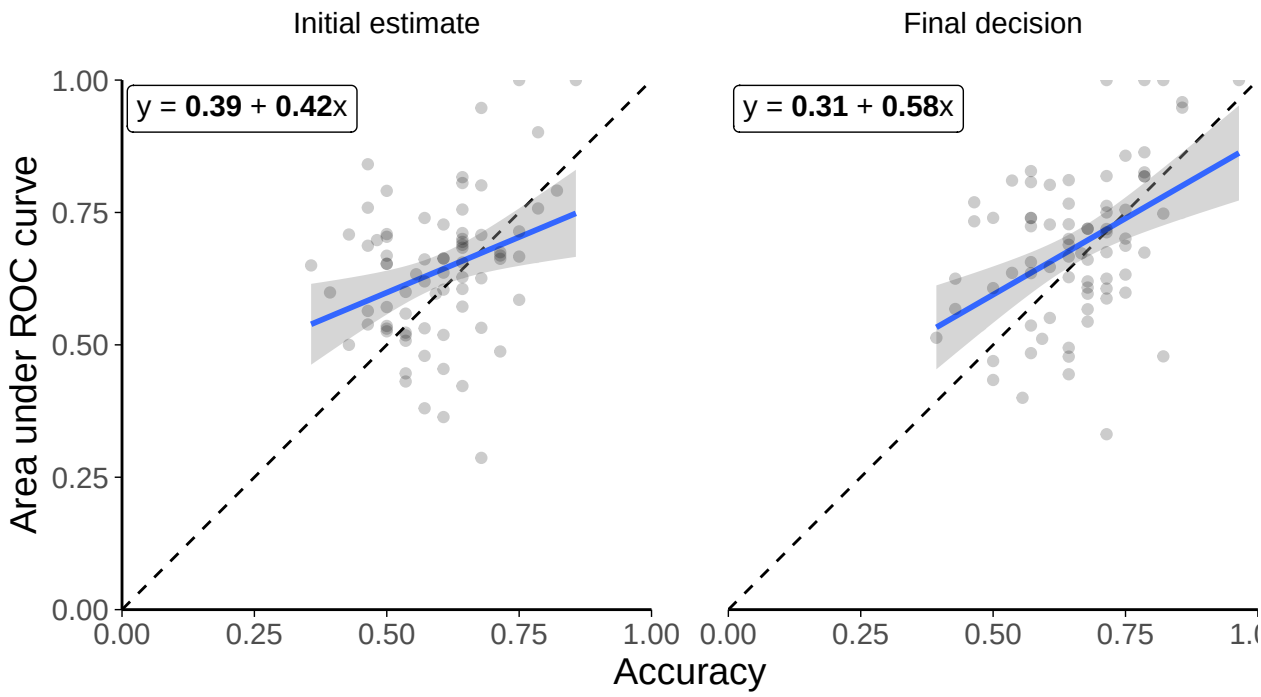


Figure 5.38: AUROC-accuracy correlation for the Dots task with High/Low agreement advisors.
Points show individual participant data for their area under the receiver operator characteristic (ROC) curve and their accuracy on initial estimates and final decisions. The blue lines and equation text show best-fit regression, and the shaded area gives its standard error. The equations give the regression equation plotted in blue, with bold coefficients being significant at p = .05.

actual advice experienced by the participants.

Advisor accuracy As shown in Figure 5.39, the advisors were similarly accurate on average as expected. Nevertheless, there was a broad range of experiences between participants, both in the absolute accuracy of advisors and in their relative accuracies. This variation is due to the relatively low number of trials on which advice was received from each advisor.

Advisor agreement Figure 5.31 shows the agreement rates experienced by each participant. Most participants experienced a higher agreement rate from

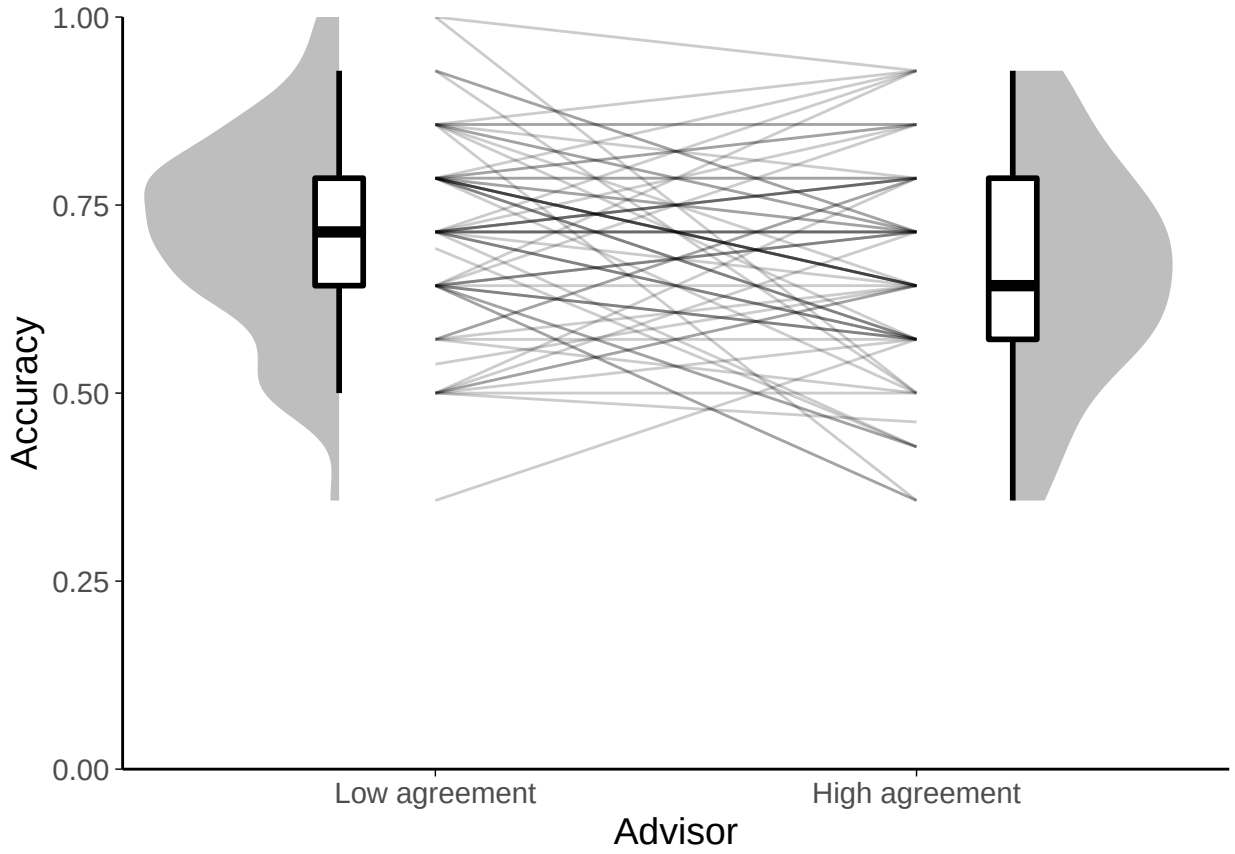


Figure 5.39: Advisor accuracy for Dots task with High/Low agreement advisors. Coloured lines show the average accuracy of the advisors as experienced by an individual participant. Box plots and violins show the distribution of the participant means.

the High agreement advisor than from the Low agreement advisor, as per the manipulation in the experimental design.

Should this break down agreement by initial in/correct as per the experiment design?

Advisor influence The Low agreement was systematically more influential for participants in both Feedback and No feedback conditions. This is presumably due to the fact that disagreeing advice is generally more influential, in part due to the assymetry of the response scale once an initial estimate has been made. Any influence effects in play during the Advisor choice phase may be more pronounced, since these data pertain to the Familiarization phase only, in which participants were still learning about the usefulness of their advisors.

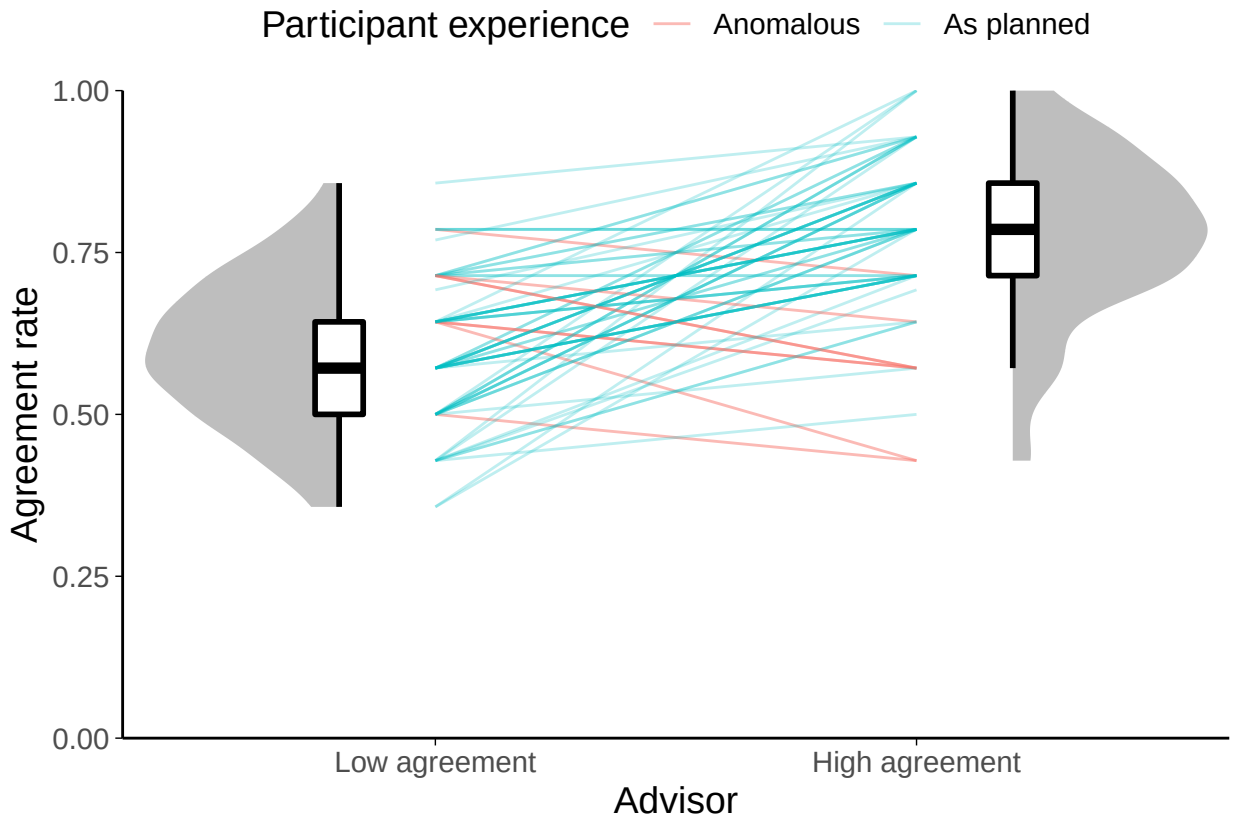


Figure 5.40: Advisor agreement for Dots task with High/Low agreement advisors. Faint lines show the average agreement rate of the advisors as experienced by an individual participant. The colour of the line indicates whether the higher agreement advisor agreed with the participant more often as per the experiment design. Box plots and violins show the distribution of the participant means.

```
## Warning: Problem with ‘mutate()‘ input ‘bf‘.
## i data coerced from tibble to data frame
## i Input ‘bf‘ is ‘map(...)‘.
## Warning: data coerced from tibble to data frame
## Warning: Problem with ‘mutate()‘ input ‘bf‘.
## i data coerced from tibble to data frame
## i Input ‘bf‘ is ‘map(...)‘.
## Warning: data coerced from tibble to data frame
```

Hypothesis test

```
## Warning: data coerced from tibble to data frame
```

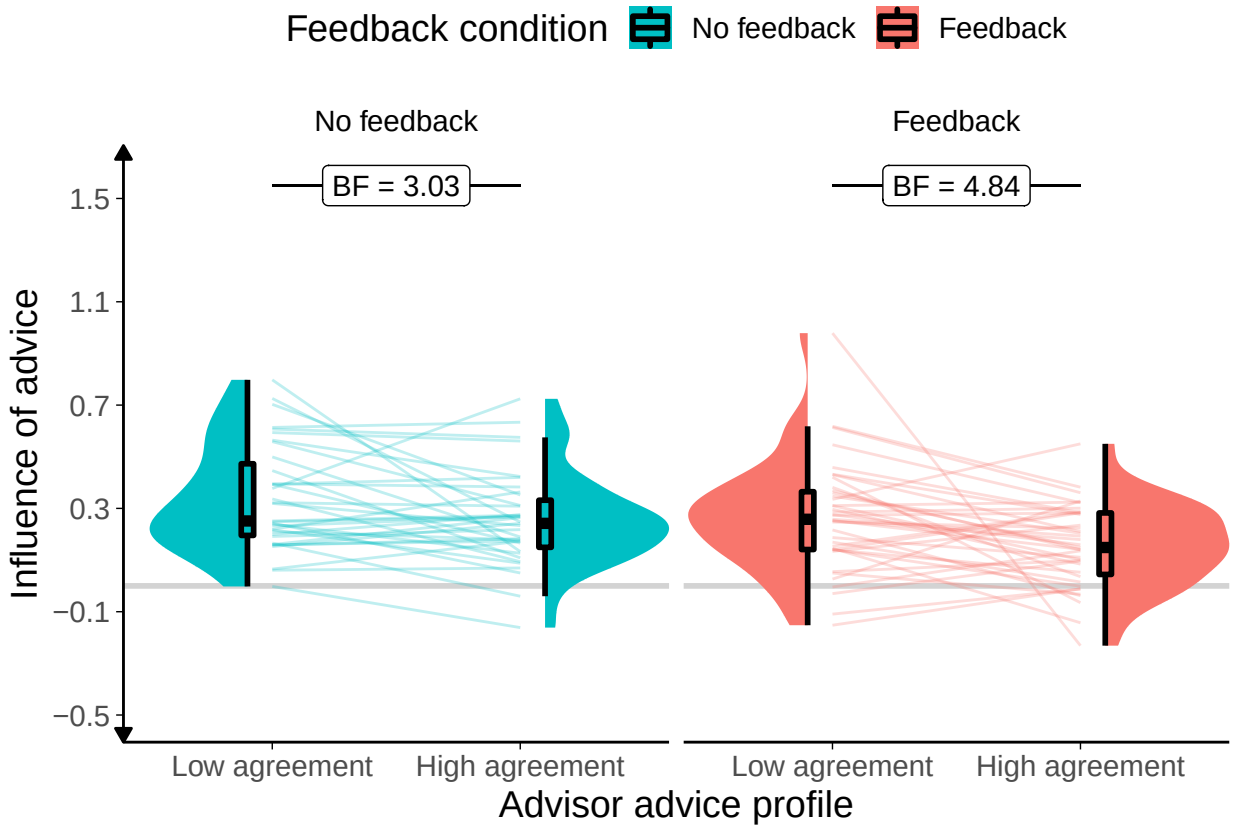


Figure 5.41: Date task advisor influence for High/Low agreement advisors. Participants' weight on the advice for advisors in the Familiarization phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

Consistent with the result from the Dots task, in the No feedback condition participants' preferences for receiving advice from the High agreement advisor were greater than chance ($t(34) = 2.62$, $p = .013$, $d = 0.44$, BF = 3.39; $M_{Nofeedback} = 0.63$ [0.53, 0.73], $\mu = 0.5$). The modal preference was to select the High agreement advisor on every Choice trial, and although some participants still showed a preference for hearing advice from the Low agreement advisor, preferences for the High agreement advisor were generally stronger and more frequent (Figure 5.42).

In the Feedback condition, the mean of the participants' selection rates was equivalent to random picking ($t(38) = 0.46$, $p = .648$, $d = 0.07$, BF = 1/5.24; $M_{Feedback} = 0.52$ [0.42, 0.62], $\mu = 0.5$). This is consistent with a strategy which attempts to maximise the accuracy of final decisions, because neither advisor would

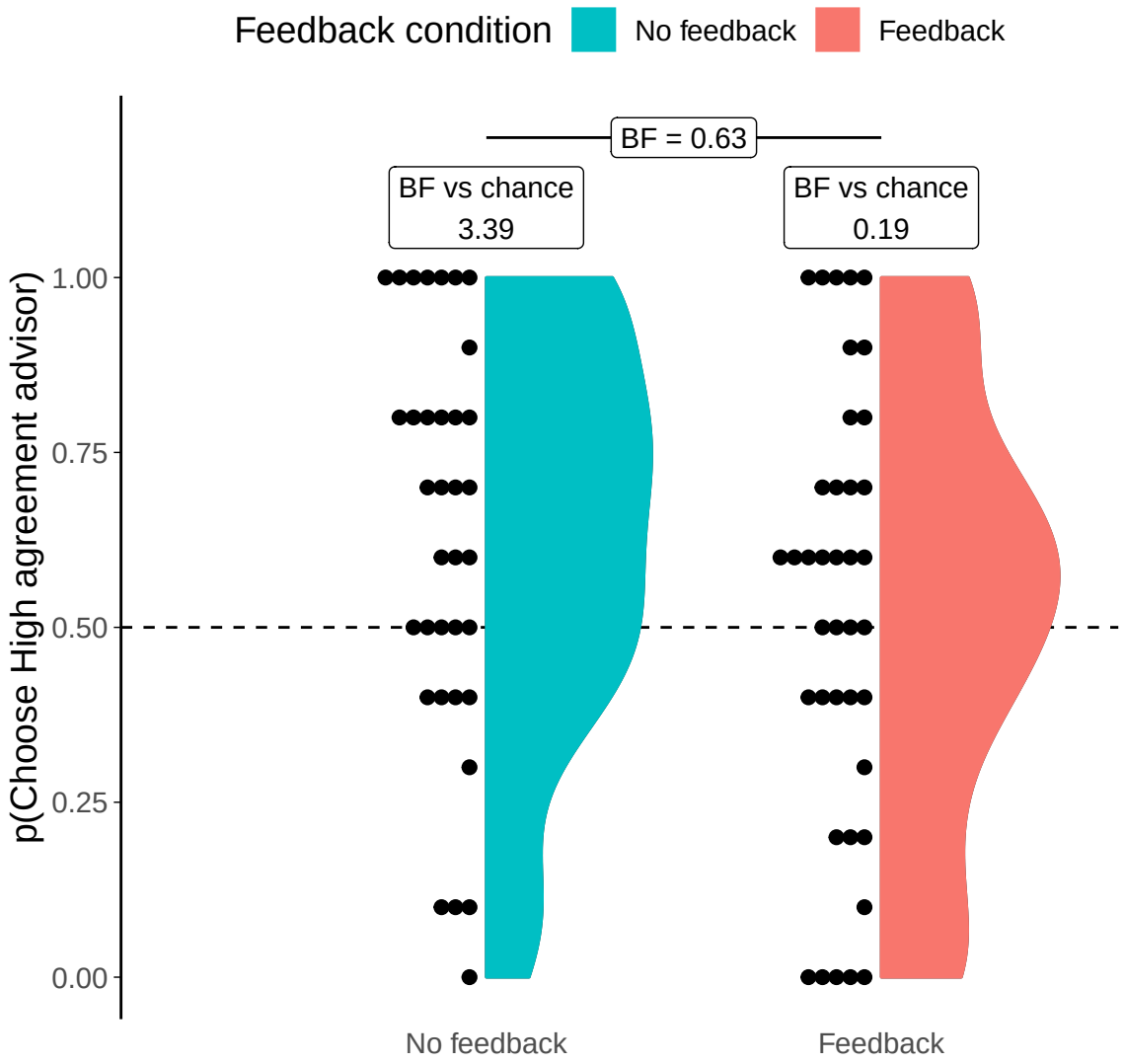


Figure 5.42: Dates task advisor choice for High/Low agreement advisors. Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line. Participants in the Feedback condition received feedback during the Familiarization phase, but not during the Choice phase.

help with this task systematically. The null result here does indicate, however, that there is no strong and clear preference for agreement over and above its accuracy benefits.

5.2.3 Discussion

Where feedback is provided on advisors' performance, participants prefer high agreement advisors to low agreement advisors.

At least in the Dates task the low agreement advice is probably more interesting - we already know what the high agreement advisor is going to say!

re: influence - An alternative explanation for the data is that the 'offbrand' advice selected for analysis was more surprising when coming from the advisor who usually agrees, and that this increased salience increased its influence rather than increased trust in the advisor.

If that were the case then the same ought to be true for agreeing advice from a disagreeing advisor, and this is not the case \mccorrect{!TODO[Do we have/can we produce evidence for this? Perhaps look at our participants where advisors disagree with them by chance more than e.g. 70% of the time and check for increased influence on agreement trials vs advisors who disagree less than 70% of the time]}.

5.3 Accuracy vs. agreement (Date estimation)

The High versus Low agreement advisor experiment showed that participants tended to prefer to receive agreeing advice when they did not receive feedback. The advisors did not differ in their accuracy (by design), which meant that participants could not increase their performance by selecting one advisor over another. Here we introduce a discrepancy between advisors' objective performance (accuracy) and their subjective performance from the judge's perspective (agreement). By playing off accuracy against agreement we can explore whether participants continue to prefer agreement when there is a cost associated with agreement through a reduction in overall accuracy. We expect that participants provided with feedback

will be able to identify the accurate advisor and ignore the agreeing advisor, while participants denied feedback will gravitate towards the agreeing advisor who, from their perspective, appears more accurate.

5.3.1 Dots Task

Open scholarship practices

 `https://osf.io/f3k4x`

 !TODO[OSFify data for these studies]

 `https://github.com/oxacclab/ExploringSocialMetacognition/blob/c18c26b5da3622988e2261433cf256aae4d19f39/AdvisorChoice/ava.html`

Unanalysed data Two versions of this experiment have unanalysed data (Pilot data, v1 Mixed design). The Pilot data was collected to ensure the study functioned properly. The v1 Mixed design data was conducted and run as a proper experiment in which participants learned about both advisors simultaneously (preregistered at `https://osf.io/5z2fp`), but there were no effects in the data. We suspected the absence of effects was because participants had difficulty distinguishing the advisors when they were presented together. The version of the experiment reported here presented one advisor per block in the familiarisation phase. The 69 participants whose data was collected in these versions is not included in analysis.

Method

!TODO[clarify any methodological differences from the main methods chapter]

Advice profiles The two advisor profiles used in the experiment were High accuracy and High agreement. The advisor profiles were not balanced for overall agreement or accuracy rates.

Table 5.9: Advisor advice profiles for Dots task Accuracy/agreement experiment

	Probability of agreement			
Advisor	Participant correct	Participant incorrect	Overall	Overall accuracy
High accuracy	.800	.200	.626	.800
High agreement	.800	.800	.800	.626

Table 5.10: Participant exclusions for Dots task Accuracy vs Agreement experiment

Reason	Participants excluded
Accuracy too low	1
Accuracy too high	0
Missing confidence categories	6
Skewed confidence categories	18
Too many participants	14
Total excluded	**39**
Total remaining	**50**

Results

Exclusions Participants' data could be excluded from analysis where they have an average accuracy below 0.6 or above 0.85, do not have trials in all confidence categories, have fewer than 12 trials in each confidence category, or finish the experiment after 50 participants have already submitted data which passed the other exclusion tests. Overall, 39 participants were excluded, with the details shown in Table 5.10.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times

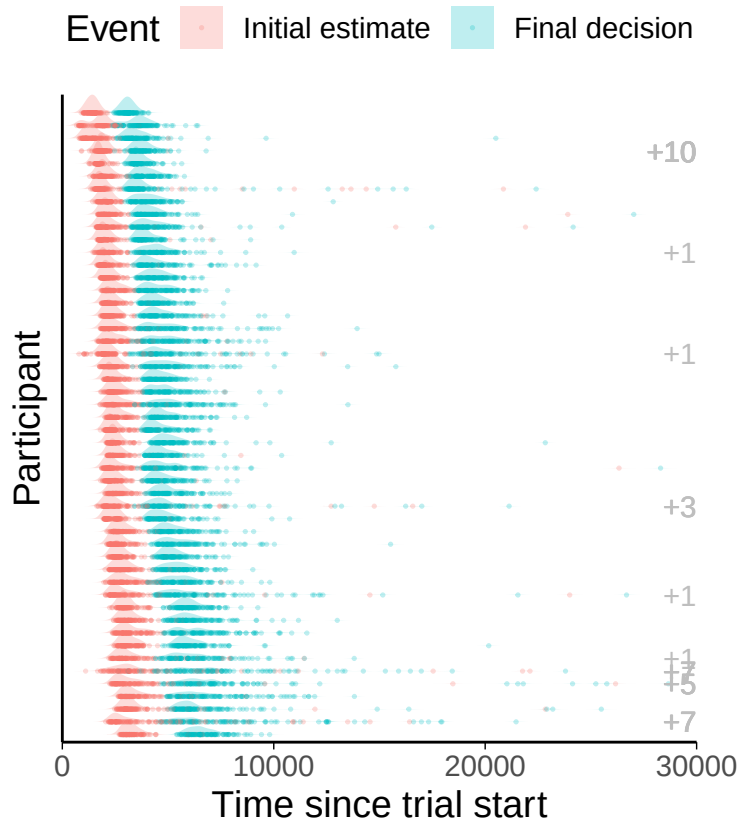


Figure 5.43: Response times for the Dots task with high accuracy/agreement advisors. Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

for both initial and final decisions can be seen in Figure 5.43.

```
## Picking joint bandwidth of 224
```

All participants had similar patterns: initial and final responses were approximately normally distributed, with final responses having a higher variance (because they include the variance for the initial response). Most participants show some trials on which initial or final responses took substantially longer than usual.

!TODO[Perhaps this plot would be better showing individual participants' distributions and box-plots/3SD markers, especially if we want to exclude trials on the basis of taking too long (we don't currently). Perhaps tying final response time to final response start would be better, too, because then initial and final

decisions can be more sensibly compared.]

Accuracy Accuracy of initial decisions was controlled by a staircasing procedure which aimed to pin accuracy to 71%. The accuracy of final decisions was free to vary according to the ability of the participant to take advantage of the advice on offer. As Figure 5.44 shows, participants' accuracy scores for initial decisions were close to the target values (partly because participants whose accuracy scores diverged considerably were excluded). Participants tended to improve the accuracy of their responses following advice from High accuracy advisors, while the evidence was unclear as to whether there was any difference in response accuracy with High agreement advice. The distributions of participant accuracy for trials with the High agreement advisor are slightly unusual, with a bimodal structure, although the medians are somewhat higher than the target value. There is no obvious reason why this should be the case.

The bimodal structure seems to show up quite a bit - perhaps there is a reason for it?

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions.

Metacognitive ability As shown by Figure 5.46, most participants showed above-chance metacognitive sensitivity for initial estimates and final decisions. Participants generally showed higher metacognitive sensitivity for final decisions, although this may be an artefact of a change in metacognitive bias. Participants' metacognitive sensitivity was not particularly high !TODO[What are typical values we might expect in the dots task and similar tasks? Is there a useful mapping between meta-d' and Type II ROC to compare with e.g. Roualt's stuff?]. Somewhat surprisingly, participant metacognition as captured by the area under the ROC

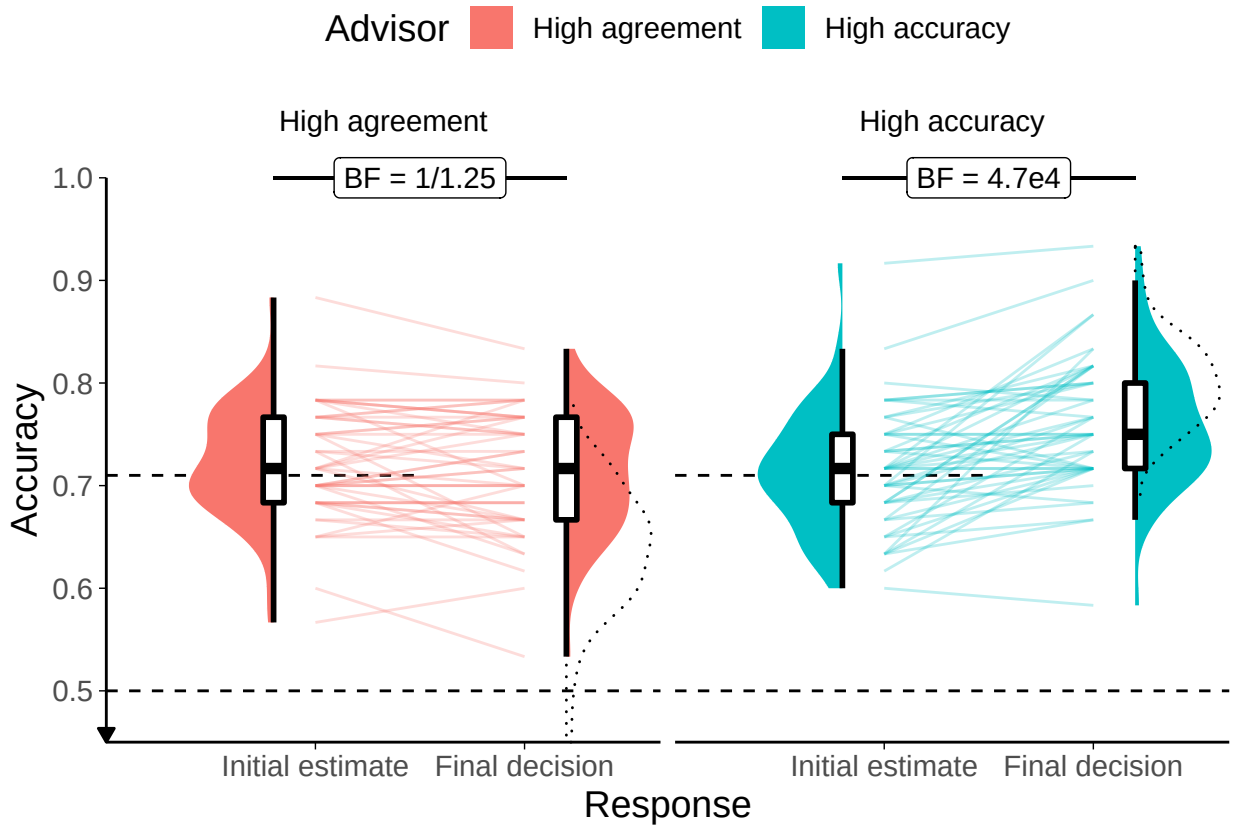


Figure 5.44: Response accuracy for the Dots task with high accuracy/agreement advisors.
Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

curve was significantly correlated with participant performance for final decisions. This is somewhat understandable because the accuracy of final decisions was not held constant in the way that the accuracy of initial estimates was.

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

Experience with advisors The advice is generated probabilistically from the rules described previously in Table 5.9. It is thus important to get a sense of the actual advice experienced by the participants.

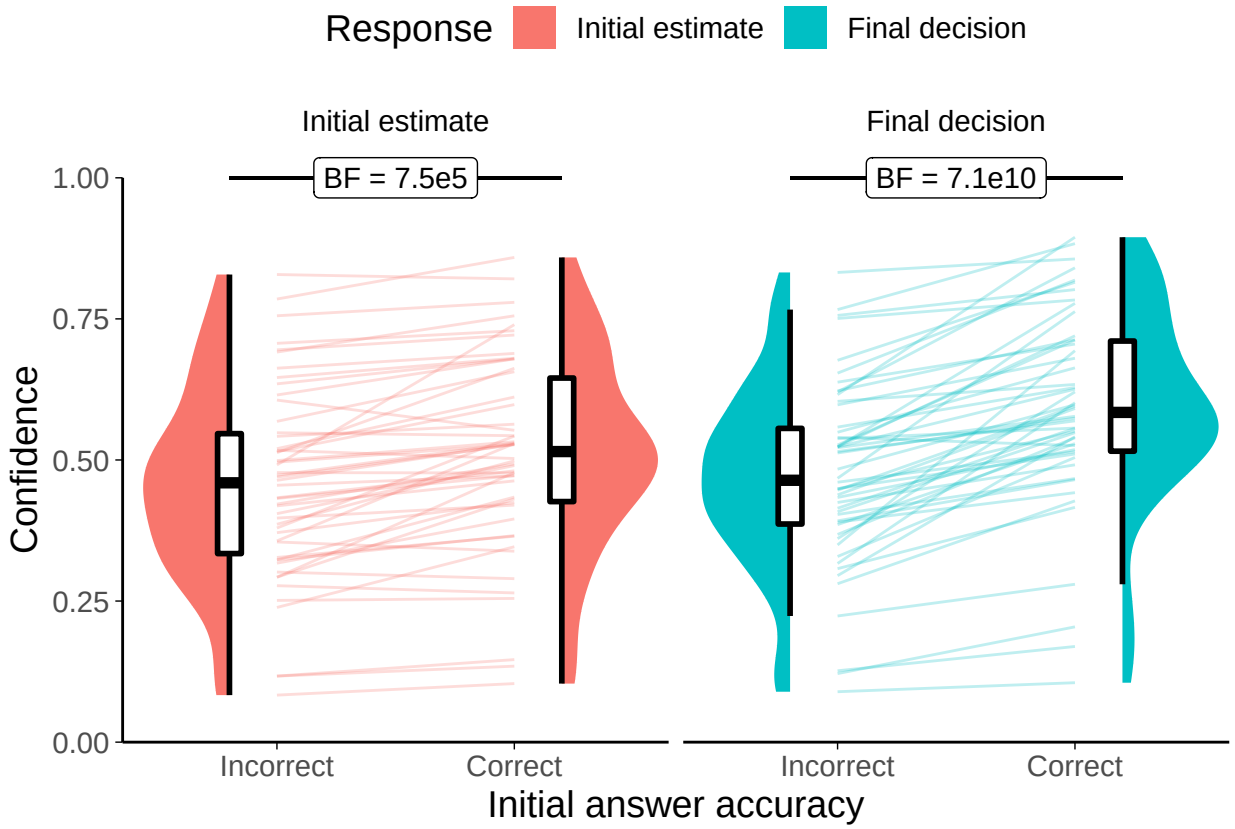


Figure 5.45: Confidence for the Dots task with high accuracy/agreement advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions.

Advisor accuracy As shown in Figure 5.48, all participants experienced the High accuracy advisor as providing more accurate advice than the High agreement advisor, as intended in the experiment design.

Advisor agreement Figure 5.49 shows the agreement rates experienced by each participant. All participants but one experienced a higher agreement rate from the High agreement advisor than from the High accuracy advisor. Together with the unanimous experience of higher accuracy from the High accuracy advisor, this indicates that the manipulation was effective for almost all participants individually, as well as for the sample on average.

Should this break down agreement by initial in/correct as per the experiment design?

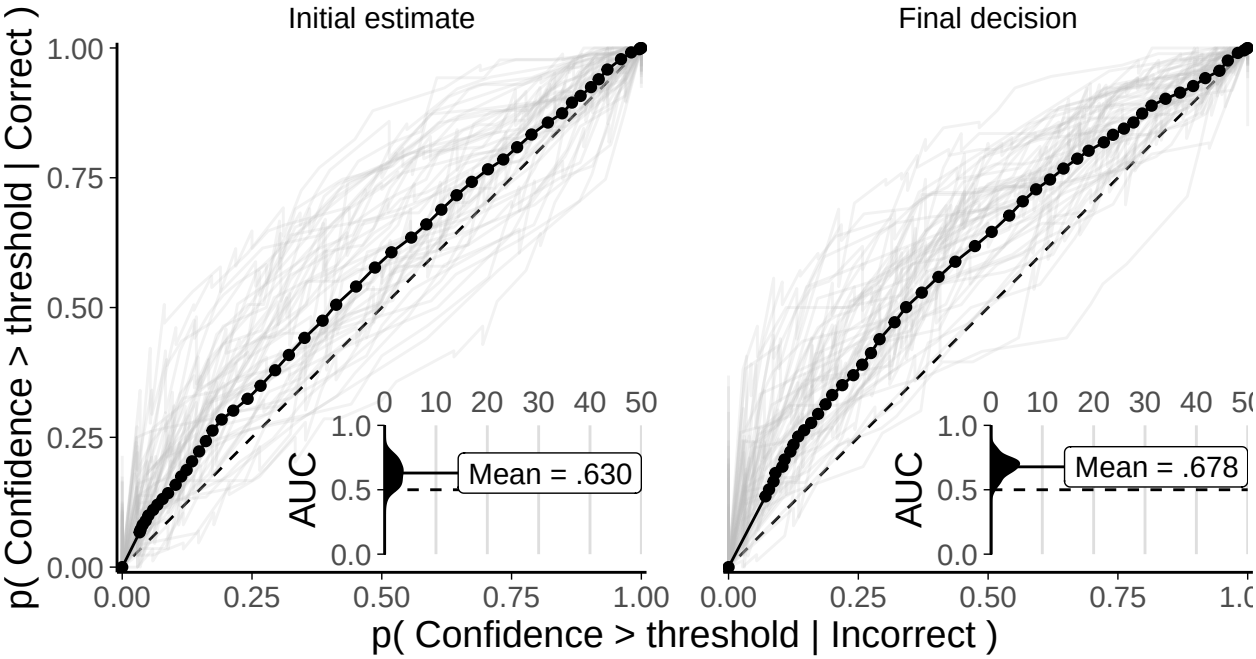


Figure 5.46: ROC curves for the Dots task with high accuracy/agreement advisors. Faint lines show individual participant data, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers.

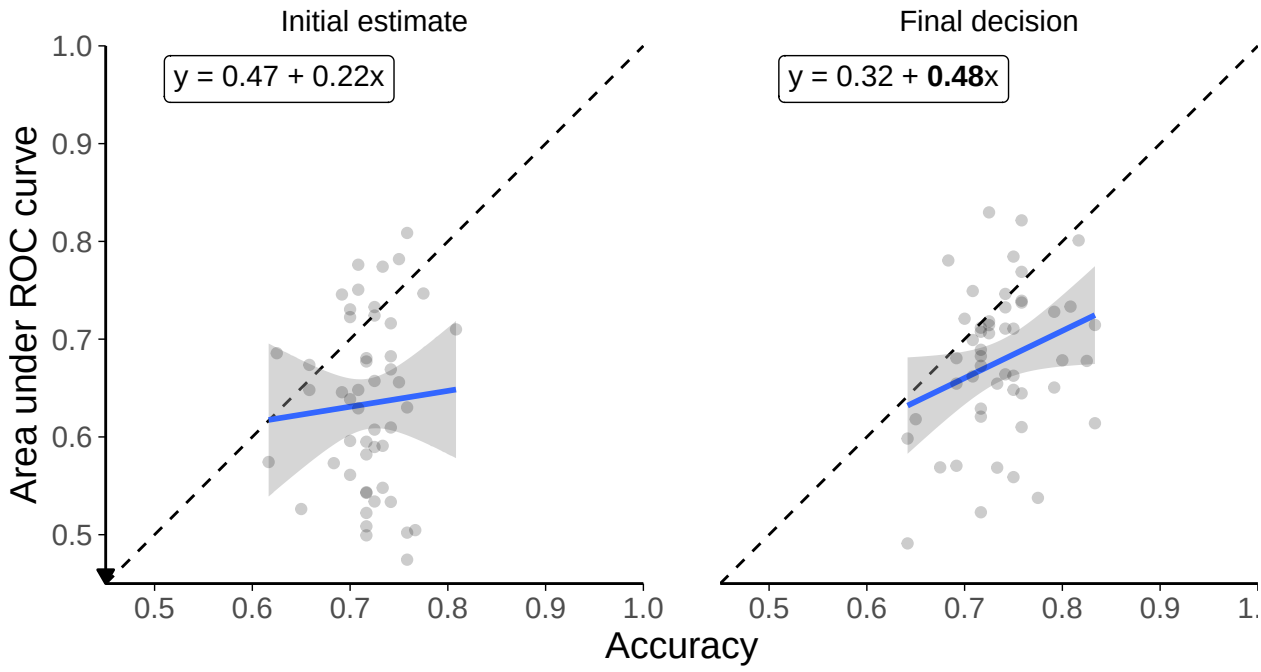


Figure 5.47: AUROC-accuracy correlation for the Dots task with high accuracy/agreement advisors.
Points show individual participant data for their area under the receiver operator characteristic (ROC) curve and their accuracy on initial estimates and final decisions. The blue lines and equation text show best-fit regression, and the shaded area gives its standard error. The equations give the regression equation plotted in blue, with bold coefficients being significant at p = .05.

Advisor influence Figure 5.50 shows that the participants were consistently more influenced by the High accuracy advisor during the Familiarization phase of the experiment. Despite not receiving feedback, participants learned to adjust their responses more in light of answers from the High accuracy advisor. The differences for individual participants are mostly small, but the direction is quite consistently in favour of the High accuracy advisor.

Hypothesis test Despite the influence differences observed above, Figure 5.51 shows that there was no consistent picking preference in favour of either the High accuracy or the High agreement advisor. While several participants did develop

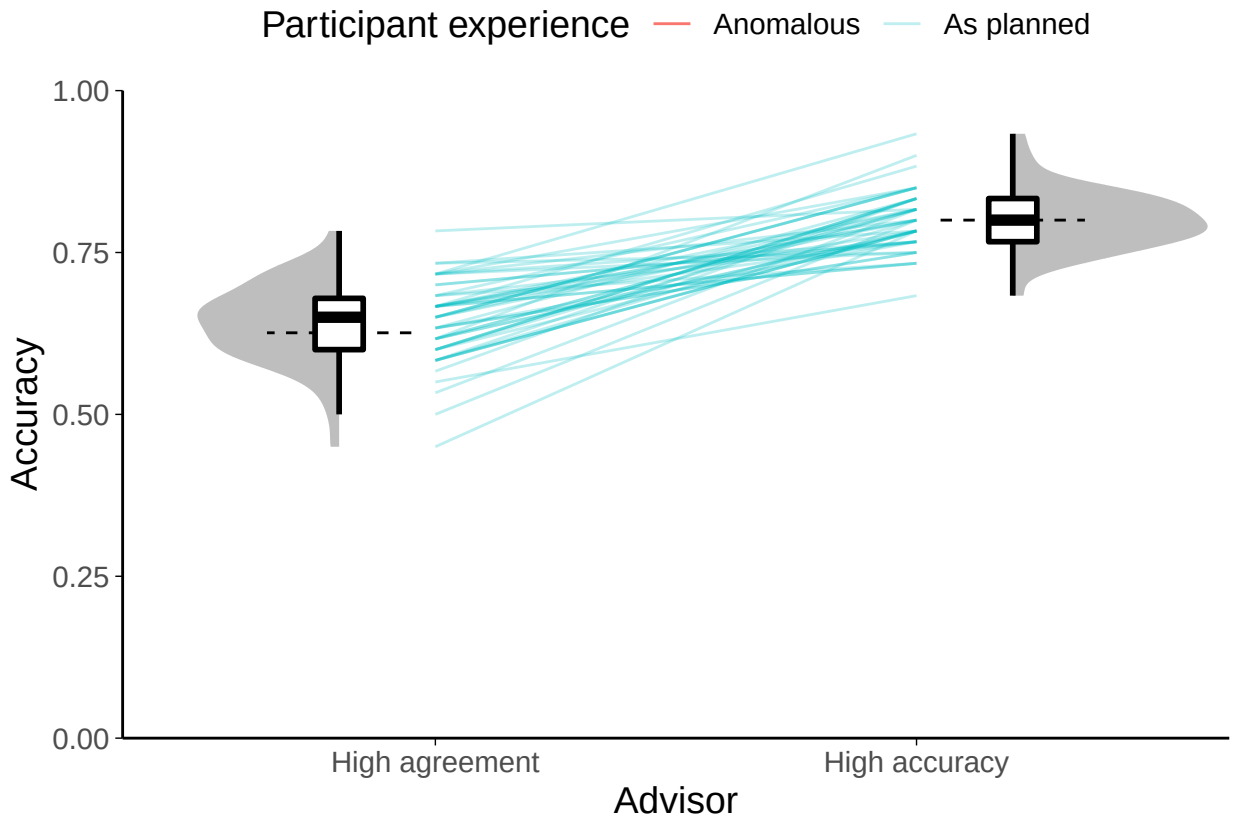


Figure 5.48: Advisor behaviour for Dots task with high accuracy/agreement advisors. Coloured lines show the average accuracy of the advisors as experienced by an individual participant. The colour of the line indicates whether the more accurate advisor was more accurate as per the experiment design. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the accuracy level for the advisors specified in their design.

very strong preferences, picking one or the other advisor nearly all the time, these preferences were not systematically oriented towards either advisor ($t(49) = 0.95$, $p = .345$, $d = 0.13$, BF $= 1/4.23$; $M = 0.54$ [0.45, 0.63], $\mu = 0.5$).

5.3.2 Dates Task

Open scholarship practices

`https://osf.io/nwmx5`

!TODO[OSFify data for these studies]

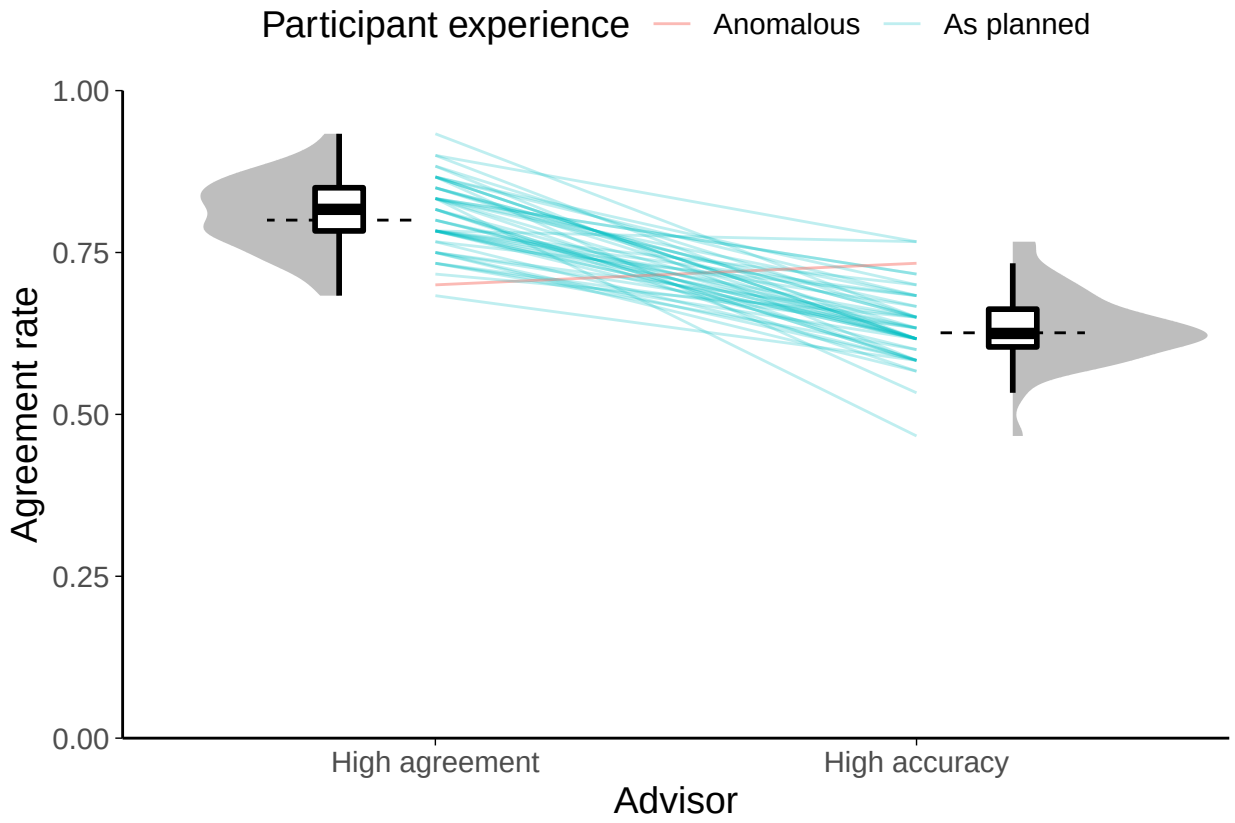


Figure 5.49: Advisor behaviour for Dots task with high accuracy/agreement advisors. Faint lines show the average agreement rate of the advisors as experienced by an individual participant. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the agreement level for the advisors specified in their design.

https://github.com/oxacclab/ExploringSocialMetacognition/blob/ed13951c488e1996df7ff53d48629843bacfd074/ACv2/ac.html

Unanalysed data The first version of this experiment (v0-0-5) included a bug in which the advisor choice options were not recorded, making it difficult or impossible to work out which trials included a choice of advisor. The second version (v0-0-6) worked perfectly, and the data reported below belong to a preregistered replication of this version. The 42 participants whose data was collected in these versions is not included in analysis. The v0-0-5 participants could be included in a study which was agnostic about advisor choice, and the v0-0-6 participants could be included in an aggregated analysis, but here I report only the preregistered replication.

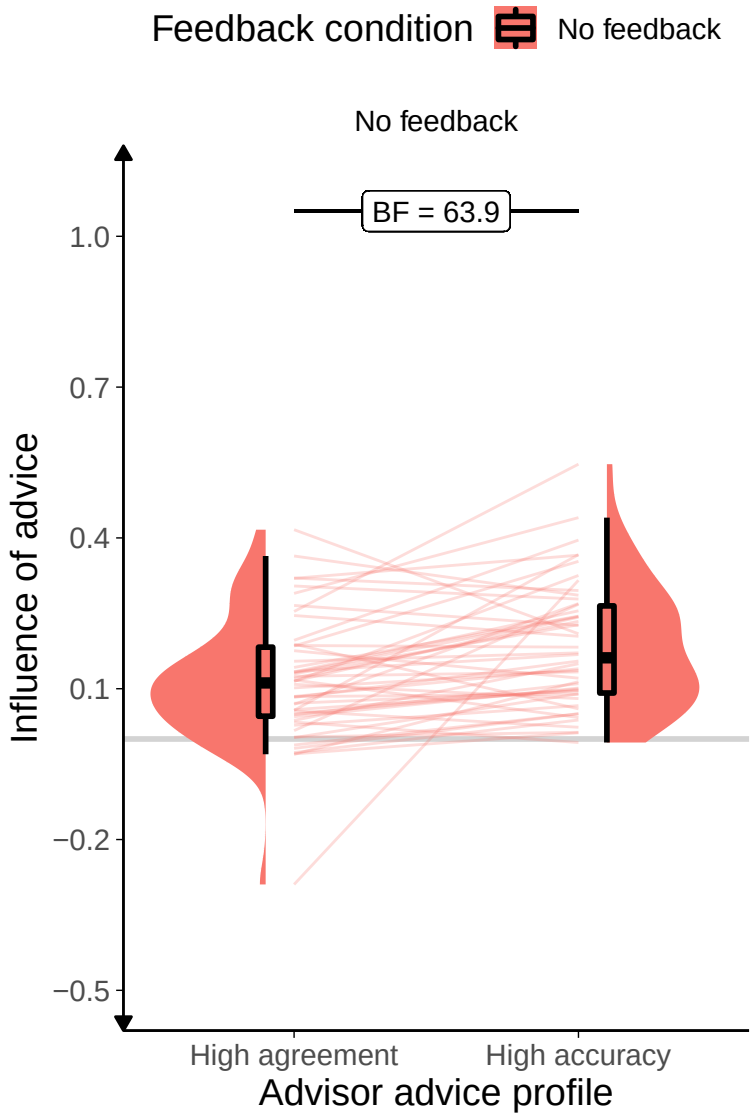


Figure 5.50: Dot task advisor influence for high accuracy/agreement advisors. Participants' weight on the advice for advisors in the Familiarization phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

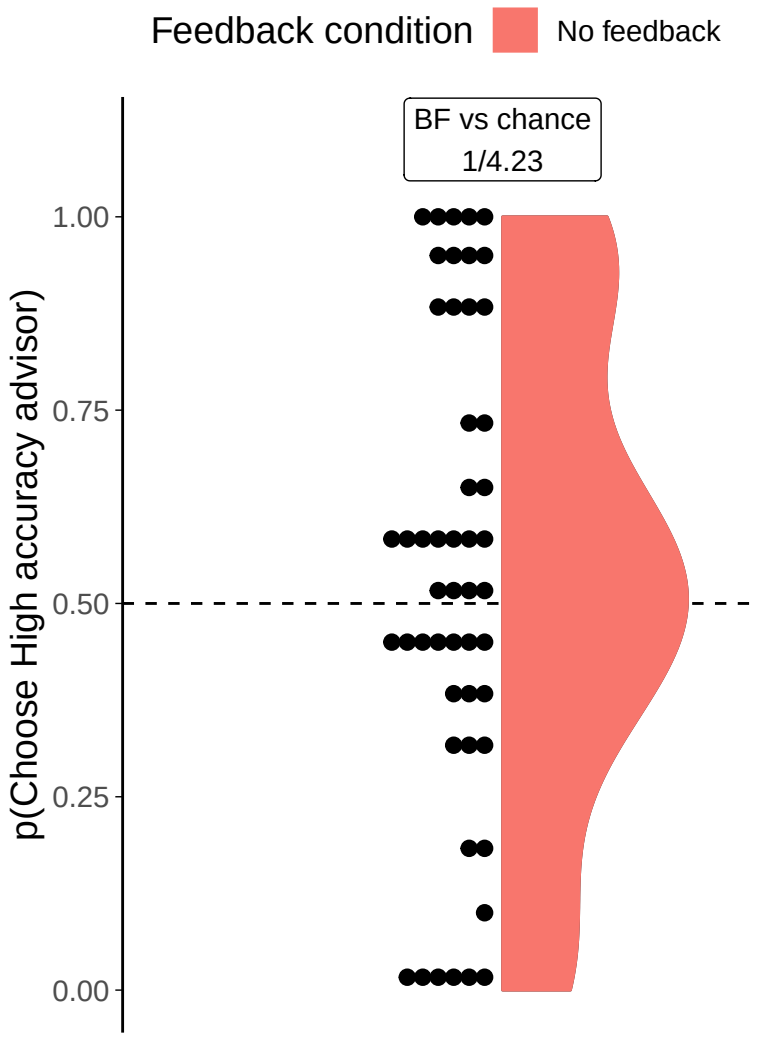


Figure 5.51: Dot task advisor choice for high accuracy/agreement advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

Table 5.11: Participant exclusions for Dates task Accuracy vs Agreement experiment

Reason	Participants excluded
Too few trials	0
Insufficient advice taking	0
Too few choice trials	0
Wrong markers	2
Non-numeric advice	0
Total excluded	**2**
Total remaining	**33**

Method

This study used the continuous version of the Dates Task (§3.1.3).

Advice profiles The High accuracy and High agreement advisor profiles defined marker placements based on the timeline based on the correct answer and the participant's initial estimate respectively. Both advice profiles provided 'on-brand' advice on 12/15 trials which was normally distributed around the target point with a standard deviation of 5 years. On the remaining 3/15 trials both advice profiles issued 'off-brand' advice which was both wrong and distant from the participant's initial estimate. These 'off-brand' trials allow the response to the advisors to be disentangled from the advisors' advice itself.

Results

Exclusions Individual trials were screened to remove those that took longer than 60s to complete. Participants were then excluded for having fewer than 11 trials remaining, fewer than 8 trials on which they had a choice of advisor, or for giving the same initial and final response on more than 90% of trials. Participants were also excluded for technical problems with the experiment and data: unexpected marker widths or having corrupt values for the advisors' advice. Overall, 2 participants were excluded, with the details shown in Table 5.11.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial and final decisions can be seen in Figure 5.52. Compared to the Dots task, response times are not only longer, but they are also much more varied within participants.

```
## Picking joint bandwidth of 2110
```

Accuracy In both the original and replication, participants seemed to reduce their error (i.e. improve their response accuracy) only following High accuracy advice. It was not clear whether the High agreement advice yielded equivalent or different mean error for initial estimates and final decisions.

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Confidence can be measured by the width of the marker selected by the participant. Where participants are more confident in their response, they can maximise the points they receive by selecting a thinner marker. Where participants are unsure, they can maximise their chance of getting the answer correct by selecting a wider marker. Participants' error was lower for each marker width in final decisions than initial estimates (Figure 5.54). For both initial estimates and final decisions, error was higher for wider markers than for narrower ones.

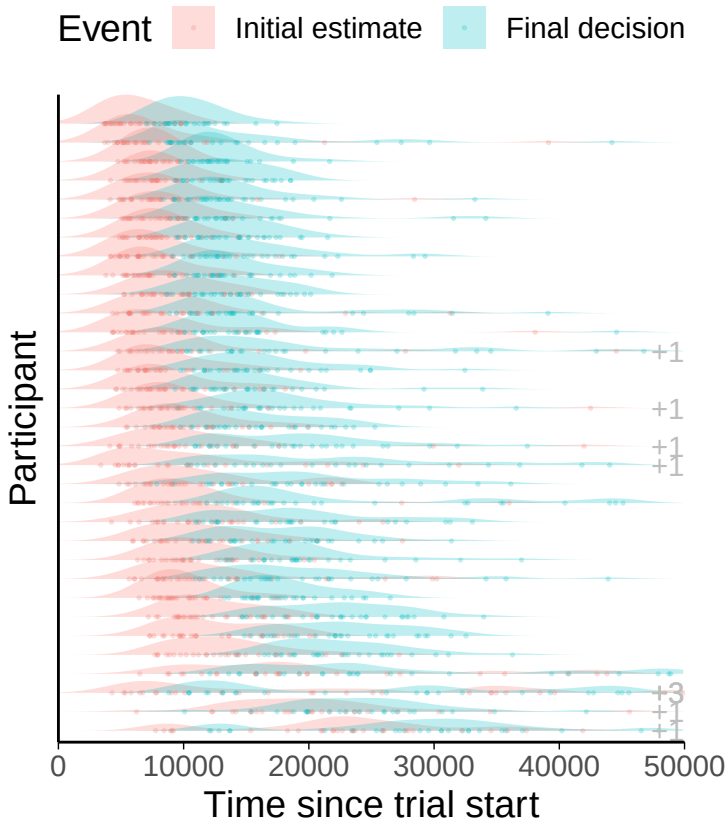


Figure 5.52: Response times for the Dates task with high accuracy/agreement advisors. Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

Experience with advisors The advice is generated probabilistically from the rules described previously in Table 5.9. It is thus important to get a sense of the actual advice experienced by the participants.

Advisor accuracy As shown in Figure 5.55, most participants experienced the High accuracy advisor as providing more accurate advice than the Low accuracy advisor, as intended in the experiment design. This indicates that the manipulation was effective for most participants individually, as well as for the sample on average. Participants experienced a much wider range of mean error for advice from the High agreement advisor.

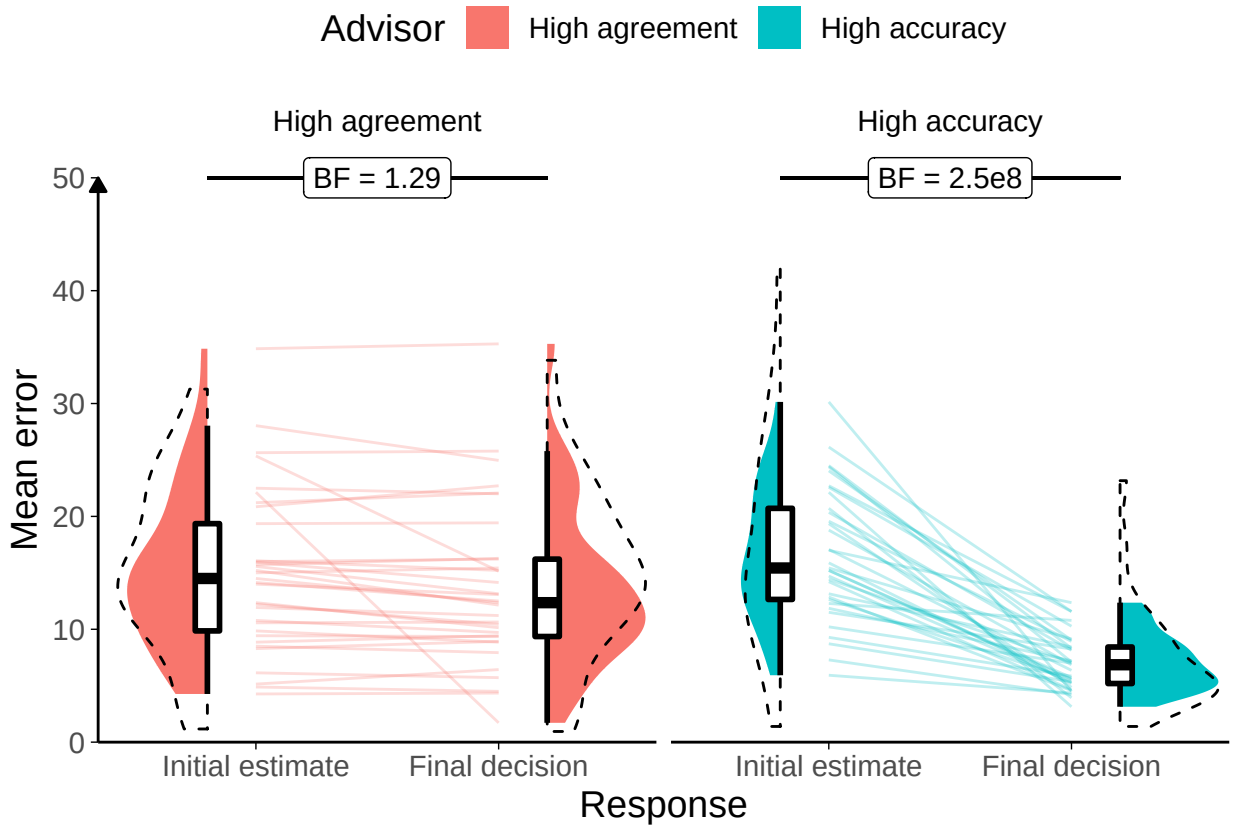


Figure 5.53: Response error for the Dates task with high accuracy/agreement advisors. Faint lines show individual participant mean error (the absolute difference between the participant's response and the correct answer), for which the violin and box plots show the distributions. The dashed line indicates chance performance. Dotted violin outlines show the distribution of participant means on the original study which this is a replication. The theoretical limit for error is around 100.

Advisor agreement Figure 5.56 shows the agreement rates experienced by each participant. There was a mixture of participants who experienced a higher agreement rate each advisor. Participants experienced a much wider range of agreement rates from the High accuracy advisor, presumably due to the variation in participants' own accuracies. Together with the lower error for the High accuracy advisor, the higher agreement for the High agreement advisor indicates that the manipulation worked well overall, and for most participants individually.

Advisor influence In the Feedback condition, the High accuracy advisor was more influential than the High agreement advisor, although there was no conclusive

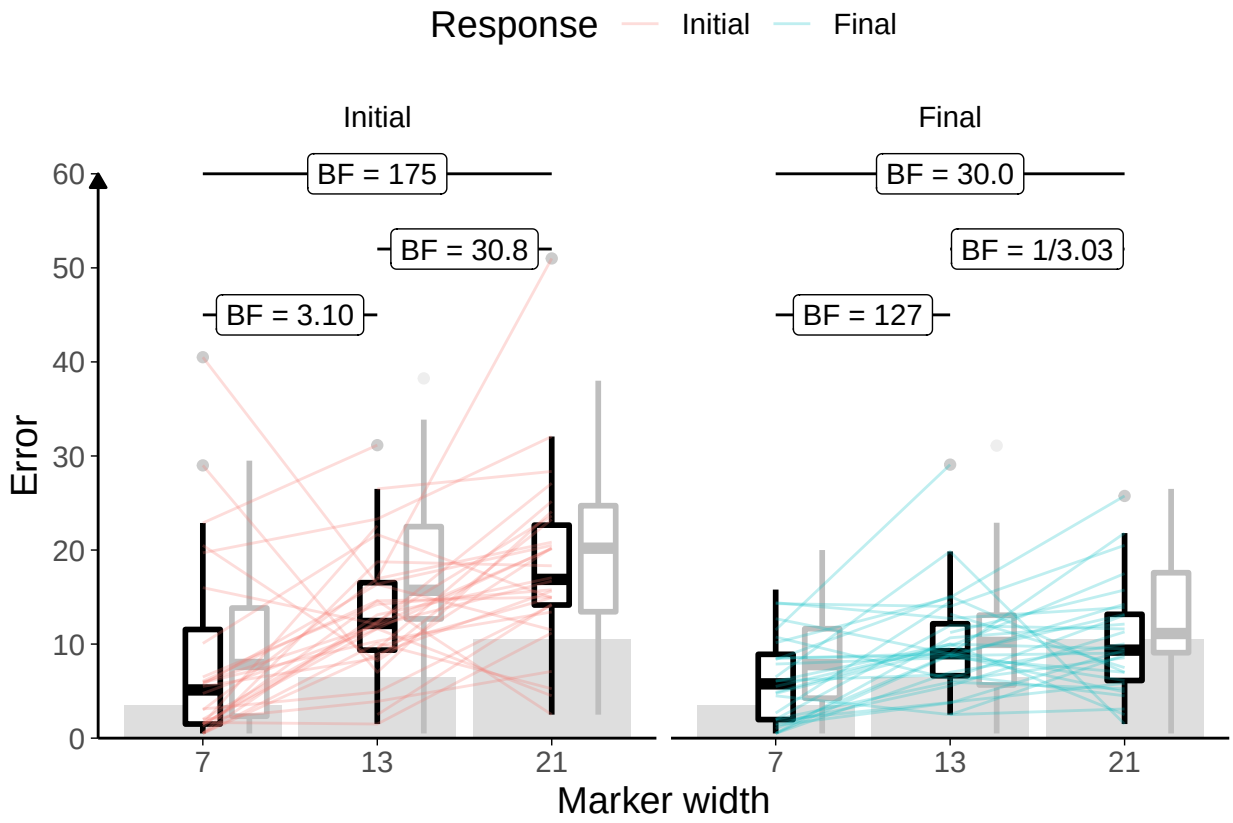


Figure 5.54: Error by marker for the Dates task with high accuracy/agreement advisors. Faint lines show individual participant mean error (distance from the centre of the participant's marker to the correct answer) for each width of marker used, and box plots show the distributions. Some participants did not use all markers, and thus not all lines connect to each point on the horizontal axis. The dashed box plots show the distributions of participant means in the original experiment of which this is a replication. The faint black points indicate outliers.
Grey bars show half of the marker width: mean error scores within this range mean the marker covers the correct answer.

evidence either way as to whether this effect was found in the No feedback condition (Figure 5.57). During the Familiarization phase, participants are learning about the advisors, so this design is not optimal for studying influence. Nevertheless, participants in the Feedback condition appear to have learned rapidly that the advice of the High accuracy advisor is worth following, and the advice of the High agreement advisor is not informative. There is a slight suggestion from the individual participant data that many of the participants in the No feedback condition may have been creeping towards this conclusion, but the statistics are

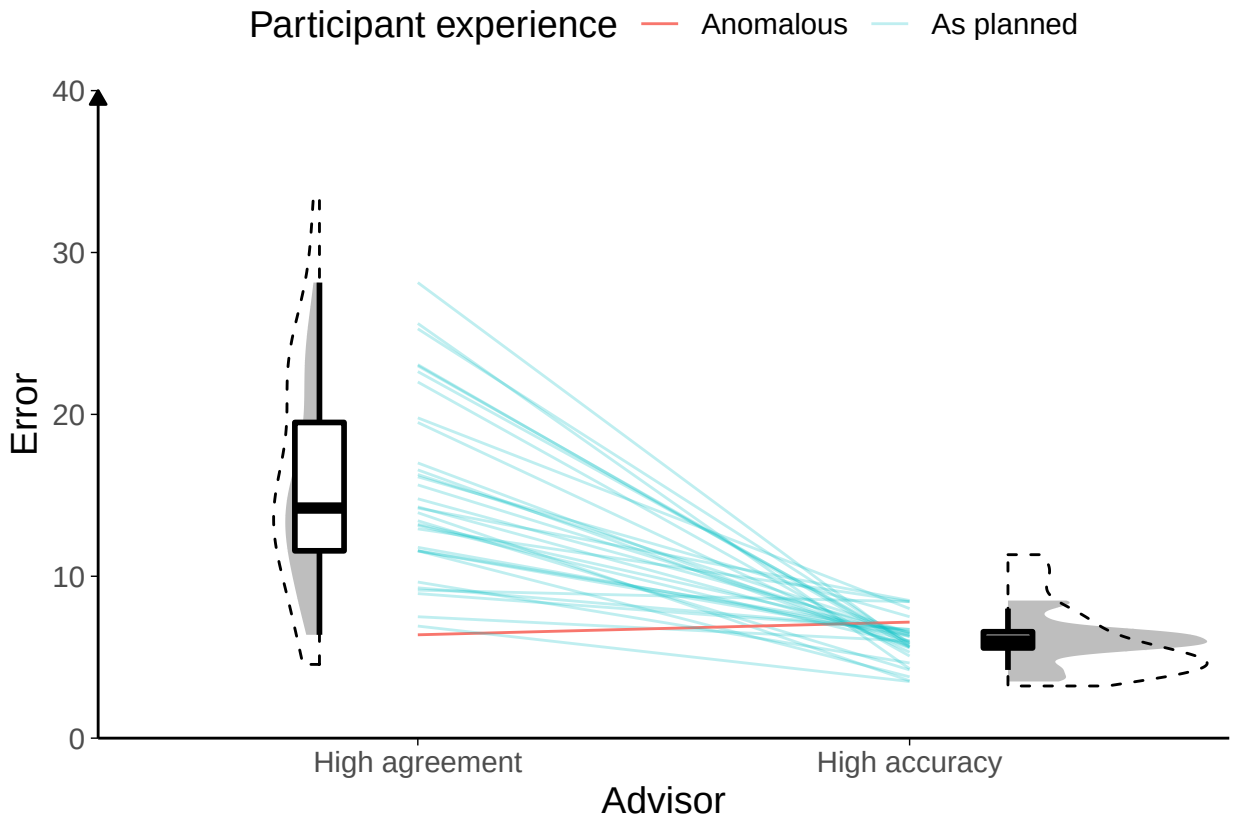


Figure 5.55: Advisor error for Dates task with high accuracy/agreement advisors. Coloured lines show the average error of the advisors as experienced by an individual participant. The colour of the line indicates whether the more accurate advisor was more accurate as per the experiment design. Box plots and violins show the distribution of the participant means, while the dashed violins show data from the original study of which this is a replication.

uninformative on the question.

Hypothesis test

```
## Warning: data coerced from tibble to data frame
```

Consistent with the result from the Dots task, in the No feedback condition participants' preferences for receiving advice from the High accuracy advisor were not different from chance ($t(13)$ = -0.16, p = .879, d = 0.04, BF = 1/3.66; $M_{Nofeedback}$ = 0.49 [0.29, 0.68], μ = 0.5). Participant preferences in the No feedback condition were almost perfectly evenly distributed, both in terms of which advisor was preferred and the strength of that preference, in both the original study and

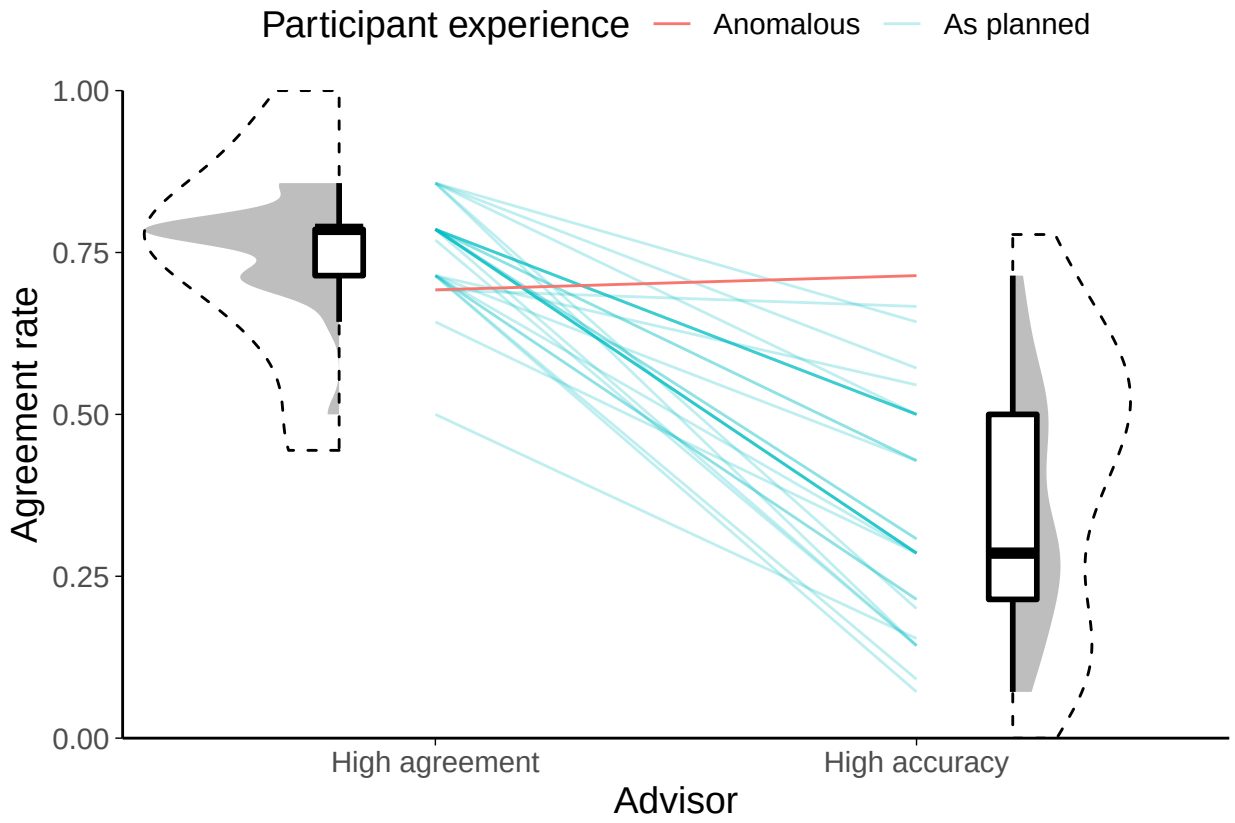


Figure 5.56: Advisor agreement for Dates task with high accuracy/agreement advisors. Faint lines show the average agreement rate of the advisors as experienced by an individual participant. Participants and advisors agree if there is overlap between the initial estimate and advice markers. Box plots and violins show the distribution of the participant means, while dashed violins show data from the original study of which this is a replication.

the replication (Figure 5.58).

In the Feedback condition, the mean of the participants' selection rates clearly favoured the High accuracy advisor ($t(18) = 5.00$, $p < .001$, $d = 1.15$, BF = 297; $M_{Feedback} = 0.81$ [0.68, 0.94], $\mu = 0.5$). This is consistent with a strategy which attempts to maximise the accuracy of final decisions.

5.3.3 Discussion

The influence results and the choice results were both clear in the Feedback condition, indicating that people are capable of attending to challenging but useful information provided they have a chance to learn that the information is actually useful. Where people are not able to learn about the information they receive, there does not

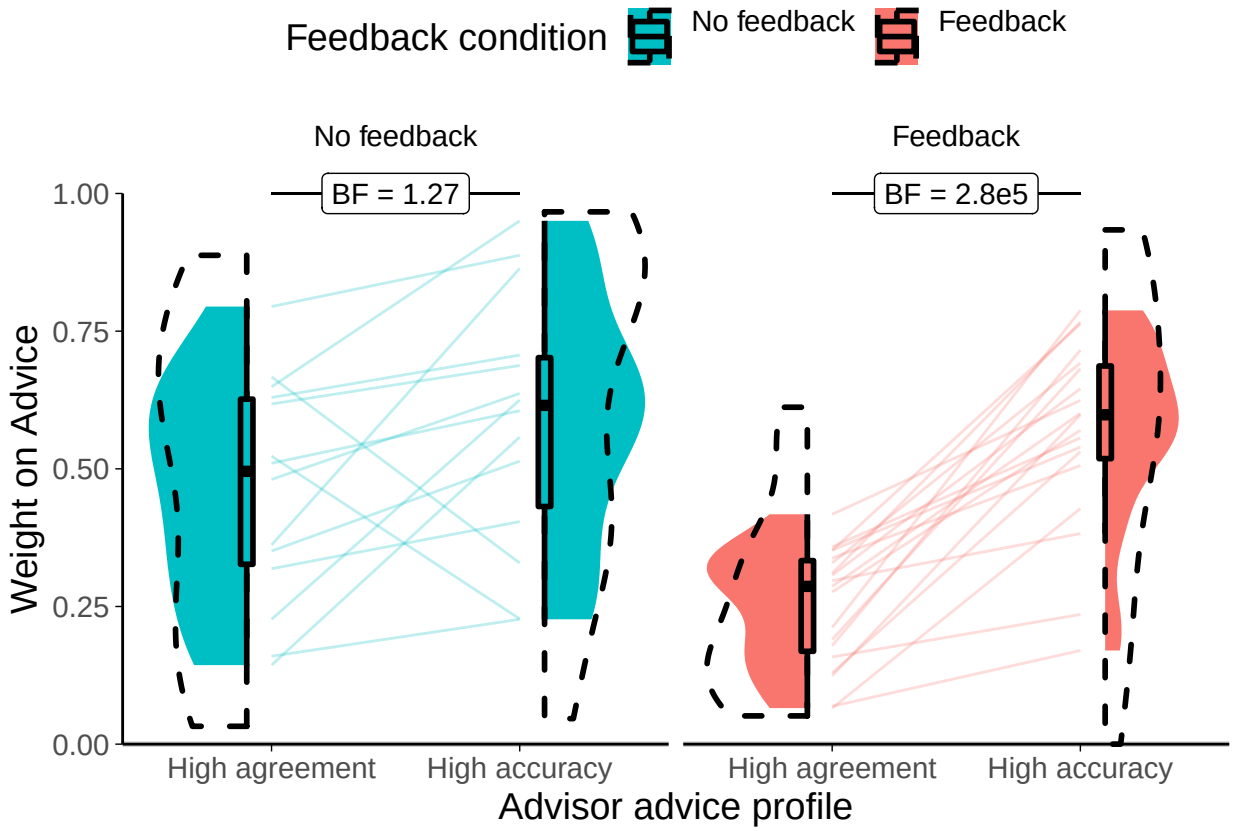


Figure 5.57: Date task advisor WoA for high accuracy/agreement advisors. Participants' weight on the advice for advisors in the Familiarization phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The dotted outline indicates the distribution of participant means in the original experiment of which this experiment is a replication. The theoretical range for influence values is [-2, 2].

seem to be a systematic response: some people seek agreement while others seek alternate perspectives, and the extent to which each strategy is pursued to the exclusion of the other is also highly variable. It is an open question whether a person's strategy choice in the absence of useful cues as to the utility of the information they receive is due to random selection or related in a meaningful way to their personality or cognitive style.

The results of the source selection behaviour in this experiment conceptually replicate the advisor influence results in the previous study (§4). While this study was not set up to examine influence rigorously, the influence results in this study

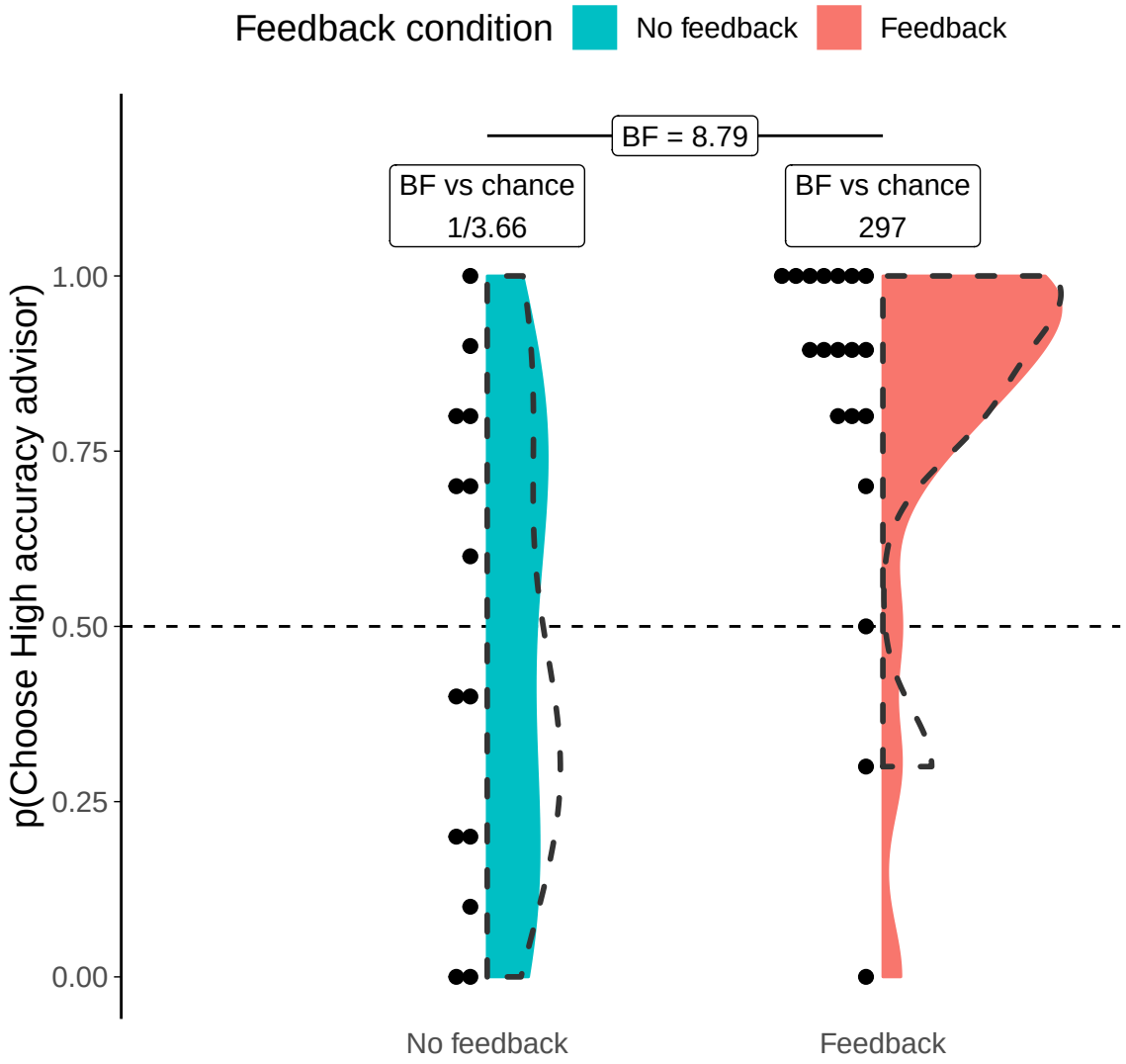


Figure 5.58: Dates task advisor choice for high accuracy/agreement advisors. Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line. Participants in the Feedback condition received feedback during the Familiarization phase, but not during the Choice phase. The dotted outline indicates the distribution of participant means in the original experiment of which this experiment is a replication.

are nonetheless compatible with the influence results in that previous study.

5.4 Confidence-contingent advice

Pescetelli and Yeung's model of advisor evaluation (§2.0.3) of advisor evaluation weights the updating of trust in an agreeing advisor using the confidence of the judge's initial estimate. This model can be directly tested by using advisors who have different agreement rates contingent upon the judge's initial estimate confidence. Pescetelli provided evidence for this model in the domain of advisor influence (!TODO[Reference Niccolo's actual paper]), and here a very similar design is used to explore this effect in the domain of advisor choice.

5.4.1 Dots Task

Open scholarship practices

 `https://osf.io/h6yb5`

 !TODO[OSFify data for these studies]

 `https://github.com/oxacclab/ExploringSocialMetacognition/blob/90c04ff21d3a2876beaddd9ee35c577a821e5727/AdvisorChoice/index.html`

Unanalysed data Two versions of this experiment have unanalysed data (2 - Bad initial exposure, 2b - Wrong agreement). The first of these had a bug which meant that the advisors were identical in the familiarisation phase. Both versions had a bug in which advisors which were supposed to agree with the participant's initial estimate gave the correct answer rather than agreeing. The 110 participants whose data was collected in these versions is not included in analysis. These participants' data could be included in an analysis which used a participant's actual experience of advice to predict their advice-taking and source selection behaviour, provided appropriate care was taken to reconstruct the advice data from the raw values instead of relying on the reported summaries.

Table 5.12: Confidence-contingent advisor advice profiles

	Initial decision confidence	Probability of agreement (%)	
		Bias Sharing	Anti Bias
Participant correct	High (top 30%)	90	50
	Medium (middle 40%)	70	70
	Low (bottom 30%)	50	90
Participant incorrect	Any	30	30
Total agreement	Participant correct	70	70
	Participant incorrect	30	30

Table 5.13: Participant exclusions for Dots task Confidence-contingent agreement experiment

Reason	Participants excluded
Accuracy too low	0
Accuracy too high	0
Missing confidence categories	3
Skewed confidence categories	1
Too many participants	0
Total excluded	**4**
Total remaining	**50**

Method

!TODO[clarify any methodological differences from the main methods chapter]

Advice profiles The two advisor profiles used in the experiment were Bias sharing and Anti-bias. The advisors are balanced for their overall accuracy and agreement rates, but the Bias sharing advisor agrees more frequently with participants when their initial estimate is correct and made with relatively high confidence. The Anti-bias advisor agrees more frequently with participants when their initial estimate is correct and made with relatively low confidence (Table 5.12).

Results

Exclusions Participants' data could be excluded from analysis where they have an average accuracy below 0.6 or above 0.85, do not have trials in all confidence categories, have fewer than 12 trials in each confidence category, or finish the experiment after 50 participants have already submitted data which passed the

other exclusion tests. Overall, 4 participants were excluded, with the details shown in Table 5.13.

Task performance Before exploring the interaction between the participants' responses and the advisors' advice, and the participants' advisor selection behaviour, it is useful to verify that participants interacted with the task in a sensible way, and that the task manipulations worked as expected. In this section, task performance is explored during the Familiarization phase of the experiment where participants received advice from a pre-specified advisor on each trial. There were an equal number of these trials for each participant for each advisor.

Response times Participants made two decisions during each trial. Neither of these decisions had a maximum response time. Each participant's response times for both initial and final decisions can be seen in Figure 5.59.

```
## Picking joint bandwidth of 248
```

Accuracy Accuracy of initial decisions was controlled by a staircasing procedure which aimed to pin accuracy to 71%. The accuracy of final decisions was free to vary according to the ability of the participant to take advantage of the advice on offer. As Figure 5.60 shows, participants' accuracy scores for initial decisions were close to the target values (partly because participants whose accuracy scores diverged considerably were excluded). Participants tended to improve the accuracy of their responses following advice from Anti-bias advisors, while their response accuracy was equivalent following advice from Bias sharing advisors.

Confidence Generally, we expect participants to be more confident on trials on which they are correct compared to trials on which they are incorrect. Participants were systematically more confident on correct as compared to incorrect trials for both initial estimates and final decisions.

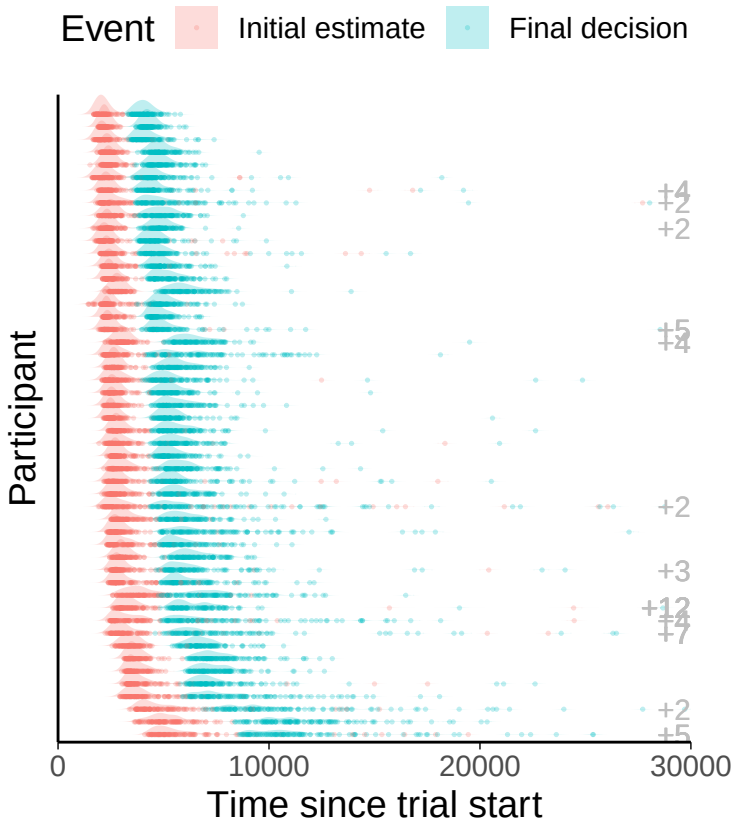


Figure 5.59: Response times for the Dots task with bias sharing/anti-bias advisors. Each point shows a response relative to the start of the trial. Each row indicates a single participant's trials. The ridges show the distribution of the underlying points, with initial estimates and final decisions shown in different colours. The grey numbers on the right show the number of trials whose response times were more than 3 standard deviations away from the mean of all final response times (rounded to the next 10s).

Metacognitive ability As shown by Figure 5.62, most participants showed above-chance metacognitive sensitivity for initial estimates and final decisions. Participants generally showed higher metacognitive sensitivity for final decisions, although this may be an artefact of a change in metacognitive bias. As expected given the controlled performance level on the task, participants' metacognition scores were not significantly correlated with their performance on the underlying task.

Experience with advisors The advice is generated probabilistically from the rules described previously in Table **??**. It is thus important to get a sense of the actual advice experienced by the participants.

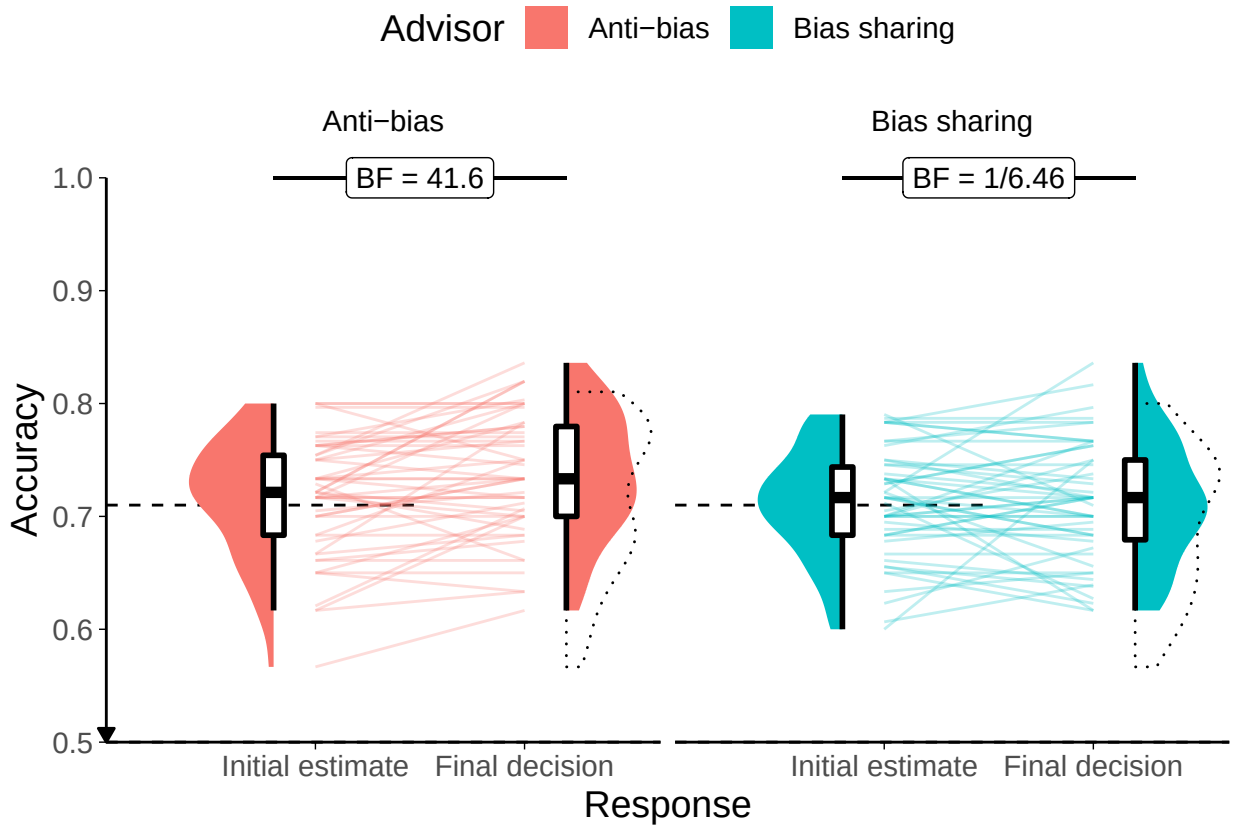


Figure 5.60: Response accuracy for the Dots task with bias sharing/anti-bias advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions. The half-width horizontal dashed lines show the level of accuracy which the staircasing procedure targeted, while the full width dashed line indicates chance performance. Dotted violin outlines show the distribution of actual advisor accuracy.

Advisor accuracy As shown in Figure 5.64, the advisors were designed to have the same accuracy, and the actual variation experienced by participants was relatively slight.

Advisor agreement Figure 5.65 shows the agreement rates experienced by each participant. As with the advisors' accuracy, this was supposed to be balanced between advisors by design. For both accuracy and agreement, this balancing approach appears to have worked well.

Should this break down agreement by initial in/correct as per the experiment design?

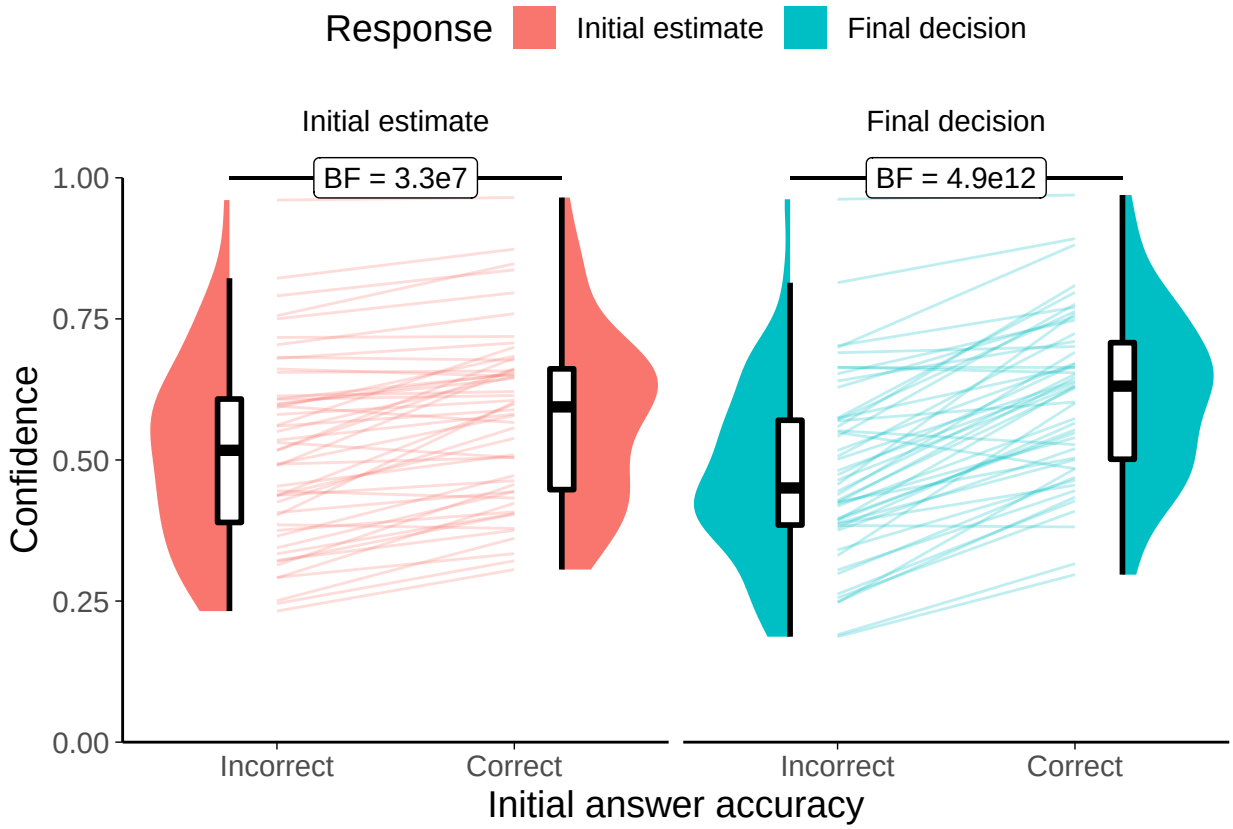


Figure 5.61: Confidence for the Dots task with bias sharing/anti-bias advisors. Faint lines show individual participant means, for which the violin and box plots show the distributions.

Advisor influence The advisors were similarly influential (Figure 5.66). Influence is only measured here during the Familiarization phase, when participants are forced to see the advice of one or another advisor, and participants are still learning about advisors' advice styles while this measure is being taken. This means that no strong conclusions should be drawn from this analysis.

Hypothesis test There was a strong tendency for participants to express no, or slight preferences between advisors ($t(49)$ = -1.52, p = .134, d = 0.22, BF = 1/2.20; M = 0.47 [0.44, 0.51], μ = 0.5). Intriguingly, almost all participants who expressed a stronger preference expressed it towards the Anti-bias advisor: in the direction counter to that hypothesised (Figure 5.67). The statistics are not conclusive here because the experiment was terminated prematurely. The good

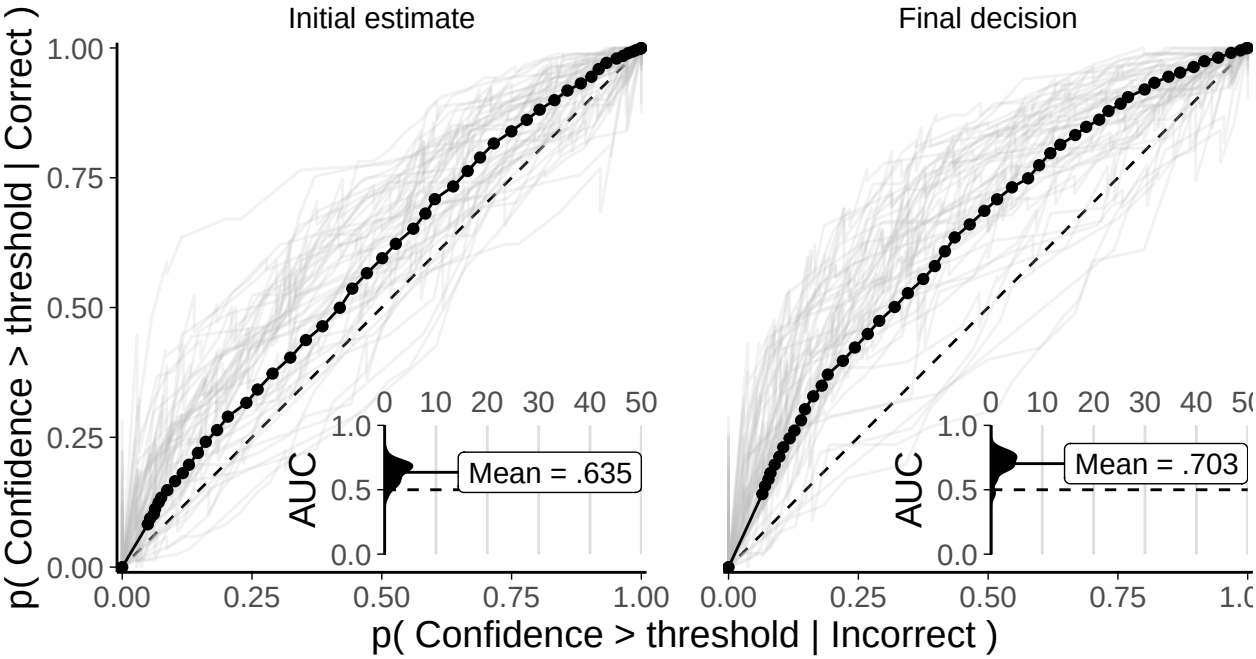


Figure 5.62: ROC curves for the Dots task with bias sharing/anti-bias advisors. Faint lines show individual participant data, while points and solid lines show mean data for all participants. Each participant's data are split into initial estimates and final decisions. For correct and incorrect responses seperately, the probability of a confidence rating being above a response threshold is calculated, with the threshold set to every possible confidence value in turn. This produces a point for each participant in each response for each possible confidence value indicating the probability of confidence being at least that high given the answer was correct, and the equivalent probability given the answer was incorrect. These points are used to create the faint lines, and averaged to produce the solid lines. The dashed line shows chance performance where the increasing confidence threshold leads to no increase in discrimination between correct and incorrect answers.

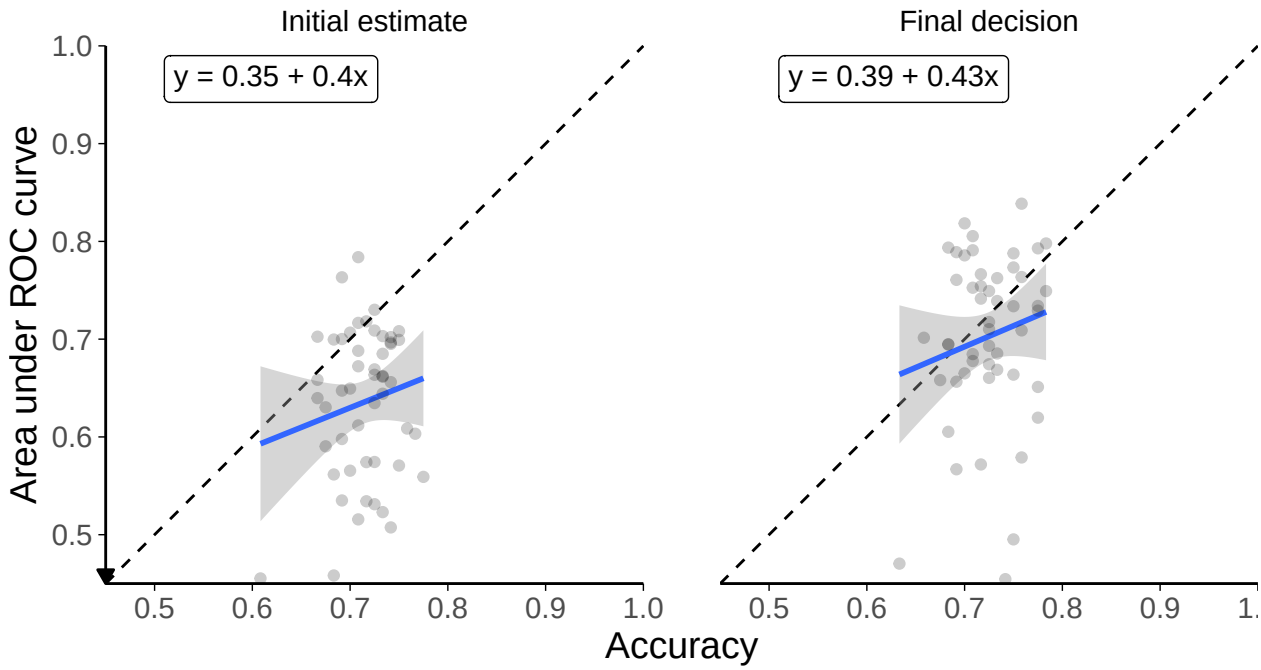


Figure 5.63: AUROC-accuracy correlation for the Dots task with bias sharing/anti-bias advisors.
Points show individual participant data for their area under the receiver operator characteristic (ROC) curve and their accuracy on initial estimates and final decisions. The blue lines and equation text show best-fit regression, and the shaded area gives its standard error. The equations give the regression equation plotted in blue, with bold coefficients being significant at p = .05.

thing about Bayes stats is we can quite happily go collect more data. Maybe we should do this because there's a chance we come out with a counter-to-expectations result here to do some talking and thinking about.

Discussion

While source selection and advice-taking are different domains, the previous experiments have shown strong similarities in the tendencies of participants: participants tend to be more influenced by the same kinds of advisors that they are more willing to hear from. On this basis, following Pecesetilli and Yeung CITE Niccolo and Nick, we would expect to see a preference for picking the Bias sharing advisor. We do not

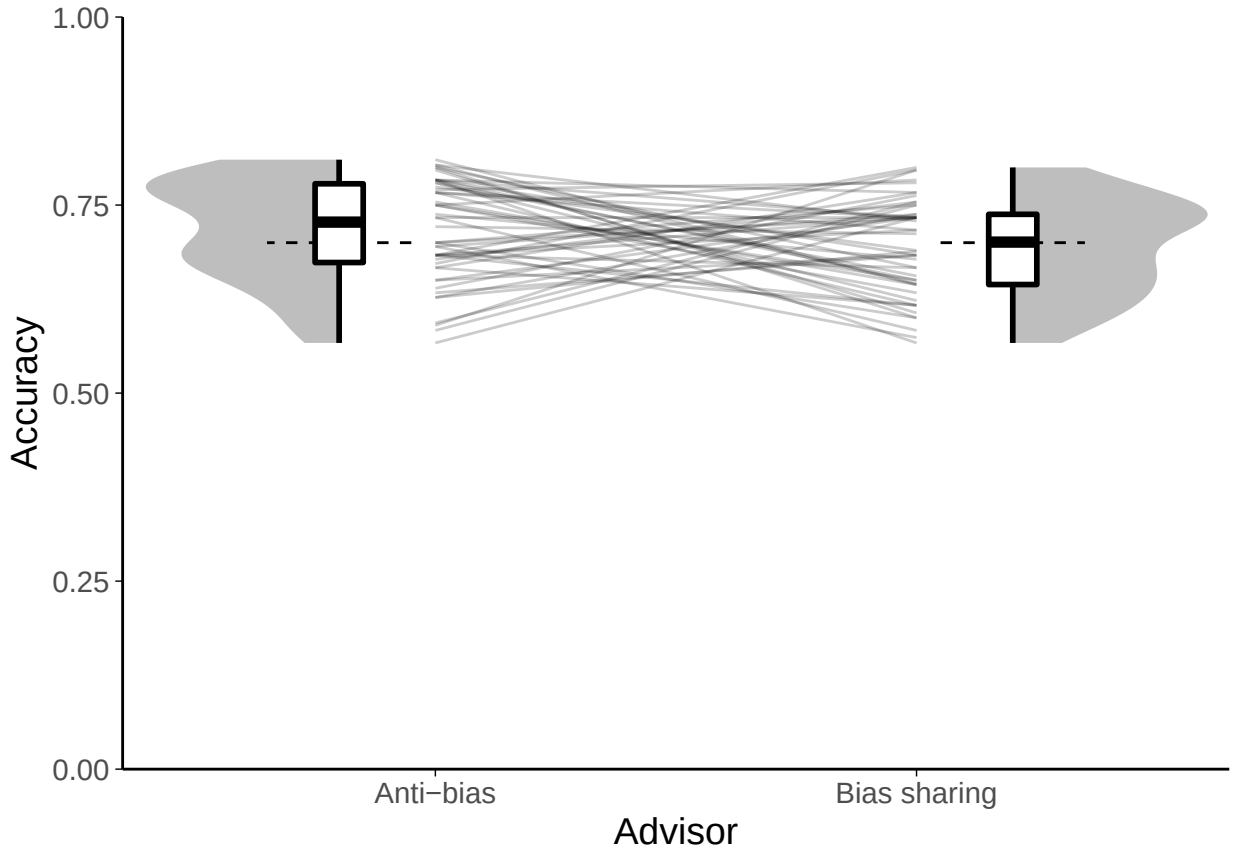


Figure 5.64: Advisor accuracy for Dots task with bias sharing/anti-bias advisors. Coloured lines show the average accuracy of the advisors as experienced by an individual participant. The colour of the line indicates whether the more accurate advisor was more accurate as per the experiment design. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the accuracy level for the advisors specified in their design.

see this preference, and, insofar as we see any preference at all, we see the opposite.

5.4.2 Dates task

The Dates task was not used to study confidence-contingent advice because such advice requires both a precise control over the relative agreement and accuracy rates of the advisors and the ability to estimate confidence in responses. The advisors' agreement (and hence accuracy) profiles depend on the participant's performance, and this is unknown a priori in the Dates task whereas it is controlled in the Dots task using a staircase procedure. Different approaches to estimating participants' confidence were trialled, including a pilot experiment in which the width of marker

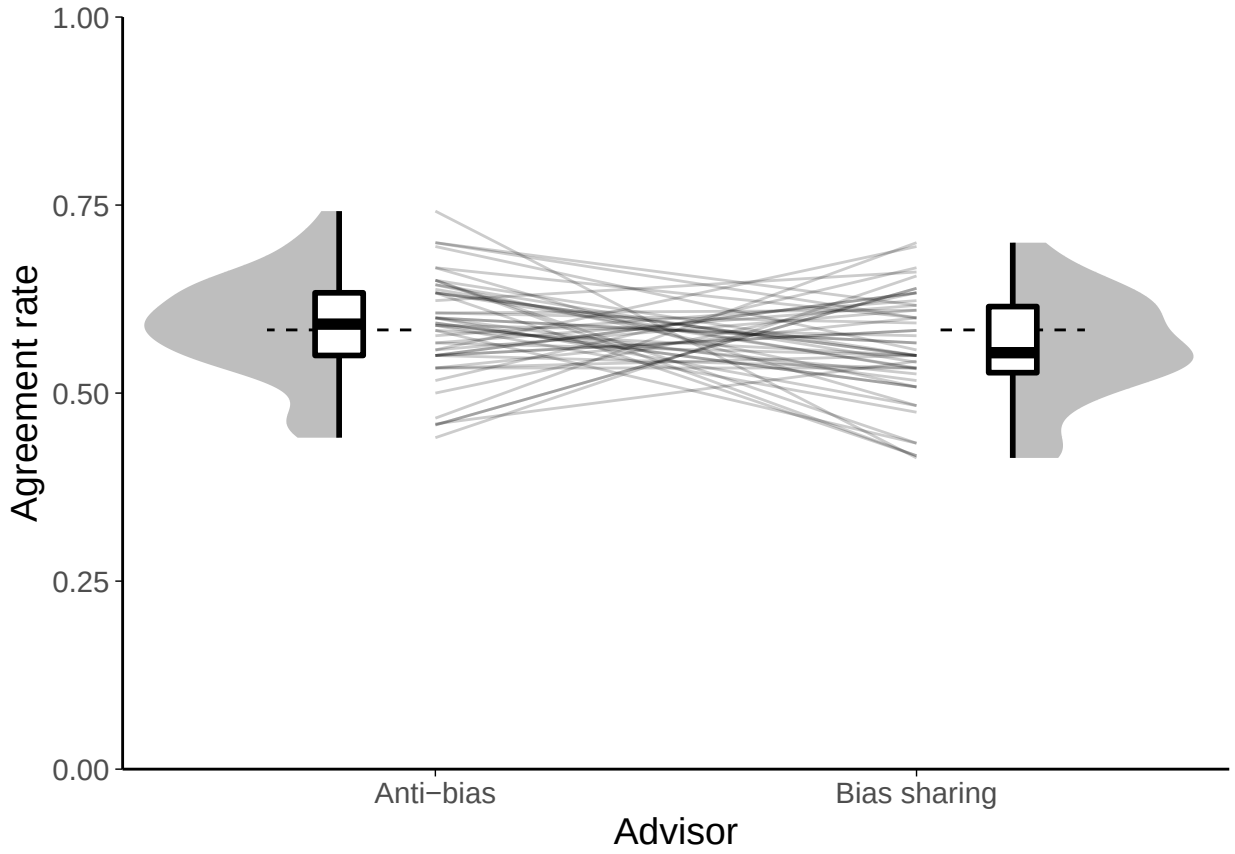


Figure 5.65: Advisor agreement for Dots task with bias sharing/anti-bias advisors. Faint lines show the average agreement rate of the advisors as experienced by an individual participant. Box plots and violins show the distribution of the participant means, while the dashed lines indicate the agreement level for the advisors specified in their design.

used by participants was used as a proxy for confidence, but none of the approaches produced any discernable effect of confidence on advisor agreement.

!TODO[If this study was worth reporting on, report on it (as a null manipulation). If not, why mention it? With the Marker use studies this probably works out as quite a lot of work and worth describing, at least. Probably worth detailing some of the attempts that were made to adapt Dates to Confidence-contingent advice.]

5.4.3 Lab study

`https://aspredicted.org/ze3tn.pdf`

`https://github.com/mjaquiery/nofeedback_trust`

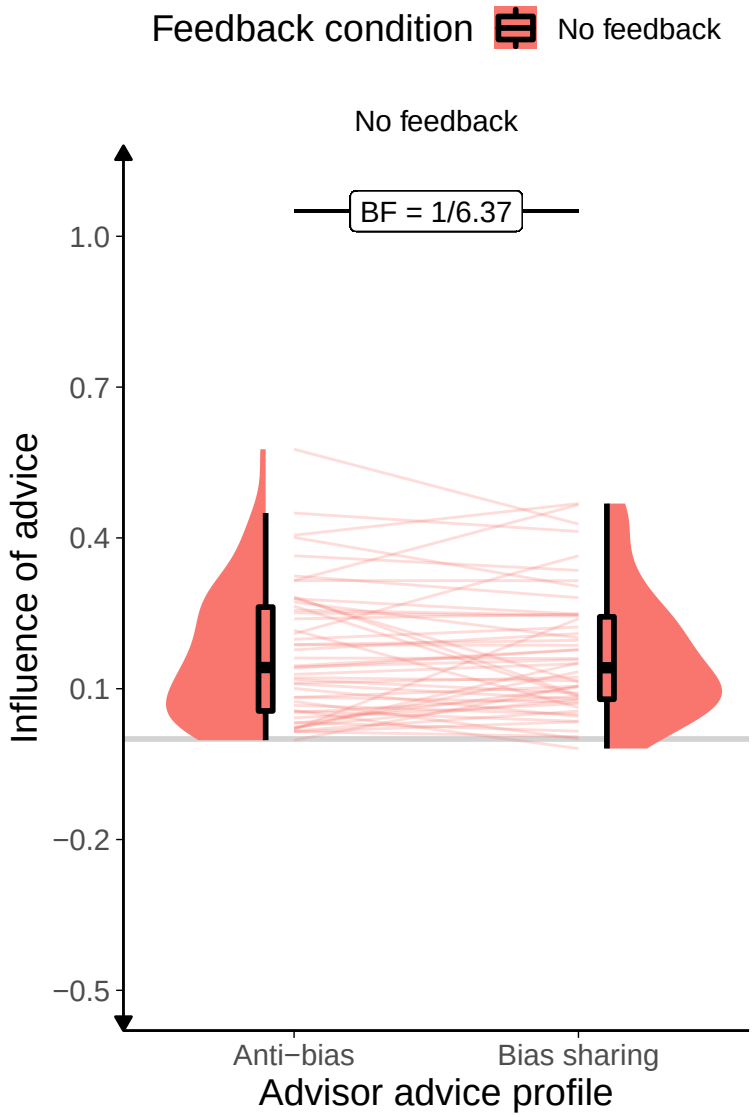


Figure 5.66: Dot task advisor influence for bias sharing/anti-bias advisors. Participants' weight on the advice for advisors in the Familiarization phase of the experiment. The shaded area and boxplots indicate the distribution of the individual participants' mean influence of advice. Individual means for each participant are shown with lines in the centre of the graph. The theoretical range for influence values is [-2, 2].

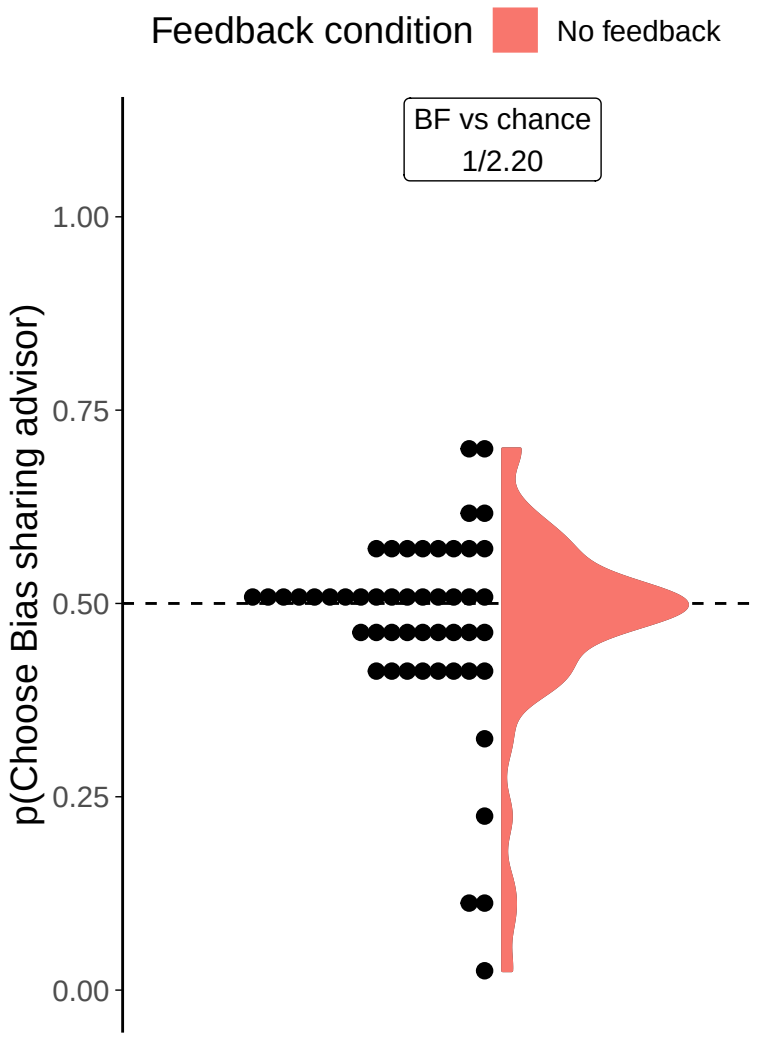


Figure 5.67: Dot task advisor choice for confidence-contingent advisors.
Participants' pick rate for the advisors in the Choice phase of the experiment. The violin area shows a density plot of the individual participants' pick rates, shown by dots. The chance pick rate is shown by a dashed line.

!TODO[Use a sensible archive format for this study data, archive on OSF, and produce data dictionary]

Pescetelli et al. (**pescetelliRoleDecisionConfidence2018**) showed that, in the absence of objective feedback, advice was more influential coming from an advisor who agrees with a participant when that participant is confident (*Bias Sharing*) than coming from an advisor who agrees with a participant when that participant is unconfident (*Anti Bias*). This provides evidence of a metacognitive sensitivity in the tracking of advice and the updating of advisor utility. Here we investigate whether these effects show up in the domain of advisor selection.

The literature on information exposure and evaluation indicates that people evaluate more favourably information which agrees with their currently-held opinion !TODO[REF], and preferentially seek out information sources which are likely to provide information which agrees with their currently-held opinion (**garrettEchoChambersOnline2009** **searsSelectiveExposureInformation1967**). If this holds true in the context of the judge-advisor system, advice from *Bias Sharing* advisers ought to be evaluated more favourably (influence should increase) and should be sought more frequently. Given the evidence in favour of the first of these propositions, we here investigate the latter: given a choice, will judges prefer to receive advice from a *Bias Sharing* advisor over receiving advice from an advisor who does not share the judge's bias?

Pescetelli et al. (**pescetelliRoleDecisionConfidence2018**) used a judge-advisor system to demonstrate that judges are influenced to a greater extent by advisers who share their biases. Participants played the role of judge in a judge-advisor system, while the advisers were virtual agents whose advice-giving was dependent upon the confidence and correctness of the judges' initial decisions. The advisers were balanced for overall agreement with the judge and objective correctness of advice. We place participants in a similar paradigm in which they are given a choice between advisers, and hypothesise that they will more frequently seek advice from the *Bias Sharing* advisor than from the *Anti Bias* advisor.

Method

!TODO[remove local dependencies in this script. Maybe add data to esmData or make a new package.]

Participants 24 participants (M_{age} = 22 ±SD 4.7, 5 male, 19 female, 0 other) recruited from the University of Oxford participant recruitment platforms took part in the experiment. An additional 2 participants attended experimental sessions but their data were not analysed. Participants were compensated for their time with either course credit for a psychology degree, or 10GBP.

Procedure The experiment consisted of a judge-advisor system with a perceptual decision task (Figure 5.68). Participants played the role of the judge, and the advisers were played by virtual agents whose answers depended upon the confidence with which the judge reported the initial decision. In the majority of trials (92%), participants were offered advice from virtual advisers. In one third of these trials ('choice trials'), participants chose which advisor to receive advice from by clicking on their respective portraits appearing at the top and bottom of the screen. On the remaining two thirds of trials ('forced trials'), participants were forced to take advice from one of the two advisers (equiprobably). On these trials, the forced advisor's portrait appeared at the top or bottom of the screen, with a red cross appearing in the other location, which was not selectable. On the remaining 8% of trials, participants received no advice and were given no opportunity to revise their initial decision. These 'catch trials' were included to encourage participants to attend to the initial decisions.

Each participant completed 363 trials (51 practice trials over 2 blocks + 12 x 26-trial experimental blocks) in which they had to identify the box with the most dots (Figure 5.68). The difficulty of the task was continually adjusted throughout the experiment using a 2-down, 1-up staircase procedure to keep the participant's initial decision accuracy at 72%. At the end of each block, participants were notified as to their (final decision) accuracy in the block and given the opportunity to rest

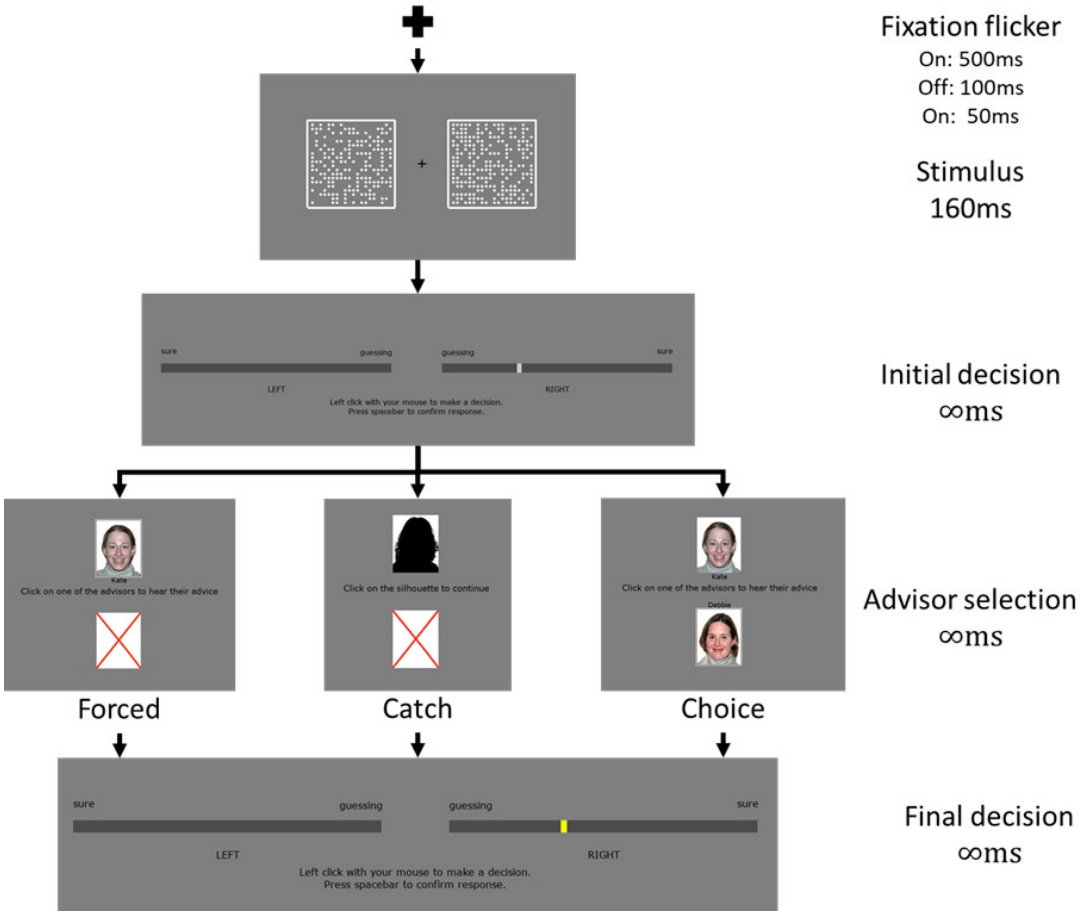


Figure 5.68: Experiment 1 procedure.
The task began with a blank screen containing only a fixation cross and progress bar. Momentarily prior to the onset of the stimuli the fixation cross flickered. The stimuli, two rectangles containing approximately 200 dots each, appeared for 0.16s, one on either side of the fixation cross. Once the stimuli disappeared, a response-collection screen appeared and prompted the participant to indicate their initial decision and its confidence by selecting a point within one of two regions. The left region indicated a decision that the target was on the left, and increasingly-leftwards points within that region indicated increasing confidence in that decision. The right region indicated a decision that the target was on the right, and increasingly-rightwards points within that region indicated increasing confidence in that decision.
Next, the participant was presented with a choice screen. The choice screen displayed two images, one at the top of the screen and one at the bottom. The images were one of the following: an advisor portrait, a silhouette, or a red cross. The red cross was not selectable, forcing participants to choose the other option. The silhouette offered no advice, and was only ever offered as a forced choice. Selecting an advisor image provided the participant with the opinion of that advisor on the trial.
Having heard the advice, the participant was again presented with the response-collection screen, with a yellow indicator marking their original response. A second (final) judgement was collected using this screen (except on catch trials), and the trial concluded.

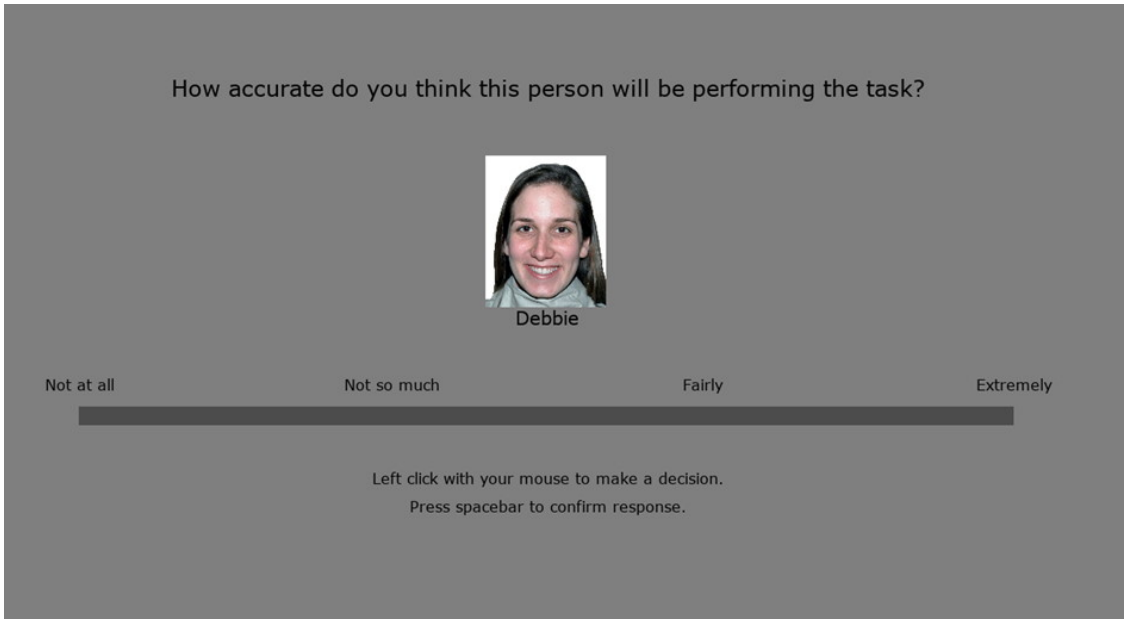


Figure 5.69: Experiment 1 advisor questionnaire.
Participants rated advisors on a number of different dimensions.

for as long as they wished. Throughout the experiment a progress bar provided a graphical indication of the number of trials remaining in the experiment. After each block participants were told what percentage of the (final) answers they had provided were correct and allowed to take a short, self-paced break. Prior to the first experimental block, after the final experimental block, and after the 4th and 8th experimental blocks, participants were presented with a questionnaire (Figure 5.69). The questionnaire contained 4 questions for each advisor. The questions asked for the judge's assessment of the advisor's likeability, trustworthiness, influence, and ability to do the task. The questions presented before the first experimental block were worded prospectively (e.g. 'How much are you going to like this person?' as opposed to 'How much do you like this person?'). Answers were provided by moving a sliding scale below the advisor's portrait towards the right for more favourable responses (marked 'extremely') or towards the left for less favourable responses (marked 'not at all').

Each participant attended the experiment individually, was welcomed and briefed on the experimental procedure, and had their informed consent recorded, before the experiment began. They were seated a comfortable distance in front of a 24'

Table 5.14: Experiment 1 advisor advice profiles

	Initial decision confidence	Probability of agreement (%)	
		Bias Sharing	Anti Bias
Participant correct	High (top 30%)	80	60
	Medium (middle 40%)	70	70
	Low (bottom 30%)	60	80
Participant incorrect	Any	30	30
Total agreement	Participant correct	70	70
	Participant incorrect	30	30

(1440x900 resolution) computer screen in a small, quiet, and dimly-lit room. The experiment took place wholly on the computer, and lasted around 45 minutes.

The experiment was programmed in MATLAB R2017b (**MATLAB2017**) using the Psychtoolbox-3 package (**kleiner2007s**).

Advisor advice profiles The advisers are virtual agents whose probability of agreeing with the participant's decision varies as a function of the participant's confidence and correctness in the initial decision phase. Table 5.14 illustrates how this relationship functions, and shows that the overall correctness and agreement rates of the advisers is equivalent overall. Importantly, on largest minority of trials, the middle 40%, the advisers are exactly equivalent, meaning these trials can be compared directly without confounds arising from agreement rate and initial confidence.

Analysis Data analysis was performed using R (**rcoreteamLanguageEnvironmentStatistical20**
For a full list of packages and software environment information, see !TODO[figure out where to include this stuff. Appendix? Also link to a containerized version of this.]

Bayes Factors (BF) are presented alongside p values and test statistics. A BF $<$ 0.33 indicates decisive evidence in favour of the null hypothesis over the alternative hypothesis (with lower values being increasingly clear), BF $>$ 3 indicates decisive evidence of the alternative over the null (with higher values being increasingly clear), and 0.33 <= BF <= 3 indicates there is insufficient evidence to reach a conclusion.

Capped influence Influence, the dependant variable in some analyses, is calculated as the extent to which the judge's initial decision is revised in the direction of the advisor's advice. The initial (C_1) and final (C_2) decisions are made on a scale stretching from -55 to +55 with zero excluded, where values <0 indicate a 'left' decision and values >0 indicate a 'right' decision, and greater magnitudes indicate increased confidence. Influence (I) is given for agreement trials by the shift towards the advice:

$$I|\text{agree} = f(C_1)\begin{cases} C_2 - C_1 & C_1 > 0 \\ -C_2 + C_1 & C_1 < 0 \end{cases} \tag{5.1}$$

And by the inverse of this for disagreement trials:

$$I|\text{disagree} = -I|\text{agree} \tag{5.2}$$

The confidence scale excludes 0, and thus the final decision can always be more extreme when moving against the direction of the initial answer than when moving further in the direction of the initial answer. A capped measure of influence was used to minimise biases arising from the natural asymmetry of the scale. This measure was calculated by truncating absolute influence values which were greater than the maximum influence which could have obtained had the final decision been a maximal response in the direction of the initial answer (Figure 5.70).

The capped influence measure I_{capped} is obtained by:

$$I_{\text{capped}} = f(C_1)\begin{cases} \min(I, 2C_1 - 55) & C_1 > 0 \\ \max(I, 2C_1 + 55) & C_1 < 0 \end{cases} \tag{5.3}$$

The explicit measure of trust is obtained using questionnaires. The questionnaires are delivered at 4 time points, and consist of 4 questions per advisor which are answered on a 1-100 scale.

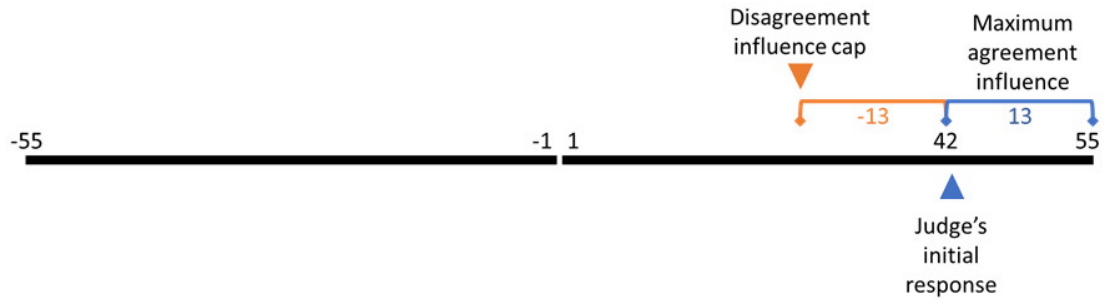


Figure 5.70: Capping influence to avoid scale bias.
In this example the judge's initial response is 42, meaning that their final decision could be up to 13 points more confident or up to 97 points less confident. Any final decision which is more than 13 points less confident is therefore capped at 13 points less confident.

Result

Descriptive statistics 26 participants took part in the study. One participant was unable to complete the experiment due to technical difficulties. Preregistration of the study analysis stated that data would be collected from 24 participants, so the final (overbooked) subject tested was excluded from analysis. Descriptive statistics for the 24 participants included in the analysis are presented in Table 5.15.

```
## `summarise()` regrouping output by 'cor1' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'cor2' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'advisorId' (override with `.groups` argument
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` regrouping output by 'step' (override with `.groups` argument)
```

Table 5.15: Descriptive statistics for Experiment 1

					95% CI	
			Target	Mean	Low	High
Participant proportion correct	**Initial estimate**	Bias Sharing	.71	.68	.67	.70
		Anti Bias	.71	.71	.69	.72
		Both	**.71**	**.70**	**.68**	**.71**
	Final decision	Bias Sharing	-	.70	.69	.71
		Anti Bias	-	.73	.72	.75
		Both	**-**	**.72**	**.71**	**.72**
Advisor-participant agreement rate	**Bias Sharing advisor**	Low confidence	.80	.60	.57	.63
		Medium confidence	.70	.70	.68	.72
		High confidence	.60	.79	.76	.82
		Initial wrong	.70	.30	.29	.32
		Initial correct	.30	.69	.68	.71
		All	**-**	**.57**	**.56**	**.58**
	Anti Bias advisor	Low confidence	.60	.79	.75	.82
		Medium confidence	.70	.70	.67	.74
		High confidence	.80	.64	.60	.67
		Initial wrong	.70	.32	.29	.34
		Initial wrong	.30	.71	.69	.73
		All	**-**	**.59**	**.57**	**.61**
Mean initial confidence	**Initial judgement**	Wrong	-	19.1	15.4	22.7
		Correct	-	22.8	19.1	26.5
		Both	**-**	**21.7**	**18.0**	**25.4**
Mean final confidence	**Final decision**	Wrong	-	18.2	14.8	21.6
		Correct	-	23.6	20.0	27.1
		Both	**-**	**22.0**	**18.6**	**25.5**
	Advisor agrees	Bias Sharing	-	26.1	22.6	29.6
		Anti Bias	-	24.6	21.2	28.1
		Both	**-**	**22.0**	**18.6**	**25.5**
	Advisor disagrees	Bias Sharing	-	16.6	12.9	20.2
		Anti Bias	-	18.3	14.4	22.3
		Both	**-**	**22.0**	**18.6**	**25.5**

```
## `summarise()` regrouping output by 'cor1' (override with `.groups` argument)

## `summarise()` ungrouping output (override with `.groups` argument)

## `summarise()` regrouping output by 'step' (override with `.groups` argument)

## `summarise()` regrouping output by 'cor1' (override with `.groups` argument)

## `summarise()` ungrouping output (override with `.groups` argument)
```

The descriptive statistics demonstrate the contingent agreement of the advisers,

with the Bias Sharing and Anti Bias advisers both agreeing at close to the target rate for most participants. While the ranges overall patterns are as designed, the variation in individual experience means some participants may have experienced by chance an advisor profile which contradicted the generative pattern (e.g. a Bias Sharing advisor who agreed on fewer high-confidence trials than mid-confidence trials). This is especially likely to be true of contingencies with fewer trials per participant, such as the incorrect trials. Overall, however, both the mean and 95% confidence intervals suggest the pattern was as desired for most participants most of the time.

Participants' revisions to their confidence were mostly in the direction of the advice, symmetrical across left and right responses, and usually relatively small, especially in agreement trials (Figure 5.71).

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

Advisor selection We hypothesised that the participants would display different pick rates for the Bias Sharing advisor versus the Anti Bias advisor. This hypothesis was evaluated by calculating the proportion of choice trials on which each participant picked the Bias Sharing advisor, and then testing these values as a one-sample t-test against the null hypothesis that the pick rates would be 0.5. No support was found for this hypothesis (; Figure 5.72), although the Bayesian test indicated that the data were not sufficient to conclude that no effect was present. There was considerable variability across participants in the overall pick rate for the Bias Sharing advisor (range = [.11, .82]).

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

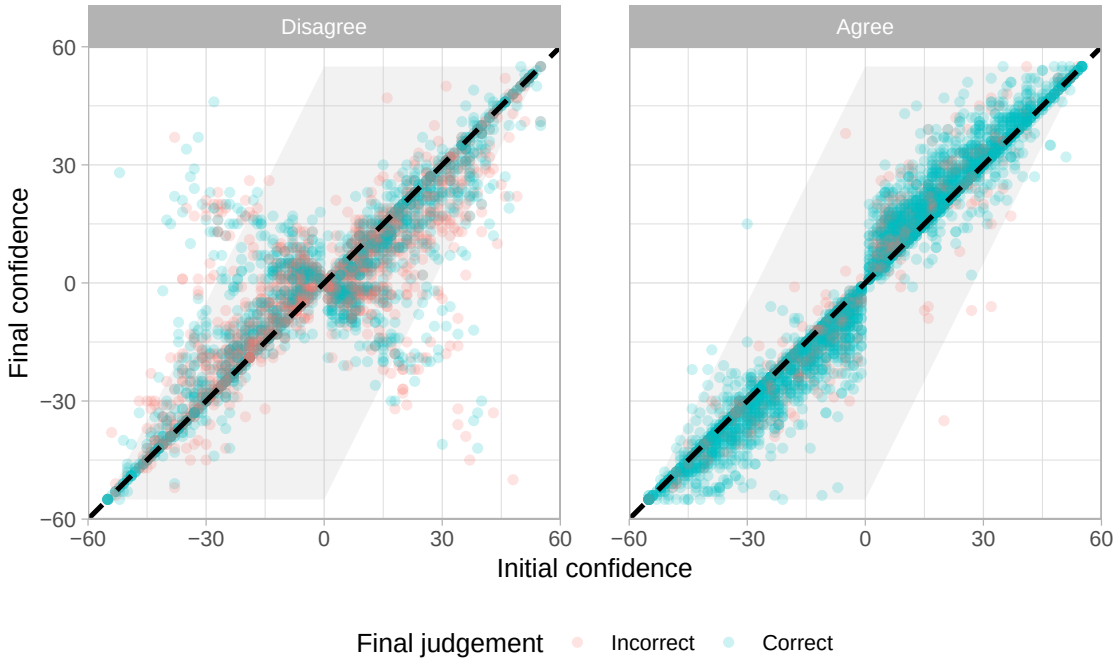


Figure 5.71: Initial vs final confidence.
Influence of the advisors is evident in the deviation from the dashed y = x line. Points lying below the line indicate a more leftward response from initial to final judgement. Points above the line indicate a more rightward response in the final judgement. The further away from the y = x line, the greater the change from initial to final judgement. Separate plots show agreement vs disagreement trials (between the advisor and judge), and separate colours indicate whether the judge's final decision was correct or incorrect. The shaded area indicates the boundary for the symmetrical influence measure. Points outside this area are truncated by moving them vertically until they meet the grey area.

Advisor selection on medium-confidence trials The advisers differed in their advice-giving as a function of the judge's initial confidence. In trials where the judge's initial decision was made with medium confidence, however, the advisers were equal on judge confidence and agreement rate. Comparing selection rates for these trials alone revealed a clear preference for the Bias Sharing advisor (; Figure 5.72 "Medium" confidence category), although the Bayesian analysis again indicated an insensitive result, albeit in the hypothesised direction.

Advisor influence Previous work in our lab demonstrated that the agree-in-confidence advisor exerted greater influence on the judges' final decisions than the

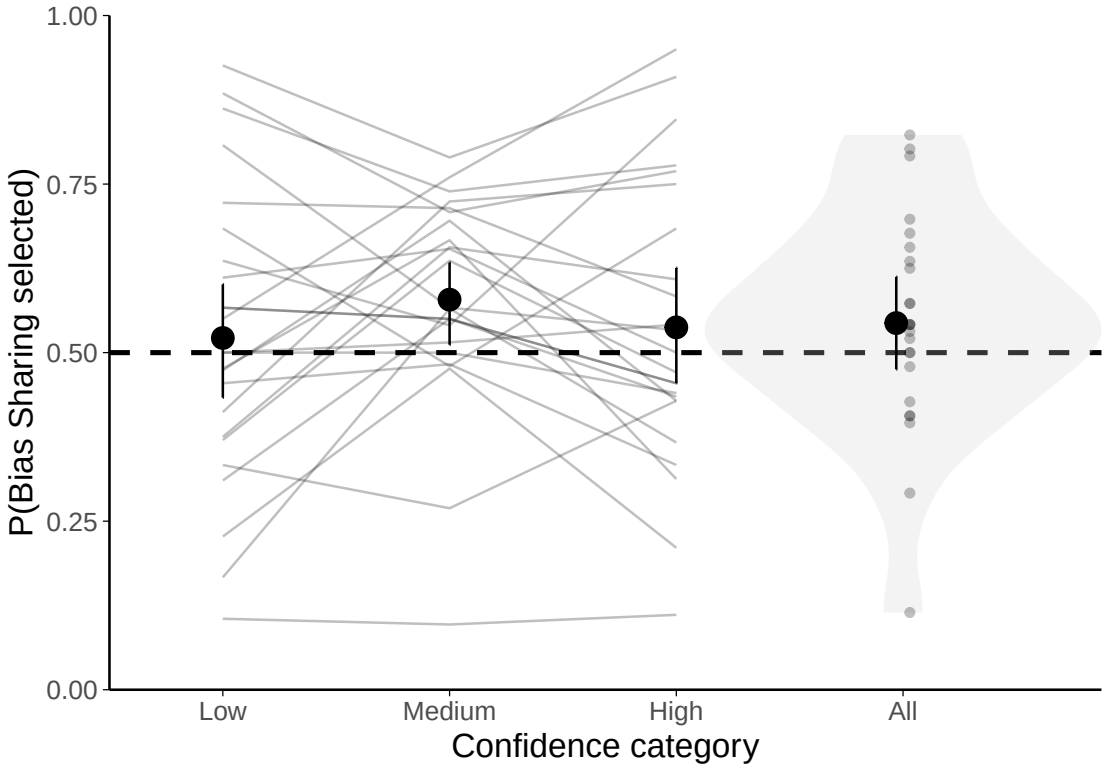


Figure 5.72: Advisor selection.
Proportion of the time each participant picked the Bias Sharing advisor. Faint lines and dots indicate data from individual participants, while the large dot indicates the mean proportion across all participants. The dashed reference line indicates picking both advisors equally, as would be expected by chance. Error bars give 95% confidence intervals.

agree-in-uncertainty advisor (**pescetelliUseMetacognitiveSignals2017**). Influence was examined with a 2x2x2 (Bias Sharing versus Anti Bias advisor; choice versus forced trials; agreement versus disagreement trials) ANOVA (Figure 5.73). No main effect was found for advisor !TODO[stats], meaning that the previous finding was not replicated. As shown in Table 5.16, the only statistically significant effect was the main effect of agreement, with disagreement producing higher influence than agreement !TODO[marginal means].

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

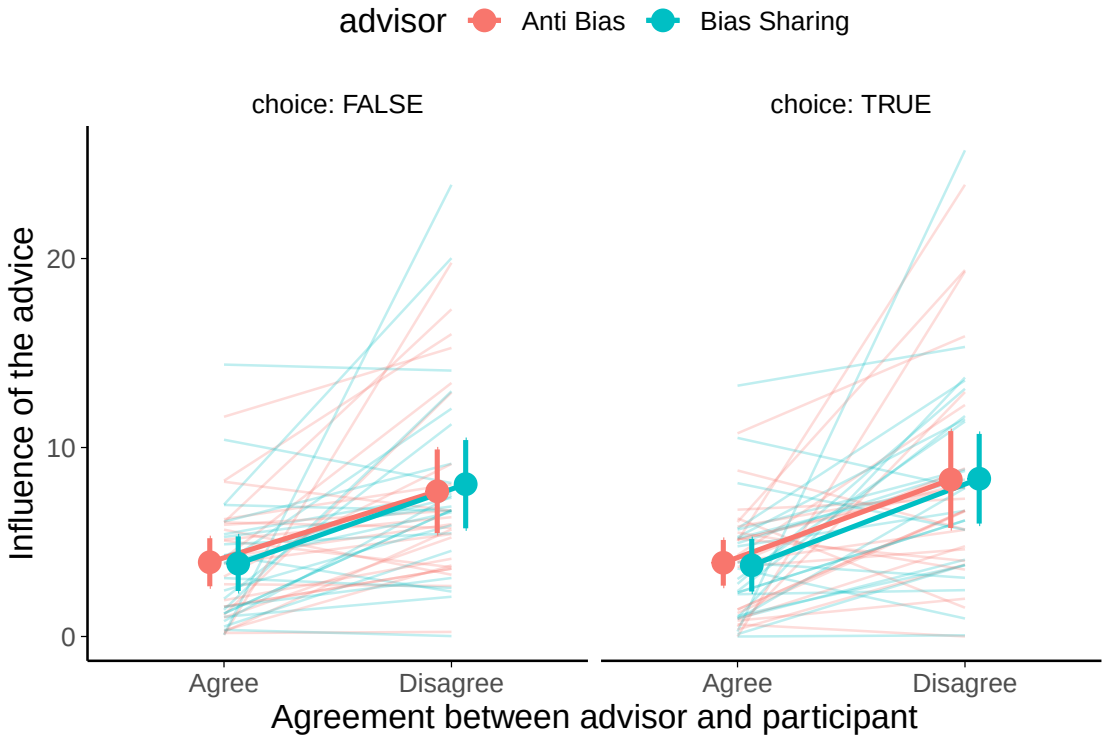


Figure 5.73: Advisor influence.
Influence of advice from each advisor by advisor, agreement, and trial type. Faint lines and indicate data from individual participants, while the dots indicate the mean proportion across all participants. Error bars give 95% confidence intervals.
Note: vertical axis is truncated to show group differences more clearly, the theoretical maximum influence given the scale is 110. The minimum is 0 as shown.

Table 5.16: ANOVA of Advisor influence in Experiment 1

Effect	$F(1, 23)$	p		η^2
AiC	0.28	.602		.001
agree	13.88	.001	*	.175
hasChoice	4.23	.051		.001
AiC:agree	0.75	.395		.001
AiC:hasChoice	0.04	.842		$<$ .001
agree:hasChoice	1.99	.172		$<$ .001
AiC:agree:hasChoice	0.01	.935		$<$ .001

Degrees of freedom: 1, 23

Table 5.17: ANOVA of Advisor influence on medium confidence trials in Experiment 1

Effect	$F(1, 21)$	p		η^2
AiC	0.88	.358		.005
agree	9.32	.006	*	.119
hasChoice	0.50	.485		.001
AiC:agree	3.06	.095		.010
AiC:hasChoice	0.64	.433		.002
agree:hasChoice	0.21	.650		.001
AiC:agree:hasChoice	0.01	.919		$<$ .001

Degrees of freedom: 1, 21

Advisor influence on medium confidence trials The agree-in-confidence and agree-in-uncertainty advisers differed by design in the frequency with which they agree with the participant as a function of the participant's confidence in their initial estimate. To control for the effects of initial confidence on influence, the above analysis was repeated using only those trials on which the initial estimate was correct and given with medium confidence. Two participants were missing data in this analysis: one participant had zero mid-confidence choice trials in which the agree-in-confidence advisor disagreed with them; and the other had zero mid-confidence choice trials in which the agree-in-uncertainty advisor disagreed with them. These participants were removed from this analysis. As before, the only effect which was statistically significant was agreement. The low number of trials in some of the intersections meant the some participants had to be dropped. The analysis was rerun with forced trials only, !TODO[stats]

Subjective assessment of advisers Participants answered questionnaires about their trust in the advisers at four time points during the experiment. We hypothesised that this subjective trust measure would change over the course of the experiment, with the agree-in-confidence advisor becoming more preferred over time. As indicated by Table 5.18, however, no such effects were found: subjective assessments of the advisers did not differ post-experiment.

```
## Warning: unnest() has a new interface. See ?unnest for details.
## Try `df %>% unnest(c(test, bf))`, with `mutate()` if needed
```

```
## Usually it is recommended to use column_spec before collapse_rows, especially ir
```

Sensitivity to the manipulation Finally, we planned to investigate the hypothesis that participants' choice of advisor would be sensitive to the differential agreement strategies of the advisers, e.g. participants might preferentially select the advisor with the greater likelihood of agreement given their initial confidence. This was investigated by testing the participants' mean bias sharing advisor pick rate in low- versus high-confidence trials. Pick rates did not differ (); Figure 5.72).

Exploratory analyses Below are reported analyses which were not part of the preregistration, but which were pursued to gain a greater insight into the behaviour of participants in the experiment.

Effect of initial agreement The hypothesised effect of the different advice-giving profiles of the advisers on their pick rates was not found, and we hypothesised that initial exposure to the advisers may have overshadowed information in subsequent blocks. To investigate this possibility, we examined the effect that agreement in the first experimental block had upon choices throughout the rest of the experiment (Figure 5.74). A simple regression (Table 5.19) indicated that the extent to which the Bias Sharing advisor agreed with the judge more frequently than the Anti Bias advisor in the first experimental block (Block 3) was a significant predictor of the preference for picking the Bias Sharing advisor in subsequent blocks.

The relationship between early experience of advisor agreement and overall advisor preference may be modified by an initial preference for one or the other advisor. We investigated this by adding advisor pick rate into the regression giving:

$$\text{PickRate}_{block>3} = \beta_1 \cdot \text{AgreementDifference}_{block=3} + \beta_2 \cdot \text{PickRate}_{block=3} + \beta_3 \cdot \text{AgreementDifference}_{block=3} \cdot \quad (5.4)$$

This model (Table 5.20) also fit the data well enough for interpretation, and represented an improvement on the previous model ($F(20,\ 22) = 7.2990236$, p

Table 5.18: Questionnaire responses pre- and post-experiment

		Bias sharing			Anti bias					
			95% CI			95% CI				
	Question[a]	Mean	Low	High	Mean	Low	High	BF	$t(23)$	p
Pre-experiment	How **accurate** do you think this person was when performing the task?	61.0	56.9	65.1	59.9	55.5	64.4	0.30	1.1	.536
	How much are you **influenced** by the opinions of this person?	51.8	45.1	58.4	50.4	43.5	57.2	0.30	1.4	.642
	How much do you **like** this person?	60.5	56.0	64.9	61.3	56.1	66.6	0.30	-0.9	.613
	How **trustworthy** are the opinions of this person?	61.1	56.1	66.1	61.7	57.4	66.0	0.29	-0.6	.770
Post-experiment	How **accurate** do you think this person was when performing the task?	56.9	49.1	64.6	55.0	46.9	63.0	0.30	1.9	.607
	How much are you **influenced** by the opinions of this person?	54.0	44.7	63.4	53.3	43.1	63.6	0.29	0.7	.831
	How much do you **like** this person?	57.2	48.9	65.5	58.7	48.8	68.5	0.29	-1.5	.529
	How **trustworthy** are the opinions of this person?	57.2	50.5	64.0	56.0	47.8	64.2	0.29	1.3	.700

[a] Emphasis added.

Table 5.19: Linear regression of pick rate in later blocks by initial agreement difference

Effect	Estimate	SE	t	p	
(Intercept)	0.54	0.03	16.92	$< .001$	$*$
agreeRateDifference.block3	0.37	0.17	2.27	.034	$*$

Model fit: F(5.1, 1) = 22; p = .034; R^2_{adj} = .152

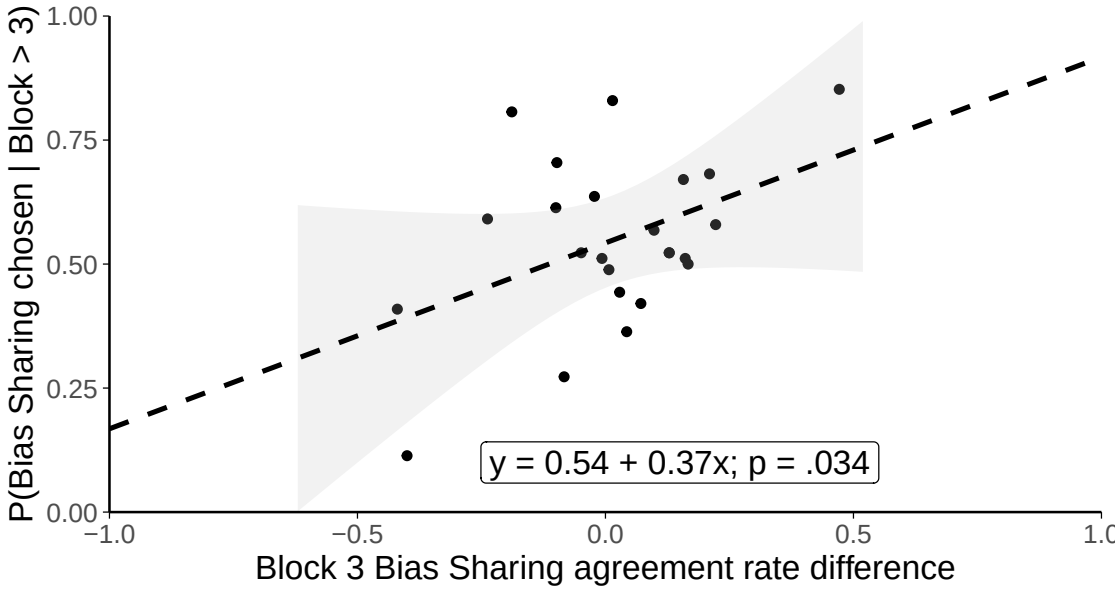


Figure 5.74: Initial agreement and subsequent preference.
figure shows the relationship between the agreement rate of the Bias Sharing advisor in the first experimental block (relative to the Anti Bias advisor) and the proportion of the time the participant picked the Bias Sharing advisor in later blocks. The dashed line shows the best-fit regression line, with shaded 99% confidence intervals.

Table 5.20: Linear regression of pick rate in later blocks by initial agreement difference and preference

Effect	Estimate	SE	t	p	
(Intercept)	0.41	0.06	6.41	$< .001$	$*$
agreeRateDifference.block3	1.19	0.28	4.29	$< .001$	$*$
aicPickRate.block3	0.27	0.12	2.27	.034	$*$
agreeRateDifference.block3:aicPickRate.block3	-2.43	0.72	-3.37	.003	$*$

Model fit: F(7.6, 3) = 20; p = .001; R^2_{adj} = .461

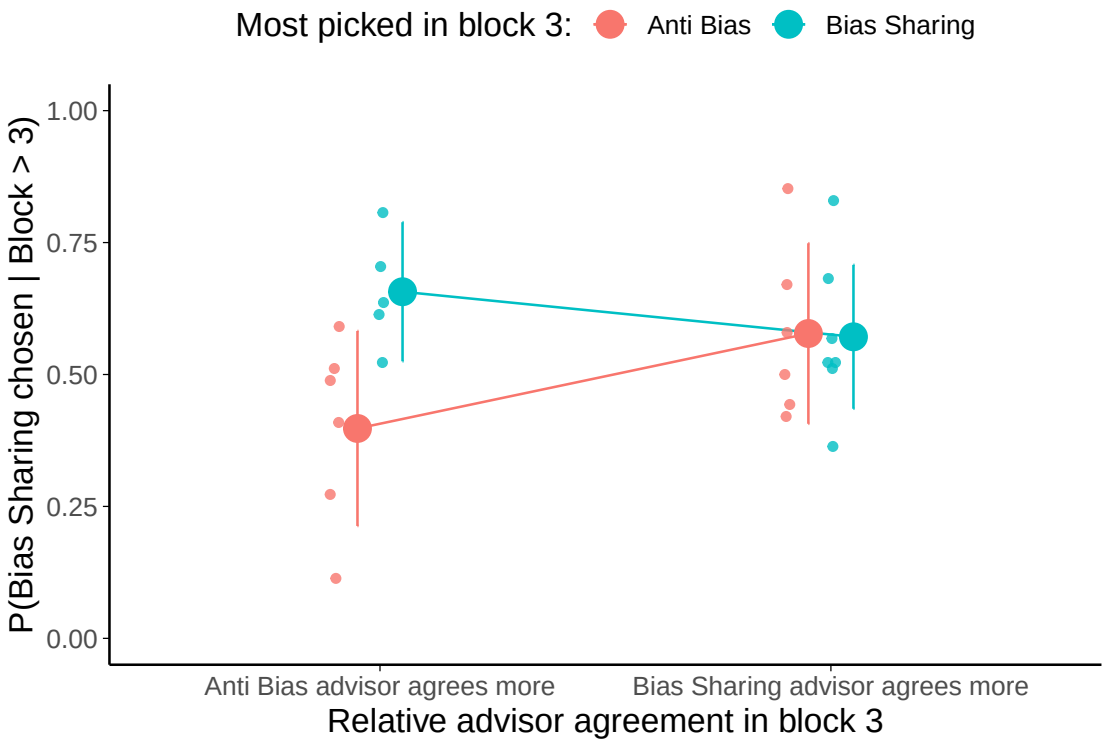


Figure 5.75: Initial agreement and preference predicting subsequent preference. Figure shows the relationship between whether the agreement rate of the Bias Sharing advisor in the first experimental block (relative to the Anti Bias advisor) and the proportion of the time the participant picked the Bias Sharing advisor in later blocks, split by whether the participant picked the Bias Sharing or Anti Bias advisor more often in the first block. Splits are based on the sample means, and error bars give 95% confidence intervals.

= .004). Early agreement difference remained predictive, and relative pick rate in block 3 was also predictive. Finally, the interaction between these predictors was also important to the model, although the negative sign of the interaction beta (β_3 = -2.43) was unexpected, and required further exploration. Figure 5.75 shows a marginal means plot for the interaction with the predictors collapsed to binary measurements using a mean split.

```
## Warning: ‘fun.y‘ is deprecated. Use ‘fun‘ instead.
```

```
## Warning: ‘fun.y‘ is deprecated. Use ‘fun‘ instead.
```

The theoretical mechanism by which judges evaluate advisor accuracy in the absence of feedback is through their own metacognitive awareness. Thus, we would expect that, given preferences appear to be set by initial exposure to the advisors,

Table 5.21: Linear regression of pick rate in later blocks by initial agreement difference by confidence

Effect	Estimate	SE	t	p	
(Intercept)	0.43	0.10	4.21	.002	$*$
agreeRateDifference.highConf.block3	0.02	0.08	0.23	.825	
agreeRateDifference.medConf.block3	0.12	0.13	0.92	.382	
agreeRateDifference.lowConf.block3	0.23	0.11	2.11	.064	
aicPickRate.block3	0.37	0.21	1.74	.116	

Model fit: F(2.6, 4) = 9; p = .107; R^2_{adj} = .331

those preferences would be more affected by high confidence agreement in block 3. This was tested using a regression model in which block 3 agree-in-confidence advisor agreement was used as a predictor split according to the three different levels of confidence. Several participants (5) had no trials in block 3 for one or both advisors at one or more confidence levels; these participants were removed from the analysis. Neither the overall model fit nor any of the predictors in this model (Table 5.21) were significant, but the low-confidence beta was substantially larger than the high- and medium-confidence betas.

A prediction of the high weighting of initial exposure is that this weighting should drop off over time. This does not appear to be the case. Figure 5.76 shows the regression coefficient for Bias Sharing advisor agreement in each block against the Bias Sharing overall pick proportion: no significant negative correlation was found between block number and the regression coefficient for relative agreement in that block predicting overall advisor preference.

```
## `geom_smooth()` using formula 'y ~ x'
```

Capped influence measure

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Preregistered analyses indicated that advisors are more influential when they disagree with the judge. This result may simply be an artefact of the scale: although the scale extends equally far in both directions, there is necessarily potential for adjusting answers to follow agreeing advice by virtue of the fact that the decision is not placed in the middle of the scale. To redress this balance, a capped measure

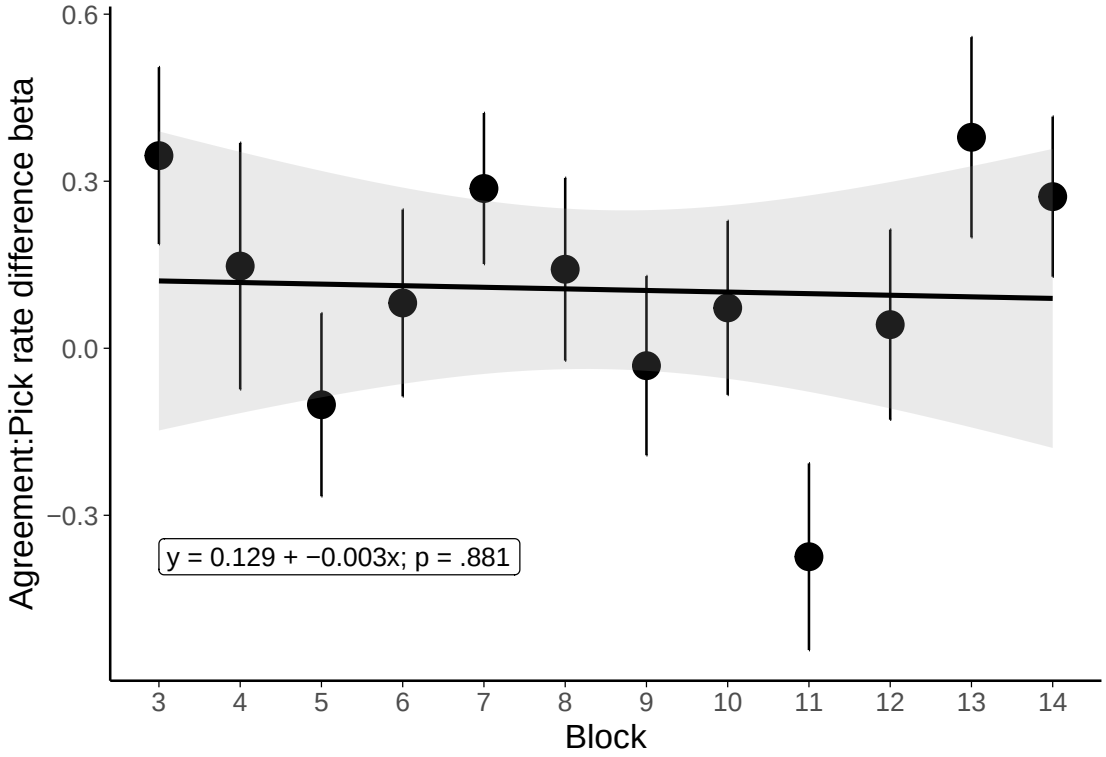


Figure 5.76: Block 3 agreement predicting pick rate in each block.
Figure shows the relationship between the agreement rate of the Bias Sharing advisor in the first experimental block (relative to the Anti Bias advisor) and the proportion of the time the participant picked the Bias Sharing advisor in each later block. Error bars give +/-1 standard error. Shaded area gives 95% confidence for the overall regression line.

of influence (§5.4.3) was used. Of the 6912 trials on which judges received advice, 193 trials (2.8%) had their influence scores capped by this process. These trials were predominantly disagreement trials[2] (178, 92.2%). The mean number of trials adjusted for each participant was $M_{TrialsAdjusted} = 8.04$ [2.55, 13.54], though the distribution was highly positively skewed and 10 participants had no trials adjusted this way (Figure 5.77). Analysis using the 2x2x2 (advice type/agreement/choice) ANOVA described above (§5.4.3) no longer showed a significant effect of agreement. As before, neither interactions nor other main effects were significant (Table 5.22).

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

[2]Agreement trials can have capped values if the final decision goes in the opposite direction to the advice.

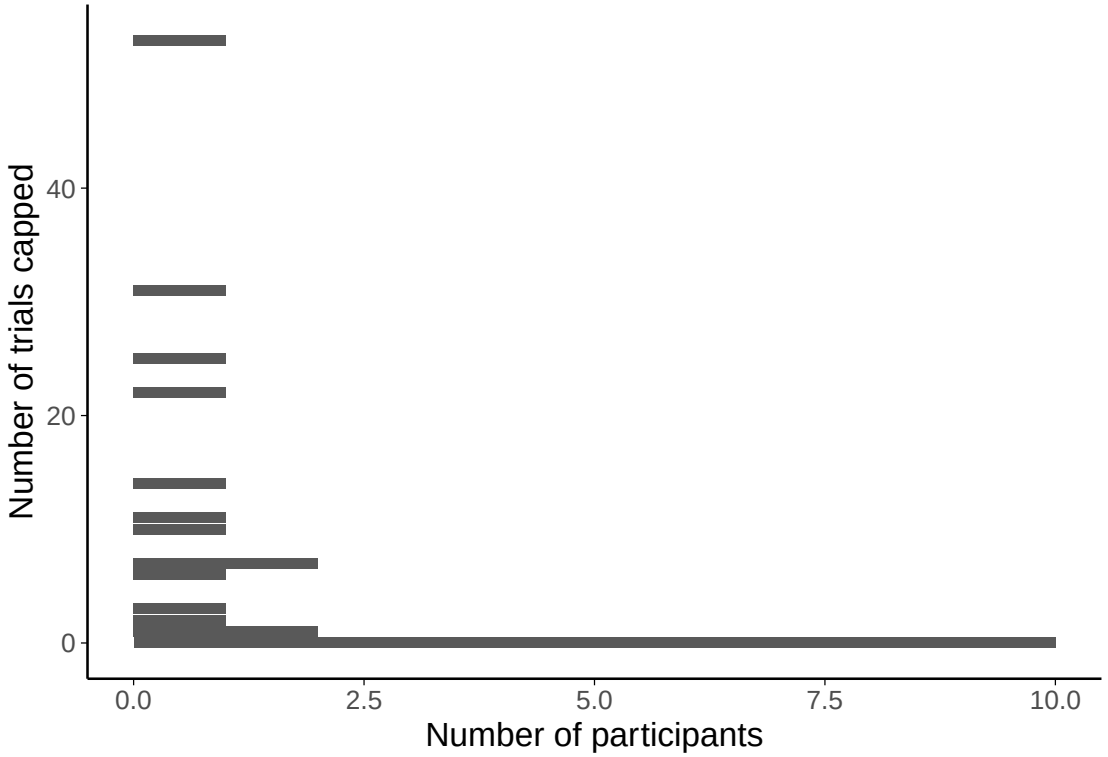


Figure 5.77: Advisor influence capping frequency.
Histogram showing the number of trials in which influence was capped for each participant.

Table 5.22: ANOVA of Advisor influence with capped values in Experiment 1

Effect	$F(1, 23)$	p		η^2
adviceType	3.67	.068		.010
agree	0.30	.588		.008
hasChoice	3.19	.087		.003
adviceType:agree	2.44	.132		.013
adviceType:hasChoice	0.70	.411		.002
agree:hasChoice	0.71	.407		.001
adviceType:agree:hasChoice	0.21	.649		$< .001$

Degrees of freedom: 1, 23

Subjective assessment and choice of advisors We did not observe the hypothesized preference for the agree-in-confidence advisor, either as measured by pick rate or by questionnaire response. Nevertheless, it was possible to examine the relationship between the behavioural measure of pick rate and the self-reported questionnaire answers by examining the extent to which participants who picked the agree-in-confidence advisor more frequently rated that advisor more favourably.

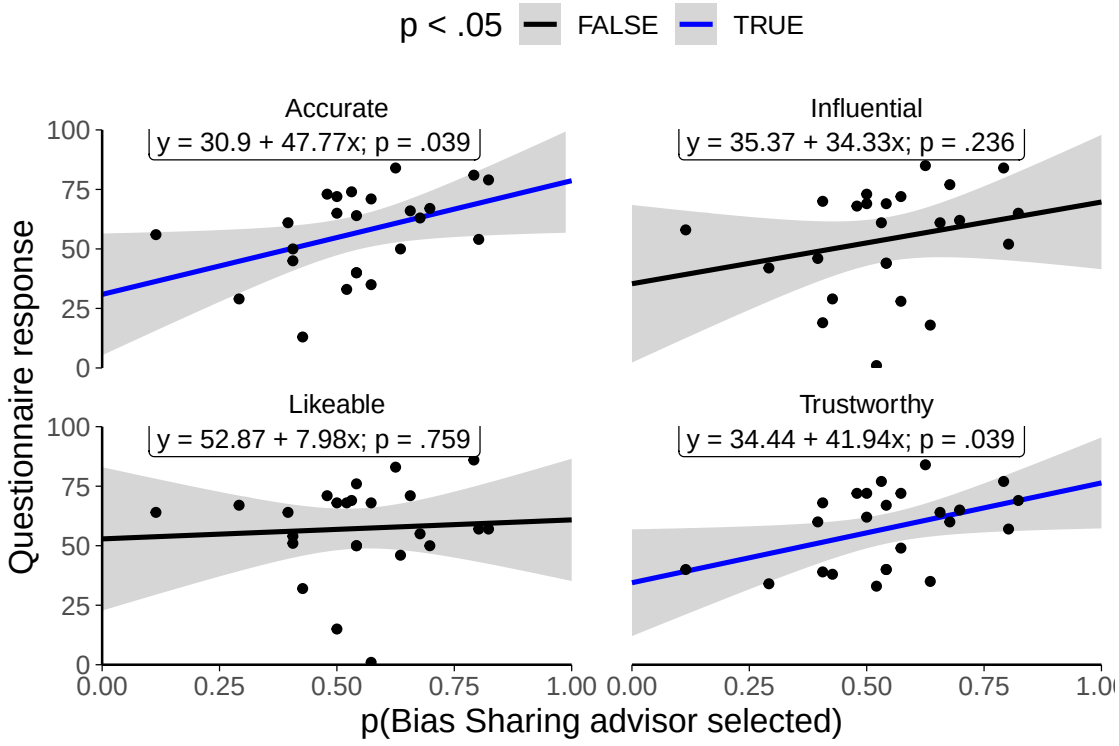


Figure 5.78: Behavioural and self-report consistency.
Relationship between questionnaire response scores on each item and overall pick rate for the Bias Sharing advisor. Lines show best-fit linear model and shaded areas give 95% confidence intervals for the parameters.

Questionnaire answers did not evolve systematically over the duration of the experiment, so the analysis was conducted on the final (post-experiment) answers. Correlations (Figure 5.78) indicated that the hypothesized effect was present for the questionnaire dimensions of accuracy and trustworthiness, but not for influence or likeability.

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `geom_smooth()` using formula 'y ~ x'
```

We supported this raw score analysis by comparing the pick rate for the Bias Sharing advisor (which is by construction comparative with the Anti Bias advisor) to a comparative version of the questionnaire responses obtained by subtracting scores for the Bias Sharing advisor from the equivalent score for the Anti Bias advisor. The results were highly similar (Figure 5.79), with the explicit trust measures for the Bias Sharing advisor relative to the Anti Bias advisor correlating with the

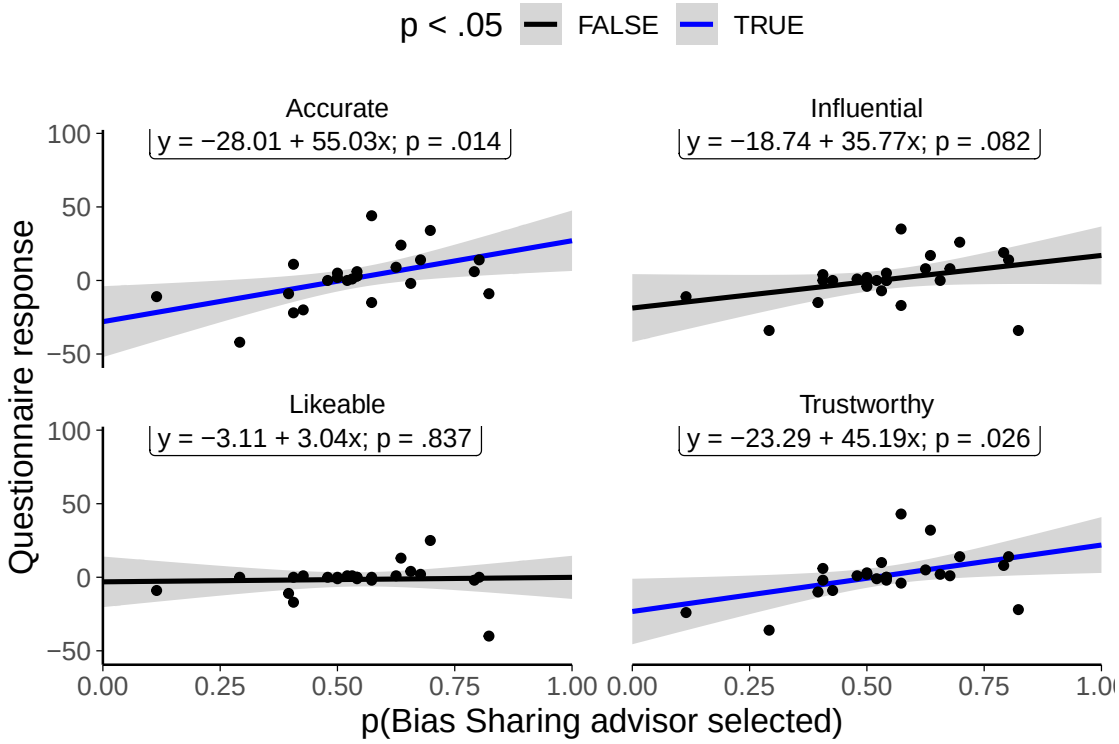


Figure 5.79: Behavioural and self-report consistency.
Relationship between questionnaire response scores on each item for the Bias Sharing advisor minus scores for the Anti Bias advisor and overall pick rate for the Bias Sharing advisor. Lines show best-fit linear model and shaded areas give 95% confidence intervals for the parameters.

picking preference for the Bias Sharing advisor in the dimensions of accuracy and trustworthiness. Again, neither influence nor likeability showed the effect.

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `geom_smooth()` using formula 'y ~ x'
## `summarise()` regrouping output by 'participantId', 'advisorId' (override with `
```

There was a strong correlation between trustworthiness and accuracy (r = .894), and thus it was probable that the variance in the questionnaire responses explained by advisor preference was principally driven by this shared component. This was borne out by nested multiple regression Table 5.23, which demonstrated that a model using accuracy to predict advisor preference was not significantly improved by including trustworthiness.

Table 5.23: Iterative model comparison predicting advisor choice from questionnaire scores

		Model			
		1	2	3	4
(Intercept)	β	0.54	0.54	0.53	0.53
	p	$< .001$	$< .001$	$< .001$	$< .001$
accurate	β	0.00	0.00	0.00	0.00
	p	.014	.284	.232	.246
trustworthy	β		0.00	0.00	0.00
	p		.808	.587	.603
influential	β			0.00	0.00
	p			.517	.782
likeable	β				0.00
	p				.370
F	df	1 / 22	2 / 21	3 / 20	4 / 19
	F	7.20	3.48	2.40	2.00
	p	.014	.050	.098	.136
R^2_{adj}	R^2_{adj}	.212	.177	.154	.148
	Δ		-.035	-.023	-.007
	F		0.06	0.43	0.84
	p		.811	.519	.370

Subjective and objective measures of influence Measures of influence were obtained in two distinct ways: through a self-report questionnaire and through observation of behaviour. To the extent that participants have an insight into their behaviour, these measures should correlate positively with one another. This proved to be the case (using the post-experiment responses for simplicity) for both the agree-in-confidence advisor and the agree-in-uncertainty advisor, as shown in Figure 5.80.

```
## `geom_smooth()` using formula 'y ~ x'
```

5.4.4 Discussion

The data are pretty equivocal here: there's no particularly solid reason for thinking that participants select Bias Sharing advisors more frequently than Anti Bias advisors.

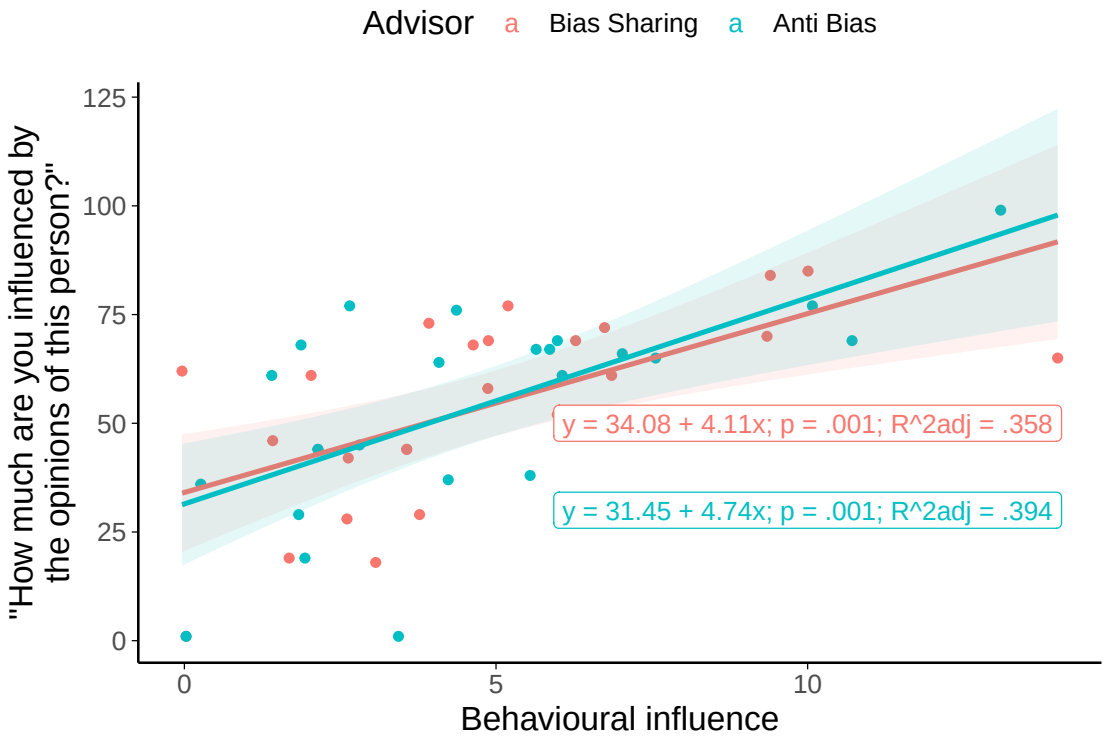


Figure 5.80: Behavioural and self-report measures of influence.
Scatterplot showing correlations between influence as measured by the change in confidence following advice vs self-report.

5.5 General discussion

The patterns observed for advice taking are also evident in advisor selection. Modelling work !TODO[lots of citations for this] indicates that biased source selection can dramatically reshape communication networks and create echo chamber effects where accurate but unpalatable information is ignored. Empirical research on source selection behaviour has found relatively little indication that people behave this way in the real world !TODO[source selection experiments citations], but here we are at least able to demonstrate in principle a psychological mechanism which could drive biased source selection. Furthermore, the mechanisms which produce biased source selection are rational and appropriate given the information available to participants.

5.5.1 Advisor choice results

The preferences for advisors were broadly consistent with the pattern expected from previous work on advisor influence. Where objective feedback could be used to calculate advisor performance, participants showed a systematic preference for picking the advisor who would provide the most accurate advice. Where objective feedback was unavailable, participants only showed a clear preference for High agreement advisors over Low agreement advisors (although in the Dates task there was also a preference for High accuracy over Low accuracy advisors). These results are consistent with an account of advisor trust updating which uses agreement as a proxy for advisor accuracy when more reliable information is not available.

5.5.2 Differences between tasks

The studies here explore the different manipulations of advice profile using two different tasks: a Dots task with a perceptual decision and a Dates task using an estimation decision. These two tasks have consistent results, but there are nevertheless identifiable differences between the tasks. Most notably, although advisor choice distributions in both tasks are roughly normally distributed, those in the Dots task results are sharper. This sharpness may be a result of many participants selecting advisors at approximately equal rates. If participants become bored or fatigued by the experiment they may disengage and select advisors in a random manner. !TODO[Did we restore advisor agreement/accuracy/etc to a baseline during Test phase? If so participants might detect a return to baseline from their favourite advisor and explore away.]

Where manipulations are effective in the Dots task (§5.3.1) they change the direction of preferences: a good many participants continue to pick advisors at approximately equal rates, but preferences in those who do express a preference systematically favour one particular advisor. Systematic differences in the Dates task are signified by distributions in which the modal preference moves to an extreme preference for the relevant advisor while the tails of the distribution

continue to cover the whole range. These differences are likely a consequence of the difference in the number of Choice trials in each Task: because the Dates task only has around 10 Choice trials it is relatively common for participants to select a single preferred advisor repeatedly, while participants in the Dots task may become tired of seeing advice from the same advisor over the !TODO[how many Choice trials in Dots?] Choice trials in the Dots task, increasing the novelty value of advice from the non-preferred advisor and thus reducing the apparent strength of preference. Long sentence: revise

The only case in which results between the No feedback condition of the Dates task and the Dots task were in conflict was High vs Low accuracy advisors. Participants in the Dots task had a systematic preference for the High accuracy advisor (§5.1.2). These differences are likely caused by the greater accuracy participants display in the Dots task making the experienced agreement profiles of the advisors more distinct. This explanation is wholly consistent with an account of advisor trust which uses agreement as a proxy for accuracy when better information is unavailable. This explanation entails a prediction that longer-term experience with the High accuracy advisor should eventually engender a systematic preference for that advisor, provided participants' guesses are more accurate than chance.

5.5.3 Strengths

5.5.4 Limitations

5.5.5 Implications for modelling work

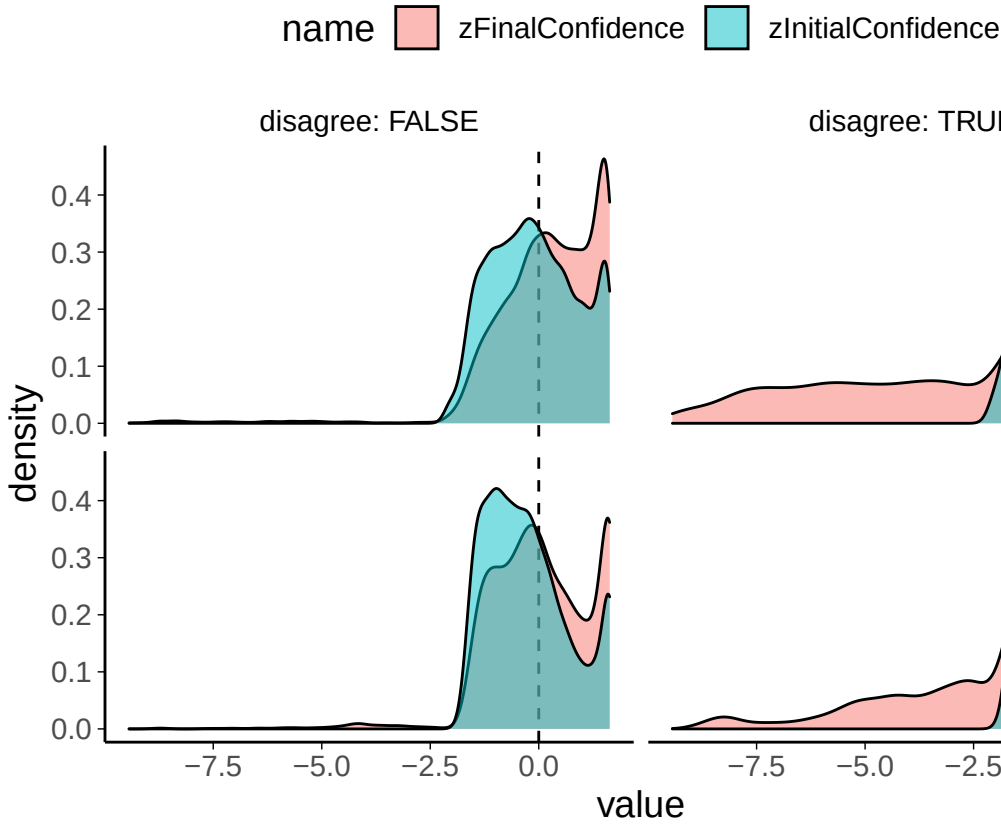
6

Modelling advice-taking behaviour

First we put all the data into the workspace.

Next we grab the data from the binary tasks. This includes all the dotstask data and a small portion of the datequiz data. !TODO[Later we can try to join them using an interoperable measure of influence.]

For what follows, we standardize initial estimate confidence and final decision confidence. The initial estimate confidence is standardized using its own mean and standard deviation. The final decision confidence is first oriented to the initial decision, meaning that it can take negative values where the endorsed answer switches from one side to the other. Increasingly negative values represent increasingly confident endorsement of the answer endorsed in the final decision. These final decision confidence values are then standardized using the mean and standard deviation from the initial estimate confidence. This means that the final decision confidence readings are z-scores based on the distribution for initial estimate confidence. It also means some of the values are much higher or lower than would be expected from the initial estimate confidence distribution because the range of possible values is wider (because negative values are possible in the final decision confidence).



binary task data-1.pdf

The distribution of advisor influence by agreement shows that final decision confidence in the initially-chosen answer is usually lower following disagreement from the advisor. This makes sense, and is to be expected. There are also a reasonable number of trials on which disagreeing advice leads to a modest *increase* in confidence.

Following agreement by an advisor, final decision confidence sometimes increases quite dramatically, but usually changes are modest and somewhat likely to be negative. Very dramatic negative shifts in confidence are rare following agreement.

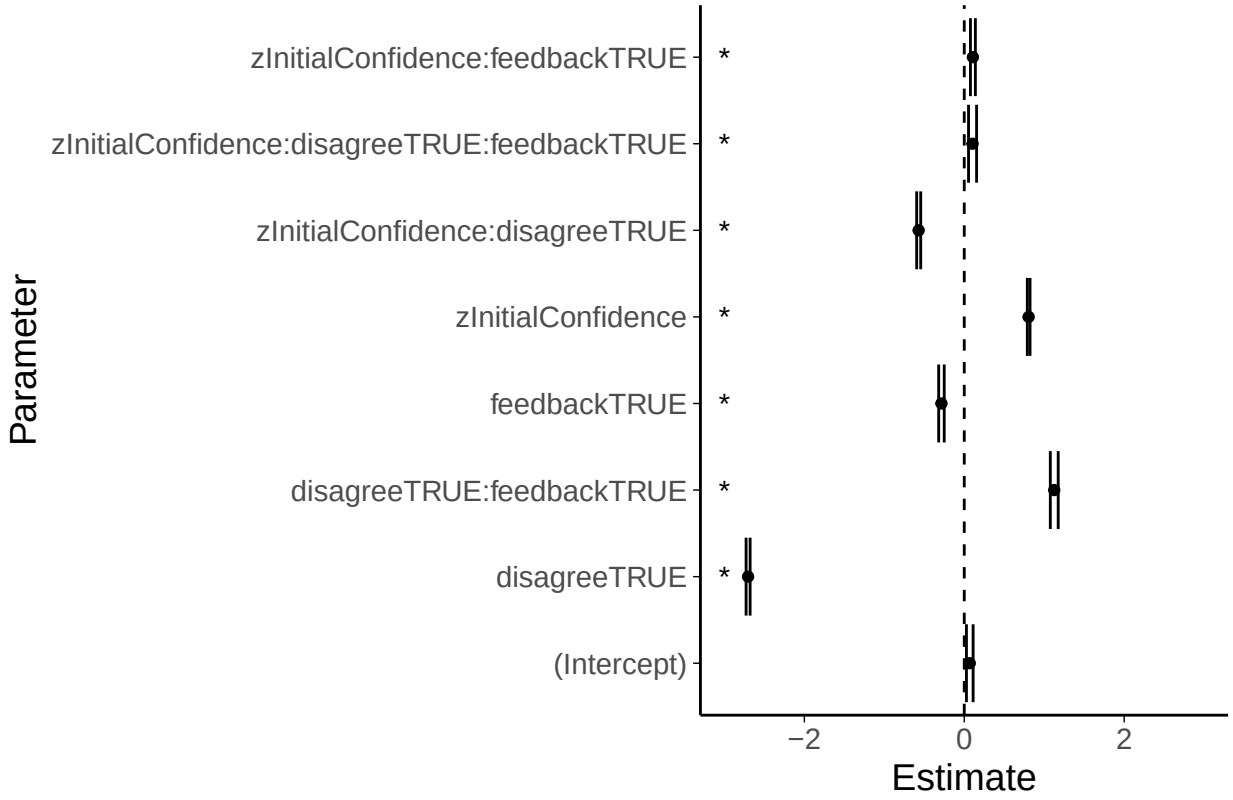
First, we run a simple multivariate linear regression predicting the final decision confidence of advice from initial estimate confidence and whether there is agreement between advice and initial estimate.

```
## Loading required package: lme4
##
## Attaching package: 'lmerTest'
## The following object is masked from 'package:lme4':
```

```
##
##     lmer
## The following object is masked from 'package:stats':
##
##     step
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: zFinalConfidence ~ zInitialConfidence * disagree * feedback +
##     (1 | pid)
##    Data: df
##
## REML criterion at convergence: 120176.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.2787 -0.2854  0.0162  0.4825  4.0975
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pid      (Intercept) 0.5301   0.7281
##  Residual             2.9263   1.7106
## Number of obs: 30487, groups:  pid, 361
##
## Fixed effects:
##                                             Estimate Std. Error         df
## (Intercept)                                6.896e-02  4.166e-02  3.889e+02
## zInitialConfidence                         8.055e-01  1.694e-02  2.882e+04
## disagreeTRUE                              -2.703e+00  2.512e-02  3.033e+04
## feedbackTRUE                              -2.852e-01  3.460e-02  2.943e+04
## zInitialConfidence:disagreeTRUE           -5.699e-01  2.525e-02  3.027e+04
```

```
## zInitialConfidence:feedbackTRUE               1.083e-01  3.087e-02  3.047e+04
## disagreeTRUE:feedbackTRUE                     1.126e+00  4.954e-02  3.018e+04
## zInitialConfidence:disagreeTRUE:feedbackTRUE  1.040e-01  5.051e-02  3.016e+04
##                                                t value Pr(>|t|)
## (Intercept)                                      1.655 0.098657 .
## zInitialConfidence                              47.542  < 2e-16 ***
## disagreeTRUE                                  -107.638  < 2e-16 ***
## feedbackTRUE                                    -8.244  < 2e-16 ***
## zInitialConfidence:disagreeTRUE                -22.567  < 2e-16 ***
## zInitialConfidence:feedbackTRUE                  3.508 0.000453 ***
## disagreeTRUE:feedbackTRUE                       22.723  < 2e-16 ***
## zInitialConfidence:disagreeTRUE:feedbackTRUE     2.058 0.039553 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##                   (Intr) zIntlC dsTRUE fdTRUE zIntlCnfdnc:dTRUE
## zIntlCnfdnc        0.021
## disagreTRUE       -0.220 -0.027
## feedbckTRUE       -0.125 -0.088  0.207
## zIntlCnfdnc:dTRUE -0.020 -0.489  0.131  0.018
## zIntlCnfdnc:fTRUE -0.009 -0.497  0.016  0.199  0.266
## dsTRUE:TRUE        0.112  0.017 -0.507 -0.418 -0.066
## zIC:TRUE:TR        0.011  0.247 -0.065 -0.095 -0.500
##                   zIntlCnfdnc:fTRUE dTRUE:
## zIntlCnfdnc
## disagreTRUE
## feedbckTRUE
## zIntlCnfdnc:dTRUE
## zIntlCnfdnc:fTRUE
```

```
## dsTRUE:TRUE       -0.095
## zIC:TRUE:TR       -0.490              0.252
```



DV = standardized directional final decision confidence; * p < .05

Where the advisor disagrees with a participant, the participant's final decision confidence is substantially lower on average than the average for initial estimate confidence. There is also a small effect of initial confidence, indicating that a participant who is initially more confident is likely to retain that increased confidence in their final decision. The interaction between the parameters indicates that the retention of initial estimate confidence is largely cancelled out by disagreement from an advisor: there is a small net positive effect from initial confidence on final confidence even where the advisor disagrees.

The presence of feedback on a trial is better regarded as a noisy proxy for feedback on previous trials than as contributing meaningfully to the final decision on any particular trial (because feedback happens *after* the final decision has been made). Feedback slightly reduces the confidence of final decisions, and noticeably

reduces the final confidence penalty from disagreeing advice. Slightly more initial confidence is retained into final decision confidence in feedback trials.

Next, we can look at the effect of adding in parameters based on the historical interactions between the advisor and the participant. The two main effects we look at are the total amount of experience a participant has had with an advisor, and the agreement rate experienced in that time.

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: zFinalConfidence ~ zInitialConfidence * disagree * feedback *
##     nthWithAdvisor * agreeRate + (1 | pid)
##    Data: dfh
##
## REML criterion at convergence: 119701.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.2109 -0.2768  0.0220  0.4603  4.1303
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pid      (Intercept) 0.4919   0.7014
##  Residual             2.8796   1.6970
## Number of obs: 30487, groups:  pid, 361
##
## Fixed effects:
##                                                                    Estimate
## (Intercept)                                                       7.091e-01
## zInitialConfidence                                                1.237e+00
## disagreeTRUE                                                     -4.221e+00
## feedbackTRUE                                                     -2.151e-01
```

```
## nthWithAdvisor                                                    2.765e-01
## agreeRate                                                        -8.269e-01
## zInitialConfidence:disagreeTRUE                                  -1.468e+00
## zInitialConfidence:feedbackTRUE                                   4.877e-02
## disagreeTRUE:feedbackTRUE                                        -7.684e-01
## zInitialConfidence:nthWithAdvisor                                 3.468e-01
## disagreeTRUE:nthWithAdvisor                                      -3.208e-01
## feedbackTRUE:nthWithAdvisor                                       4.834e-01
## zInitialConfidence:agreeRate                                     -5.967e-01
## disagreeTRUE:agreeRate                                            2.363e+00
## feedbackTRUE:agreeRate                                           -5.993e-02
## nthWithAdvisor:agreeRate                                         -4.482e-01
## zInitialConfidence:disagreeTRUE:feedbackTRUE                     -1.709e+00
## zInitialConfidence:disagreeTRUE:nthWithAdvisor                   -4.100e-01
## zInitialConfidence:feedbackTRUE:nthWithAdvisor                    3.649e-01
## disagreeTRUE:feedbackTRUE:nthWithAdvisor                         -3.798e+00
## zInitialConfidence:disagreeTRUE:agreeRate                         1.416e+00
## zInitialConfidence:feedbackTRUE:agreeRate                         8.699e-02
## disagreeTRUE:feedbackTRUE:agreeRate                               2.478e+00
## zInitialConfidence:nthWithAdvisor:agreeRate                      -4.522e-01
## disagreeTRUE:nthWithAdvisor:agreeRate                             6.696e-01
## feedbackTRUE:nthWithAdvisor:agreeRate                            -4.830e-01
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor      -3.336e+00
## zInitialConfidence:disagreeTRUE:feedbackTRUE:agreeRate            2.377e+00
## zInitialConfidence:disagreeTRUE:nthWithAdvisor:agreeRate          4.533e-01
## zInitialConfidence:feedbackTRUE:nthWithAdvisor:agreeRate         -4.320e-01
## disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate                5.018e+00
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate  4.609e+00
##                                                                  Std. Error
## (Intercept)                                                       1.697e-01
```

```
## zInitialConfidence                                                1.517e-01
## disagreeTRUE                                                      2.569e-01
## feedbackTRUE                                                      2.895e-01
## nthWithAdvisor                                                    2.157e-01
## agreeRate                                                         2.236e-01
## zInitialConfidence:disagreeTRUE                                   2.695e-01
## zInitialConfidence:feedbackTRUE                                   2.941e-01
## disagreeTRUE:feedbackTRUE                                         4.947e-01
## zInitialConfidence:nthWithAdvisor                                 2.014e-01
## disagreeTRUE:nthWithAdvisor                                       3.207e-01
## feedbackTRUE:nthWithAdvisor                                       4.105e-01
## zInitialConfidence:agreeRate                                      2.032e-01
## disagreeTRUE:agreeRate                                            3.535e-01
## feedbackTRUE:agreeRate                                            3.938e-01
## nthWithAdvisor:agreeRate                                          2.851e-01
## zInitialConfidence:disagreeTRUE:feedbackTRUE                      5.183e-01
## zInitialConfidence:disagreeTRUE:nthWithAdvisor                    3.399e-01
## zInitialConfidence:feedbackTRUE:nthWithAdvisor                    4.233e-01
## disagreeTRUE:feedbackTRUE:nthWithAdvisor                          6.419e-01
## zInitialConfidence:disagreeTRUE:agreeRate                         3.705e-01
## zInitialConfidence:feedbackTRUE:agreeRate                         3.993e-01
## disagreeTRUE:feedbackTRUE:agreeRate                               6.937e-01
## zInitialConfidence:nthWithAdvisor:agreeRate                       2.669e-01
## disagreeTRUE:nthWithAdvisor:agreeRate                             4.266e-01
## feedbackTRUE:nthWithAdvisor:agreeRate                             5.510e-01
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor       6.762e-01
## zInitialConfidence:disagreeTRUE:feedbackTRUE:agreeRate            7.207e-01
## zInitialConfidence:disagreeTRUE:nthWithAdvisor:agreeRate          4.526e-01
## zInitialConfidence:feedbackTRUE:nthWithAdvisor:agreeRate          5.675e-01
## disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate                8.702e-01
```

```
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate  9.143e-01
##                                                                              df
## (Intercept)                                                             2.359e+04
## zInitialConfidence                                                      3.029e+04
## disagreeTRUE                                                            3.016e+04
## feedbackTRUE                                                            3.036e+04
## nthWithAdvisor                                                          3.045e+04
## agreeRate                                                               3.038e+04
## zInitialConfidence:disagreeTRUE                                         3.015e+04
## zInitialConfidence:feedbackTRUE                                         3.020e+04
## disagreeTRUE:feedbackTRUE                                               3.011e+04
## zInitialConfidence:nthWithAdvisor                                       3.022e+04
## disagreeTRUE:nthWithAdvisor                                             3.016e+04
## feedbackTRUE:nthWithAdvisor                                             3.021e+04
## zInitialConfidence:agreeRate                                            3.028e+04
## disagreeTRUE:agreeRate                                                  3.016e+04
## feedbackTRUE:agreeRate                                                  3.031e+04
## nthWithAdvisor:agreeRate                                                3.045e+04
## zInitialConfidence:disagreeTRUE:feedbackTRUE                            3.011e+04
## zInitialConfidence:disagreeTRUE:nthWithAdvisor                          3.015e+04
## zInitialConfidence:feedbackTRUE:nthWithAdvisor                          3.015e+04
## disagreeTRUE:feedbackTRUE:nthWithAdvisor                                3.011e+04
## zInitialConfidence:disagreeTRUE:agreeRate                               3.015e+04
## zInitialConfidence:feedbackTRUE:agreeRate                               3.019e+04
## disagreeTRUE:feedbackTRUE:agreeRate                                     3.012e+04
## zInitialConfidence:nthWithAdvisor:agreeRate                             3.022e+04
## disagreeTRUE:nthWithAdvisor:agreeRate                                   3.016e+04
## feedbackTRUE:nthWithAdvisor:agreeRate                                   3.020e+04
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor             3.010e+04
## zInitialConfidence:disagreeTRUE:feedbackTRUE:agreeRate                  3.011e+04
```

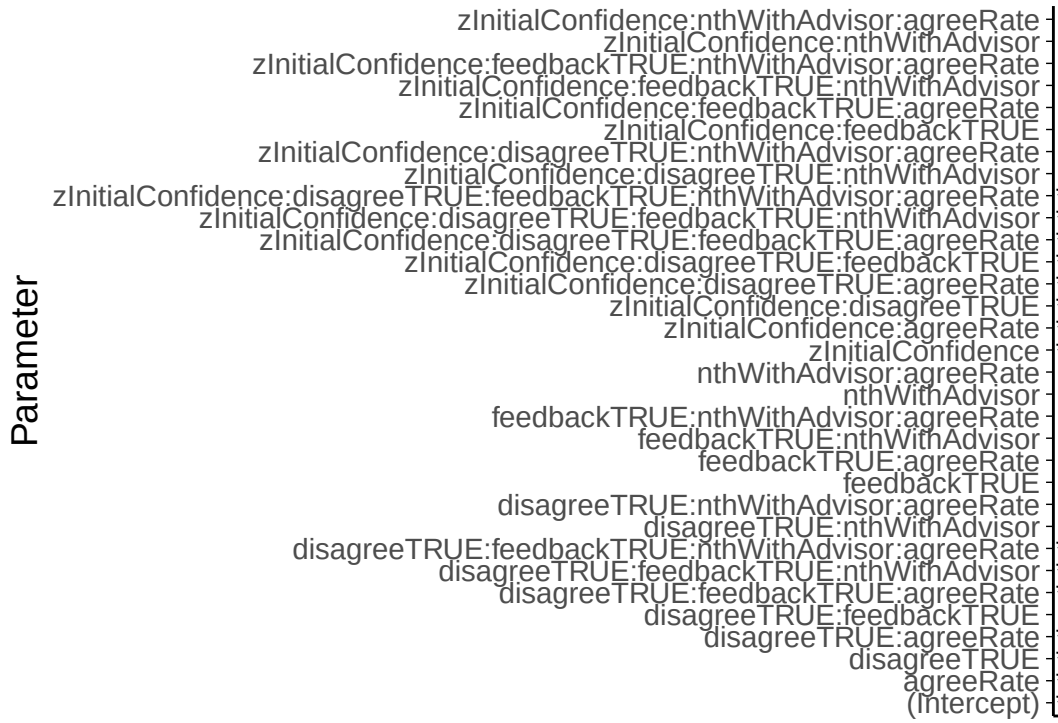
```
## zInitialConfidence:disagreeTRUE:nthWithAdvisor:agreeRate              3.015e+04
## zInitialConfidence:feedbackTRUE:nthWithAdvisor:agreeRate              3.015e+04
## disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate                    3.011e+04
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate 3.010e+04
##                                                                       t value
## (Intercept)                                                             4.179
## zInitialConfidence                                                      8.155
## disagreeTRUE                                                          -16.433
## feedbackTRUE                                                           -0.743
## nthWithAdvisor                                                          1.282
## agreeRate                                                              -3.699
## zInitialConfidence:disagreeTRUE                                        -5.448
## zInitialConfidence:feedbackTRUE                                         0.166
## disagreeTRUE:feedbackTRUE                                              -1.553
## zInitialConfidence:nthWithAdvisor                                       1.722
## disagreeTRUE:nthWithAdvisor                                            -1.000
## feedbackTRUE:nthWithAdvisor                                             1.177
## zInitialConfidence:agreeRate                                           -2.936
## disagreeTRUE:agreeRate                                                  6.684
## feedbackTRUE:agreeRate                                                 -0.152
## nthWithAdvisor:agreeRate                                               -1.572
## zInitialConfidence:disagreeTRUE:feedbackTRUE                           -3.297
## zInitialConfidence:disagreeTRUE:nthWithAdvisor                         -1.206
## zInitialConfidence:feedbackTRUE:nthWithAdvisor                          0.862
## disagreeTRUE:feedbackTRUE:nthWithAdvisor                               -5.916
## zInitialConfidence:disagreeTRUE:agreeRate                               3.822
## zInitialConfidence:feedbackTRUE:agreeRate                               0.218
## disagreeTRUE:feedbackTRUE:agreeRate                                     3.573
## zInitialConfidence:nthWithAdvisor:agreeRate                            -1.694
## disagreeTRUE:nthWithAdvisor:agreeRate                                   1.570
```

```
## feedbackTRUE:nthWithAdvisor:agreeRate                                -0.877
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor          -4.933
## zInitialConfidence:disagreeTRUE:feedbackTRUE:agreeRate                3.298
## zInitialConfidence:disagreeTRUE:nthWithAdvisor:agreeRate              1.002
## zInitialConfidence:feedbackTRUE:nthWithAdvisor:agreeRate             -0.761
## disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate                    5.766
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate  5.041
##                                                                    Pr(>|t|)
## (Intercept)                                                        2.94e-05
## zInitialConfidence                                                 3.63e-16
## disagreeTRUE                                                        < 2e-16
## feedbackTRUE                                                       0.457588
## nthWithAdvisor                                                     0.199998
## agreeRate                                                          0.000217
## zInitialConfidence:disagreeTRUE                                    5.13e-08
## zInitialConfidence:feedbackTRUE                                    0.868285
## disagreeTRUE:feedbackTRUE                                          0.120380
## zInitialConfidence:nthWithAdvisor                                  0.085018
## disagreeTRUE:nthWithAdvisor                                        0.317178
## feedbackTRUE:nthWithAdvisor                                        0.239023
## zInitialConfidence:agreeRate                                       0.003327
## disagreeTRUE:agreeRate                                             2.37e-11
## feedbackTRUE:agreeRate                                             0.879042
## nthWithAdvisor:agreeRate                                           0.116020
## zInitialConfidence:disagreeTRUE:feedbackTRUE                       0.000978
## zInitialConfidence:disagreeTRUE:nthWithAdvisor                     0.227763
## zInitialConfidence:feedbackTRUE:nthWithAdvisor                     0.388694
## disagreeTRUE:feedbackTRUE:nthWithAdvisor                           3.34e-09
## zInitialConfidence:disagreeTRUE:agreeRate                          0.000133
## zInitialConfidence:feedbackTRUE:agreeRate                          0.827537
```

```
## disagreeTRUE:feedbackTRUE:agreeRate                                    0.000354
## zInitialConfidence:nthWithAdvisor:agreeRate                            0.090253
## disagreeTRUE:nthWithAdvisor:agreeRate                                  0.116519
## feedbackTRUE:nthWithAdvisor:agreeRate                                  0.380625
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor            8.14e-07
## zInitialConfidence:disagreeTRUE:feedbackTRUE:agreeRate                 0.000976
## zInitialConfidence:disagreeTRUE:nthWithAdvisor:agreeRate               0.316561
## zInitialConfidence:feedbackTRUE:nthWithAdvisor:agreeRate               0.446511
## disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate                     8.18e-09
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate 4.65e-07
##
## (Intercept)                                                            ***
## zInitialConfidence                                                     ***
## disagreeTRUE                                                           ***
## feedbackTRUE
## nthWithAdvisor
## agreeRate                                                              ***
## zInitialConfidence:disagreeTRUE                                        ***
## zInitialConfidence:feedbackTRUE
## disagreeTRUE:feedbackTRUE
## zInitialConfidence:nthWithAdvisor                                      .
## disagreeTRUE:nthWithAdvisor
## feedbackTRUE:nthWithAdvisor
## zInitialConfidence:agreeRate                                           **
## disagreeTRUE:agreeRate                                                 ***
## feedbackTRUE:agreeRate
## nthWithAdvisor:agreeRate
## zInitialConfidence:disagreeTRUE:feedbackTRUE                           ***
## zInitialConfidence:disagreeTRUE:nthWithAdvisor
## zInitialConfidence:feedbackTRUE:nthWithAdvisor
```

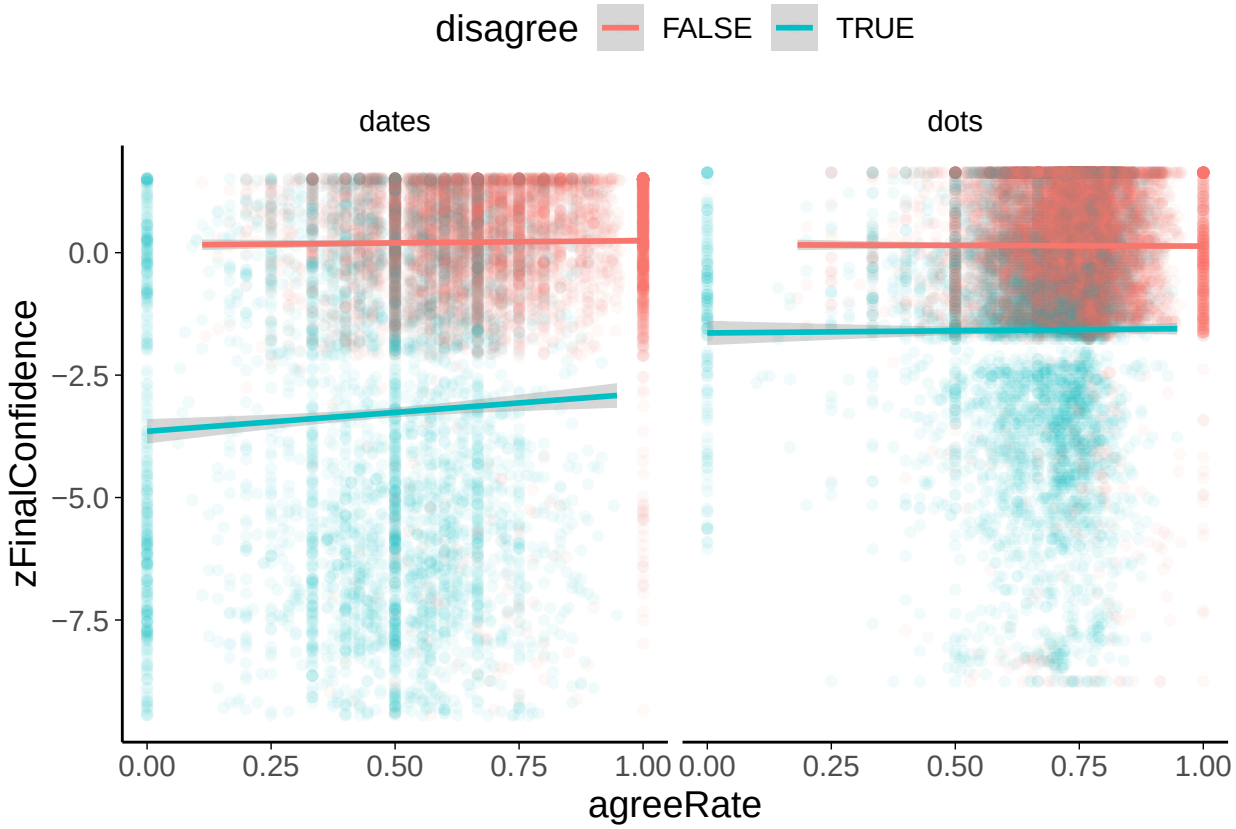
```
## disagreeTRUE:feedbackTRUE:nthWithAdvisor                                    ***
## zInitialConfidence:disagreeTRUE:agreeRate                                   ***
## zInitialConfidence:feedbackTRUE:agreeRate
## disagreeTRUE:feedbackTRUE:agreeRate                                         ***
## zInitialConfidence:nthWithAdvisor:agreeRate                                 .
## disagreeTRUE:nthWithAdvisor:agreeRate
## feedbackTRUE:nthWithAdvisor:agreeRate
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor                 ***
## zInitialConfidence:disagreeTRUE:feedbackTRUE:agreeRate                      ***
## zInitialConfidence:disagreeTRUE:nthWithAdvisor:agreeRate
## zInitialConfidence:feedbackTRUE:nthWithAdvisor:agreeRate
## disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate                          ***
## zInitialConfidence:disagreeTRUE:feedbackTRUE:nthWithAdvisor:agreeRate ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 32 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```



with history-1.pdf

```
## `geom_smooth()` using formula 'y ~ x'
```



Absent any effects, average final decision confidence is slightly higher than average initial estimate confidence in this model. As before, there is a pronounced effect of disagreement, with final confidence being notably lower following disagreement from the advisor, and a smaller effect whereby higher initial confidence increases final confidence.

The two history parameters, total experience with advisor and experienced agreement rate, go in opposite directions. The more experience a participant has had with an advisor, the higher their final confidence, while the greater the experienced agreement rate the lower the confidence. These effects are small, but puzzling; both advisor experience and experience of agreement should go the same way, and we might expect small positive effects (to be later reversed by interaction with disagreement). The interaction between experience of agreement and disagreement is also peculiar: the reduction in confidence from disagreement is largely cancelled out in an advisor who agrees very frequently, suggesting that disagreeing advice from a advisor who usually agrees is less rather than more influential. The history

parameters also interact with one another: the greater the experience with an advisor the greater the penalty to final confidence from increased experience of agreement.

Overall, the largest effects in the model come from some quite complex 3- and 4-way interactions, suggesting a nuanced structure to the data which is not well-captured simple effects. This may be because the biggest changes take place under specific conditions, or it may be an effect of the law of small numbers (because specific overlaps in conditions take place much more rarely than main effects, allowing outliers to produce large coefficient estimates). Some comfort can be taken in noting that, while the standard errors of the complex interactions with large coefficients are larger than those with smaller coefficients, they are not dramatically larger, and thus the larger coefficients are very likely to be genuinely large.

```
## refitting model(s) with ML (instead of REML)
## Data: dfh
## Models:
## m: zFinalConfidence ~ zInitialConfidence * disagree * feedback +
## m:     (1 | pid)
## mh: zFinalConfidence ~ zInitialConfidence * disagree * feedback *
## mh:     nthWithAdvisor * agreeRate + (1 | pid)
##    npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## m    10 120154 120238 -60067   120134
## mh   34 119670 119953 -59801   119602 532.65 24  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Do we want to take a descriptive model like this and use it directly as a predictive model in the network modelling (or use machine learning/decision tree/heterogeneous models)? Can we massage continuous dates task data to be compatible? Can we use a multilevel model with task[participant[trial]]] structure?

- Which parameters are important for making network agents display human-like psychology?

  - People choose in line with their experience (mostly)
    - Feedback is a big deal
    - Agreement less so?
  - Lots of variability within and between people
    - Individual differences are interesting - but what drives them/which dimensions matter?
  - What effects do all these have on the network structure?
    - Does heterogeneity lead to a stable network structure?
    - Do extremely opinionated individuals form 'echo chambers'?
      - Do they draw in others?
  - Other parameters such as
    - Choice availability
    - Hidden preference
      - How important is agree rate?
- Explore some of the parameters in these linear models
  - Visualise some of the raw data and the individual effects
- Individual differences might be able to be tracked with parameter-fitted decision-tree-type choice models between pick/average

Part III

Context of advice

7
Context of Advice

7.1 Egocentric discounting

Egocentric discounting (also known as egocentric 'advice discounting') is a phenomenon wherein advice is under-weighted during integration with the advice-taker's existing opinion, relative to a normative expectation. Most experiments that explore egocentric discounting use the Judge-Advisor System. The Judge-Advisor System has roles for a judge, usually the participant, and one or more advisors, often other participants. In a typical design, the judge offers an initial estimate for some decision, e.g. the total value of coins in a jar of change, then receives advice from the advisor, and then makes a final decision. The difference between the initial estimate and the final decision is taken as measure of how influential the advice was, typically expressed in terms of the contributions of the initial estimate and the advice to the final decision.

In these experiments, the task performance for the advisor is usually as good or better than that of the judge. This performance structure can be well captured for most tasks with Gaussian answer + error distribution where the answer supplies the mean and the error supplies the variance. When combining individual estimates from multiple distributions, optimal results are obtained by weighting the estimates according to the relative precision of their parent distributions (Soll and Larrick 2009;

Bahrami et al. 2010). This is analogous to multisensory integration (Ernst and Banks 2002; Körding et al. 2007), and many other cognitive processes argued to be modelled on Bayesian integration, such as those listed in Section 2 of **colomboBayesianCognitiveScience2015**. Where the performance of the advisor is higher than the judge, the advisor's error will be lower than the judge's, and thus the variance of the advisor's distribution will be narrower and therefore the precision of the advisor's distribution will be higher. When a judge combines their own estimate with that of an advisor who is at least as good, an optimal judge will weight the advisor's opinion at least as highly as their own. !TODO[We'll have got a whole section on models of advice integration stuff, probably, so we can point to that here.] The classic presentation of egocentric discounting is when, in these scenarios, the weight applied to the advice is lower than the optimal weight.

Egocentric discounting is a robust phenomenon in advice-taking. It is not a generic inability to combine estimates: people can accurately combine estimates that do not include their own opinion (Ilan Yaniv and Choshen-Hillel 2012), even adjusting for differences in ability between advisors (Soll and Mannes 2011). Similarly, Trouche et al. (2018) showed that when advice and initial estimates were surreptitiously switched, participants discounted their own initial estimates in favour of advice, suggesting that a person's own opinion has a privileged status.

In this chapter, I review the literature on egocentric discounting, with particular attention to manipulations that have been used and explanations that have been offered. I conclude the chapter by expounding an alternative perspective from which to view the phenomenon which, in my view, clarifies the phenomenon and opens the field for a wider range of effective explanations.

7.2 Manipulations affecting egocentric discounting

A sizeable body of research has been conducted into egocentric discounting, using a wide variety of manipulations. These manipulations fit roughly into four categories:

properties of the task, properties of the advice, properties of the advisor, and wider social factors.

Many of the experiments detailed below have not looked at egocentric discounting itself, but have instead looked at changes to the weight given to advice in integrated decisions. While egocentric discounting and advice weighting are inversely related, establishing the degree to which advice has been discounted requires a reference value from which its weighting has deviated. Many experiments do not calculate this normative value, nor even give sufficient details to establish a reference value, either an objective normative value or an expected value from the perspective of the participant.

One obvious reference value is the case where equal weight is given to one's own opinion and the advice received. This is often the value implicitly or explicitly stated. However, equal weighting is only normatively prescribed if advice is exactly as reliable as one's own initial estimate.

Nevertheless, it should be borne in mind, therefore, that depending on the circumstances, very high or low levels of advice weighting might not correspond to very low or high levels of discounting. When Schultze, Mojzisch, et al. (2017) provide participants with advisory estimates from a random number generator, ascribing any weight at all to the advice is suboptimal. Conversely there are many experiments in which expert advice is provided, and in these experiments weighting advice evenly corresponds to egocentric discounting because the advisor's estimates are more accurate on average.

As a consequence of the uncertainty about which normative weighting strategy is required by each experiment (and sometimes this strategy is different from the perspectives of the researcher and the participant), egocentric discounting is primarily examined here in terms of changes in advice weighting. Where advice weighting diminishes, egocentric discounting is said to increase, without specific comment being possible on the exact level of discounting on display. In studies where normative strategies can be determined (e.g. Ilan Yaniv and Kleinberger 2000),

advice weighting is below that predicted by the normative strategy, indicating egocentric discounting.

7.2.1 Task properties

The properties of the task chosen can affect the levels of egocentric discounting. Task difficulty is a major factor, perhaps mediated by the judge's confidence, but broader features also play a role, including how advice is provided and whether unified estimates are required at any point.

Task difficulty

The most prominent feature of the task which affects egocentric discounting is difficulty. Gino and Moore (2007) asked participants to estimate a person's weight from a clear (easy condition) or blurry (hard condition) picture, and saw less discounting on the hard task. Likewise, Wang and Du (2018) used blurring to increase the difficulty of estimating the number of coins in a photograph of a jar and found that participants discounted less in the blurry compared to the clear condition.

Judge's confidence

Wang & Du saw full mediation of their difficulty manipulation by participants' confidence on the task, while Gino and Moore saw only partial mediation. Other studies have manipulated the judge's confidence through other mechanisms. See et al. (2011) used a power manipulation which was effective in part through raising judges' confidence; and Gino, Brooks, et al. (2012) used anxiety manipulations to decrease judges' confidence. In both cases, partial or full mediation through confidence occurred such that participants' higher confidence in themselves and their decisions was associated with greater egocentric discounting. In many other experiments, including other experiments in Wang and Du (2018), confidence is not manipulated but is still associated with greater egocentric discounting. Using a similar methodology, but a different analytical approach, Moussaïd et al.

(2013) observed that highly confident participants rarely updated their views following advice.

Judge-Advisor System structure

More complex task designs, in which reflection and discussion are encouraged, can reduce discounting. Minson et al. (2011) and Liberman et al. (2012) asked dyads to take simultaneous roles as judge and advisor, providing initial estimates, exchanging advice during a discussion, and then providing final decisions on estimation tasks. Discounting was reduced, but still evident in this process, as it was in Van Swol (2011), which used a traditional Judge-Advisor System paradigm where advice was delivered face-to-face. Liberman and colleagues did manage to eliminate discounting where, between exchanging advice and providing a final decision, participants produced a single mutually satisfactory collaborative judgement, and showed that the value of this collaborative judgement was itself improved by open-minded discussion over justifying estimates or exchanging bids. Schultze, Mojzisch, et al. (2017) demonstrated that asking judges to selectively generate reasons why the advice might be correct or incorrect led to lower and higher levels of egocentric discounting respectively.

7.2.2 Advice properties

Several features of advice itself have been explored: the confidence of advice, its similarity to the initial decision, whether it is solicited, and the amount of it provided.

Confidence of the advice

When judges are more confident, they tend to be less influenced by advice, and the expected corollary of this is that when advice is expressed more confidently the advice will be more influential. Soll and Larrick (2009) measured the confidence of advice and saw that higher advice confidence was associated with higher influence of advice. Moussaïd et al. (2013) also found that differences in confidence between

judges' and advisors' estimates were useful in producing a decision tree determining the extent to which advice was taken.

Similarity of advice to the initial estimate

A frequently-manipulated property of advice, and the most interesting in the context of the first part of this thesis, is the similarity of the advice to the initial estimate. This is sometimes expressed as agreement or reasonableness of advice.[ˆNote that the influence of advice that is perfectly in accord with an initial estimate is undefined when using §Weight on Advice measures.] The evidence on the effects of advice distance on advice influence is equivocal. Some studies show that advice is less influential the further it is from the initial estimate, while other studies show a greater influence of more distant advice. Other studies have indicated that the relationship is quadratic: low weight is assigned to advice which is too near or far from the initial estimate, and a greater weight assigned to advice which is in the moderately distant.

Several studies have provided evidence for a simple agreement effect whereby advice that is nearer to the initial estimate is more influential. I. Yaniv (2004) manipulated advice to be nearer to or further away from the initial estimate and saw that the influence of advice decreased as the advice was further from the initial estimate (although this pattern did not hold for low-expertise judges in one experiment). Minson et al. (2011) found that more distant advice was associated with less advice-taking behaviour once average distance between dyad members was controlled for, although their results were not expressed using standard advice-taking metrics. Ilan Yaniv and Milyavsky (2007) observed that advice closer to the initial estimate was also more influential when combining multiple pieces of advice simultaneously.

The opposite effect was demonstrated by Hütter and Ache (2016). They found consistently higher influence for advice that was further from the initial estimate, both for single pieces of advice and for integrating multiple pieces of advice.

A non-linear, U-shaped relationship between advice distance and advice influence has been shown in other studies. Moussaïd et al. (2013) asked participants to

estimate answers to a variety of questions and give confidence ratings, both before and after receiving another person's initial estimate as advice. They identified a three-zone structure to the influence of advice according to the distance between the initial estimate and the advice. Similar advice fell into the 'confirmation zone', where opinion was unchanged but confidence increased; moderately distant advice fell into an 'influence zone' where opinion changed to accommodate the advice; and distant advice was generally ignored. Likewise, Schultze, Rakotoarisoa, et al. (2015) showed in an elegant series of Judge-Advisor System experiments that relationships between egocentric discounting and advice distance were U-shaped. Advice was given very little weight where it was close to the initial estimate, with the weight then increasing precipitously to a peak for middle-distance advice, after which it either remained stable or decreased slightly as the distance increased further. Schultze et al. also showed that confidence in final decisions was dramatically boosted by near advice, and that confidence gains decreased sharply with distance, consistent with Moussaïd et al.'s account.

Hütter & Ache used two discrete advice distances in each of their experiments, and thus could not demonstrate a U-shaped relationship with only two points. Hütter & Ache's results can be considered consistent with the studies demonstrating a U-shaped relationship, although there still remains a difficulty in integrating the results of these studies with those showing a monotonic decrease in advice weight by distance. This is consistent with the argument that high-confidence decisions reduce the processing of disconfirmatory evidence (Rollwage et al. 2020). This mechanism will function more effectively if there is a bias for processing nearby advice before more distant advice.

This mechanisms can be extended to explain the results in I. Yaniv (2004) (in which there is only a single piece of advice) by allowing the updating of confidence and the adjustment of the estimate to be separate processes. If updating is a two-stage process wherein confidence is updated first, and then the updated confidence value is used to determine the extent to which the estimate is adjusted, we can reproduce the U-shaped relationship from two monotonic relationships. First,

confidence is increased by nearby advice (and perhaps decreased by distant advice). Second, advice is more influential the nearer it is to the initial estimate. This means that, for nearby advice, the judge's confidence is boosted by agreement, meaning that the ostensibly highly-influential nearby advice is subsequently assigned very little weight (because highly confident judges ascribe little weight to advice - §Judge's confidence). For advice that lies at a middle distance, there is no boost to confidence and advice is somewhat influential. At greater distances, while the judge's confidence in their own opinion might be reduced slightly, there is little potential influence in the advice.

This unifying explanation is consistent with Moussaïd et al.'s demonstration that the effects of advice on confidence and estimate were somewhat separable. Nevertheless, it is somewhat unintuitive that the metacognitive effects of advice on confidence would happen prior to the cognitive effects of advice on estimate updating. The I. Yaniv (2004) results are also not wholly explained by this kind of process because Schultze, Rakotoarisoa, et al. (2015) included an experiment designed to reproduce Yaniv's methodology as closely as possible, the results of which were inconsistent with Yaniv's.

Related to the distance of advice is the reasonableness of advice, because where the judge has a somewhat reasonable estimate the distance serves as a reliable proxy for reasonableness. Gino, Brooks, et al. (2012) included an experiment in which (non-anxious) participants heavily discounted unreasonably high and unreasonably low advice, while discounting reasonable advice at a rate typically seen in Judge-Advisor System experiments. Similarly, Schultze, Mojzisch, et al. (2017) saw judges discount wildly implausible advice more heavily, although it was still assigned some weight, even when labelled as coming from a random number generator.

The level of agreement may act as a cue to the plausibility of both the advice and the initial estimate. Somewhat counter-intuitively, this can lead to agreeing advice being assigned less weight than we might expect because of its bolstering of the confidence in the initial estimate.

Solicitation of advice

The extent to which advice is discounted may also be related to whether advice is wanted. This is hard to disentangle from the effects of task difficulty, because people are more likely to seek advice when they find the task more difficult to do (and hence their confidence in their response is lower).

Gino and Moore (2007) compared advice-taking behaviour across two experiments using a task in which participants had to estimate people's weight from photographs. In one of these experiments participants received advice automatically and in the other participants had the option of clicking a button to receive advice. Participants opted to receive advice on almost all trials, including in an easy condition, and no differences were found in the extent to which advice was used between the compulsory and optional advice experiments. The very high rates of advice seeking in the optional experiment suggest that participants in the compulsory advice experiment may have been very welcoming of the advice due to the difficulty of the task. Gino (2008) showed that more expensive advice was sought less frequently but used more heavily, that expensive advice was used more heavily than free advice even when both are compulsory, and that paid-for advice was used more heavily than when the same advice was given for free as the result of a coin-flip. This study packaged questions together in blocks, and participants purchased advice for a whole block at once. This procedure means that the solicitation of advice is decoupled from the difficulty of the question on a trial-by-trial basis, although it is still likely that those participants who found all the questions in a block more difficult were more likely to solicit advice. In the real world, price is often an indicator of quality, albeit an imperfect one, and solicitation is likely to act as a proxy for confidence, which is again related to task difficulty: participants in the experiments may have taken more advice because they were less sure on the blocks in which they sought advice. !TODO[Michael and/or Naomi worked on seeking advice more when less confident? That's relevant here] Hütter and Ache (2016) found that participants opted to see a large number of advisory estimates in a calorific content estimation task when allowed to sample ad lib, although the influence of the advice was low.

Number of advisory estimates

Hütter and Ache (2016) !TODO[There are definitely others! Yaniv for a start? Find them.] saw levels of advice usage for multiple pieces of advice which were relatively similar to levels of advice usage for a single piece of advice. In other words, people integrating their own opinion with two advisory estimates tend to integrate the advisory estimates and then treat them as a single piece of advice for integrating with their own opinion. This may be an artefact of presentation because all these studies presented multiple estimates simultaneously as a list. If this is a real phenomenon, however, it is a critical bias: while sensible motivations for favouring one's own opinion over another's will be posited below, it is far harder to argue that the weight assigned to one's own opinion should remain the same whether being integrated with one other estimate or ten other estimates.

7.2.3 Advisor properties

A common strategy in advice-taking experiments is to manipulate the properties of the advisors (either within- or between-participants). Expertise is often manipulated, but some research has investigated factors such as the pre-existing relationship between judge and advisor, whether advisors are human or algorithmic, and whether advisors have a clear conflict of interest.

Expertise of advisors

By far the most frequently manipulated property of advisors is their ability to perform the task in question, known as their expertise. Advisor expertise can be communicated to participants in various ways, which can be broadly categorised into 'showing' approaches in which participants build up a picture of advisors' performance over a series of instances, and 'telling' approaches where participants are presented with a summary of an advisor's performance. In approaches towards the middle of the spectrum participants are offered evidence of the advisor's performance in one go, e.g. a scorecard showing an advisor's performance on practice questions. In some experiments advisor expertise manipulations are genuine, whereas in others

they are merely apparent and the participants receive equivalent advice regardless of the label or historical performance of the advisor. As noted above, the ability of the advisor to perform the task (relative to the judge) alters the optimal level of advice-taking according to the normative model. This means that discounting may still occur even when the advisor's estimate is weighted more highly than the judge's.

Ilan Yaniv and Kleinberger (2000) showed in a series of historical date estimation experiments that better-performing advisors were more influential than worse-performing advisors, especially where feedback was provided on judge's final decisions but also where it was not. Gino, Brooks, et al. (2012) used a coin-jar estimation task where participants were shown the advisor's past performance to demonstrate that people give the same advice more weight if it comes from an advisor with a history of good performance, although this effect was not visible in an anxiety arm of the experiment. Rakoczy et al. (2015) saw that 3-6 year-old children took advice from advisors who had named animals correctly more seriously than advice from advisors who acknowledged their own ignorance in a version of the Judge-Advisor System adapted for young children.

Sniezek, Schrah, et al. (2004) saw greater dependence on advice provided by specially-trained 'expert' peers, but only where the proportion of reward money for accurate judgement paid to the advisor had been decided *before* advice was provided. Soll and Larrick (2009) observed greater influence of more expert advisors across three experiments in which expertise was signalled by familiarity ratings with a university for which the graduate salary was being estimated, country of origin in a geography knowledge task, and confidence in a set of trivia questions. Tost et al. (2012) used a weight-estimation task and observed that greater weight was placed on advice from advisors labelled as experts compared to the same advice from advisors labelled as novices. Schultze, Mojzisch, et al. (2017) labelled advisors using a ranking system and saw that participants placed a higher weight on advice from highly-ranked advisors compared to low-ranked advisors, although the actual advice was (unbeknownest to the participants) the same. Wang and Du (2018) showed

that participants placed greater weight on advice from expert as opposed to novice advice in a coin-jar estimation task using advice that was genuinely expert or novice.

Önkal, Gönül, et al. (2017) reported a series of experiments using a stock price estimation task in which advisor expertise was manipulated using both labels and experience. Where experience alone was provided there was no clear effect of expertise, while labelling had a substantial effect. Further experiments crossing labelling (high or low expertise) with experience (of high or low expertise) indicated that participants responded to both labelling and experience in a rational way: the most influential advisors were both labelled and experienced as experts, followed by those labelled as having low expertise but demonstrating high expertise. Both advisors whose expertise was experienced as low were weighted less heavily, with no clear differences between them. [1]

Familiarity of advisors

So far as I can ascertain, no one has reported on a Judge-Advisor system in which comparable advice is received from friends and non-friends. Sniezek and Van Swol (2001) and Swol and Sniezek (2005) found a correlation between the amount classmates had interacted and ratings of trust in those classmates as advisors, but they did not use a measure of advice-taking which allows calculation of advice weighting. Minson et al. (2011) had long-term dance partners make estimates about their own dance performances in relation to professional assessment, and saw normal levels of egocentic discounting. [2]

It seems intuitive that advice will be weighted more highly when coming from people we know, but this does not appear to have been tested. With regard to the three-factor model of trust put forward in **mayerIntegrativeModelOrganizational1995**,

[1] Numerically, there was an indication of a backlash effect whereby the advisor labelled as high-expertise but demonstrating no expertise in practice was less influential than an advisor labelled as low-expertise and experienced as performing similarly.

[2] The partners' estimates in this study were highly correlated, meaning that they were not independent and thus did not necessarily bracket the correct answer. This means that both partners assigning positive weight to the other's advice is not likely to approach the correct answer.

knowing and trusting an advisor would be expected to increase the extent to which the advisor's advice is taken.

Humanity of the advisor

Some experimenters have examined advice weighting for non-human advisors. In a stock-market forecasting task, Önkal, Goodwin, et al. (2009) found that judges placed less weight on (identical) advice when it was labelled as coming from a statistical model versus a human expert forecaster. In their study on overweighting, Schultze, Mojzisch, et al. (2017) provided participants with advice labelled as coming from a random number generator and noted that its estimates were still assigned some weight by judges. The influence of this randomly-generated advice was roughly equivalent to that of human advisors not labelled as having high expertise.

Advisor conflict of interest

Where advisors have a conflict of interest, following the advice may benefit the advisor at the expense of the judge. Gino, Brooks, et al. (2012) observed that non-anxious judges assigned less weight to advice from advisors with a conflict of interest, while anxious judges assigned similar weight regardless of conflict of interest. Bonner and Cadman (2014) similarly saw less influence from advice from advisor with a conflict of interest in a CEO-remuneration task.

7.2.4 Wider social factors

The Judge-Advisor System in experiments is usually quite abstracted away from advice-taking in the real world, in which broader social factors are likely to offer a highly influential context dictating or suggesting norms for advice-taking. Despite being somewhat shielded from these broader factors, they are nevertheless apparent in some Judge-Advisor System experiments.

Fairness may be a fairly universal value !TODO[Do we need to back this up? There'll be some cross-cultural studies to reference, and even de Waal's research on fairness in non-human primates.], and it is certainly a lauded value in the developed

Western societies where the majority of Judge-Advisor System research takes place. Consistent with this, judges in the Sniezek, Schrah, et al. (2004) study tended to offer both novice and expert advisors equal shares in their reward money. This sense of fairness is seen to extend to advice-weighting, too. Harvey and Fischer (1997) saw expert judges consistently placed some weight on novice advice, and attributed this advice-taking behaviour to a social requirement for fairness. Likewise, Mahmoodi et al. (2015) showed that dyads making perceptual decisions would consistently over-weight the estimate of the less accurate dyad member despite evidence that this impaired performance and thus decreased the dyad's reward money.

Feeling powerful was observed to lead to decreases in advice-taking by See et al. (2011) and Tost et al. (2012), with the latter showing a partial mediation via the judge's confidence. Somewhat similarly, Gino, Brooks, et al. (2012) saw angry judges took less advice through being more self-confident, while anxious judges sought and took more advice through being less self-confident.

Finally, developmental maturity seems to be an important factor. Rakoczy et al. (2015) found that while 3-6 year old children did differentiate between knowledgeable and ignorant advisors, they nonetheless placed exceptionally high weight on advice from both advisors.

7.3 Purported explanations for egocentric discounting

Almost as numerous as the studies into egocentric discounting are the explanations offered to account for it. Despite this, no explanation has managed to withstand critical empirical scrutiny. Below, I offer a brief review of the explanations which have been put forward to date.

7.3.1 Egocentric bias

Among the earliest explanations for egocentric discounting was egocentric bias: the belief that one's judgement is superior to that of others. Harvey and Fischer

(1997) attributed egocentric discounting to such self-serving estimates of ability, such as when car drivers report on average being better than average (Svenson 1981). The explanation also suggested a mediating role of overconfidence, tying it neatly into later similar explanations from See et al. (2011), Gino, Brooks, et al. (2012), and Tost et al. (2012). For all these authors, a judge's confidence in the initial estimate is the principal driving force behind the weighting of advice. This view is appreciable from a Bayesian perspective on advice integration (Bahrami et al. 2010): as with multi-sensory integration (Fetsch et al. 2012), informational cues coalesce optimally onto the correct answer where they are weighted according to their precision. In such a framework, there is an optimal (Bayesian) integration process which is being fed faulty inputs (because the judge's own estimate is believed to have erroneously high precision).

Overconfidence certainly seems to have a role, and accounts well for many of the nuances in advice-taking experiments including the findings that advice is taken more readily for difficult tasks (where judge confidence is lower) and that judges made to feel more powerful (and hence more confident) take less advice. It is unlikely to be the whole story, however: Trouche et al. (2018) note that Soll and Mannes (2011) were able to disentangle ratings of ability and advice-taking behaviour and saw that the egocentric discounting occurred even if the judge's assessment of relative ability were taken as true.

7.3.2 Access to reasons

One of the most influential explanations of egocentric discounting has been Yaniv's argument that judges have greater access to the reasons justifying their own decisions than to those justifying the decisions of others, due to the opaqueness of other minds (Ilan Yaniv and Kleinberger 2000). This differential access to reasons suggests, from the perspective of the judge, that there is greater evidence favouring the judge's own opinion than favouring the estimate of their advisor, and that, analogous to the confidence case above, the more well-supported opinion should be given a greater weight during integration.

Trouche et al. (2018) presented judges with their own estimates labelled as advice, and with advice labelled as their own estimates, and observed that judges persisted in placing greater weight on what they *believed* was their own initial estimate rather than what was *actually* their own initial estimate. This result is a serious problem for the access-to-reasons explanation, because there is no good reason why simply relabelling estimates should change the judge's internal census of evidence supporting the estimates.

7.3.3 Anchoring

Some researchers have argued that egocentric discounting can be explained by anchoring. Anchoring is a well-established phenomenon whereby numbers clearly unrelated to a numerical estimation task can nevertheless bias estimates, as when participants asked to estimate the height of Mount Everest in feet but first asked to say whether '2,000' is higher or lower than the correct value (12,000) give far lower estimates than participants asked to say whether '45,500' is higher or lower than the correct value (Jacowitz and Kahneman 1995). Bonner and Cadman (2014) suggested that anchoring was responsible for judges' over-use of outlandishly extravagant suggestions for CEO remuneration. In a more thorough set of studies, Schultze, Mojzisch, et al. (2017) observed a consistently greater-than-zero weight on transparently useless advice, which they ascribed to anchoring to the advice. While these cases for anchoring may be admitted, they are to do with over-weighting advice rather than the under-weighting advice which characterises egocentric discounting. This is because the putative anchor is the advisory estimate.

Historically, it has been suggested that the judge's initial decision acts as an anchor (Harvey and Fischer 1997), but this was later ruled out when Harvey and Harries (2004) demonstrated egocentric discounting persevered when the labels for the judge's initial estimate and the advisor's advice were switched. If anchoring to the initial estimate were responsible for egocentric discounting the relative weighting of initial estimate and advice would have followed the actual values rather than

the labelled values, whereas the data showed that discounting occurred towards the *labelled* rather than *actual* initial estimate.

7.3.4 Sunk costs

Another general cognitive bias recruited as an explanation for egocentric discounting is the sunk costs fallacy, in which one perseveres with a poor strategy in order to justify the cost or effort that has already gone into pursuing it. Interestingly, sunk costs have been recruited to explain both egocentric discounting *and* following advice.

Gino (2008) had participants receive advice for free or for a fee depending upon the outcome of a coin flip. They found that the same advice was more influential where payment had been taken, and that the more expensive the more influential it was whether paid for or free. Assuming that participants viewed the greater cost as a marker of quality, the remaining effect contingent on whether or not payment had actually been taken can be explained by sunk costs. Similarly, Sniezek, Schrah, et al. (2004) found that, at least for expert advisors, judges who allocated a portion of their prospective reward money to the advisor *before* receiving the advice placed more weight on that advice.

Ronayne and Sgroi (2018) also invoked the sunk costs fallacy, but suggested that it could account for using less advice, i.e. discounting. These authors presented participants with the opportunity to use another participant's results rather than their own in a reward lottery, which they somewhat oddly labelled 'advice'. Despite this 'advice' being transparently better, participants frequently chose to keep their own results rather than switch, and this finding is explained on the basis that those participants were loath to forfeit the work they had done to obtain their own results by adopting another's. Extended to the Judge-Advisor System, this explanation would predict that more effortful tasks would lead to greater egocentric discounting. This is an intriguing prediction, but perhaps because effortfulness tends to covary with difficulty (which in turn decreases egocentric discounting), it does not appear to have been studied.

7.3.5 Naïve realism

A third cognitive bias invoked to explain egocentric discounting is naïve realism. Naïve realism occurs when people treat their own perceptions as reflective of shared underlying reality and others' perceptions as misguided or biased to the extent that they do not agree. Minson et al. (2011) argue strongly for a naïve realism explanation of egocentric discounting, although it is never fully explained why, on a naïve realism view, *any* adjustment to advice is warranted (because ex hypothesi the initial estimate reflects the true answer). The most creative element of the experiments involves funnelling integrated decisions (corresponding to judges' final decisions in the typical Judge-Advisor System) through a collaborative joint decision process before extracting a final individual decision. Naïve realism once again fails to provide a compelling explanation why the final individual decisions closely reflect the collaborative join decisions rather than the initial estimates: rather than continuing to endorse their own view of 'reality', participants appeared to be willing to accept the joint view once it had been established.

7.3.6 Responsibility / feeling of deserving outcomes

Some advice taking may be explicable on the basis of responsibility sharing. Harvey and Fischer (1997), Mahmoodi et al. (2015), and Ronayne and Sgroi (2018) have all suggested that taking advice can transfer some of the responsibility for the outcome of a decision onto the advisor. Where uncertainty is high, or where rewards are shared, this can be particularly useful.

While distribution of responsibility is more a reason for *reduced* rather than *increased* egocentric discounting, when combined with an account that predicts ascribing very low or no weight to advice by default (e.g. naïve realism), it can explain why advice weighting is higher than the zero that would be expected. It seems more plausible, however, that factors which promote advice-taking such as fairness, advisor expertise, and distribution of responsibility serve to place a

limit on egocentric discounting, rather than that complete discounting is a default strategy from which these factors move judges.

7.3.7 Wariness

Trouche et al. (2018) designate the above explanations 'proximal explanations', because they offer a mechanistic account of how discounting occurs. Rather than partaking in this discussion, they instead provide an 'ultimate explanation', which may explain why discounting occurs. They suggest that discounting occurs because advisors' interests do not always align with judges', and thus some level of discounting offers protection from relying too heavily on advice that may be deliberately harmful. They suggest thus that discounting is an evolved response to misaligned incentives between judges and advisors. The account I offer below is in the mode of this explanation.

7.4 A wider view of egocentric discounting

The explanations outlined above are all explanations pitched at the level of are all offered as explanations for a deviation from normative optimality. I have chosen to take a different approach, asking instead under which circumstances the observed behaviour would be an optimal policy, and exploring the plausibility of those circumstances continuing to have an influence in experimental settings where the normative behaviour might be averaging one's own opinion with advice.

As a starting point, I note that if one tells anyone who is not an advice-taking researcher that people do not take others' opinions as seriously as they take their own when making decisions, the response is likely to be a flat "of course", perhaps accompanied by a perplexity as to why such an obvious statement is being presented as a valuable insight. The approach taken here works to codify this intuition as a set of hyperpriors: expectations about the relative utility of advice as compared to one's own opinion. I argue that the normative model of advice-taking, and the pared-down experimental design with which is it entwined, seek to take the

situation out of the evolutionary history of advice, but cannot take the evolutionary history of advice out of the situation.

According to this hypothesis, the hyperpriors on advice-taking are a consequence of the many opportunities for deception and misunderstanding which apply to advice but not to one's own opinion. The most obvious of these is the opportunity for deception. As Trouche et al. (2018) point out, advisors do not always have the judge's best interests at heart when supplying advice. Consider, for example, a situation where one coworker, Sally, asks another, Hanan, whether it is a good idea to apply for a promotion. Hanan may think it would be good for Sally to apply, because Sally is well-qualified and hard-working, but nevertheless discourage Sally from applying because Hanan herself is going to apply and wishes to reduce the competition.

It is not necessary, however, for there to be misaligned incentives of this kind. Advice may be less informative than one's own opinion where the advisor is less able to perform the task, or does not perform the task as effectively. Consider, for instance, Sally asking Hanan for advice on where to go on holiday. Hanan may wish to maximise the probability Sally has a wonderful holiday by offering the best advice possible, but, because Hanan does not have a perfect knowledge of Sally's preferences, nevertheless advice Sally to select a non-optimal destination. Likewise, Sally may well have spent considerable time researching and thinking about the question, and it is not reasonable to believe that Hanan would do likewise because it is, after all, Sally's holiday.

There is room for misunderstanding even where an advisor's interests are aligned with the judge's and the advisor's ability equals the judge's. Advice must be communicated to the judge, and communication in the real world is inherently noisy. Communication of advice requires something in the mind of the advisor to be encoded into a set of signals, transmitted to the judge, and then reencoded into something in the mind of the judge, at which point it can be integrated with what the judge already believes. Information can be degraded at any of these steps, resulting in advice that is less informative than the judge's own opinion. When Sally asks Hanan what to do about a work problem, and Hanan rapidly and

confidently rattles off a suggestion, Sally may be forgiven for thinking the rapidity and confidence are a property of Hanan's confidence in the suggestion rather than an underlying characteristic of Hanan. If Sally does not adjust for the fact that Hanan is always more confident about things than Sally is, then Hanan's suggestion will be overly dominant relative to its informational value.

7.4.1 Compatability with existing explanations

As noted by Trouche et al. (2018), ultimate explanations of the kind offered here do not invalidate proximal explanations of the kind offered in the middle of this chapter (§Purported explanations for egocentric discounting). My view is most consistent with a Bayesian integration view in which advice is weighted by a range of features of advice, advisor, and context, with the further proviso that there are hyperpriors which govern the default level of discounting. I am thus content to observe the contest to provide proximal explanations for changes in the level of egocentric discounting, and only baulk at claims that egocentric discounting is 'irrational'. I suggest that, once egocentric discounting as a default is accepted, the adjustments in the level of discounting are almost all transparently rational.

7.4.2 Evidence

In the chapters that follow, I present evidence from computational agent-based evolutionary simulations and online human behavioural experiments to illustrate the plausibility of the claims made above. The evolutionary simulations demonstrate that an array of plausible factors affecting the relative utility of advice can create an environment in which egocentric discounting is adaptive, and the behavioural experiments demonstrate that some of these factors can be responded to by individual humans by adjustments in behaviour. I note that only the first of these is necessary for illustrating the plausibility of the theory: the behavioural adjustments serve more to support the extension of the argument, that changes in the level of discounting are rational adjustments.

8

Sensitivity of advice-taking to context

Advice-taking is often overly-conservative as compared to the normative level of advice-taking for a given experimental design. I argue that participants' performances in advice-taking experiments reflect both the specifics of the experimental design and prior expectations about advice-taking situations. These prior expectations may be both learned, as where individuals who grow up in less stable environments show lower propensity to trust `reference?`!TODO, and inherited. Most useful for the current argument would be a demonstration that conservatism can emerge within a population even where detrimental advice is rarely experienced, and that this can thus produce individuals who exhibit conservatism without ever experiencing detrimental advice. This demonstration is presented in the form of evolutionary modelling.

As discussed in the [previous chapter](#chapter-context), conservatism is optimal under some circumstances, and thus we expect that simulated agents allowed to evolve an advice-taking policy in those circumstances will evolve a conservative policy. I explored this tendency as a function of three plausible scenarios. The first scenario is one in which agents occasionally give deliberately poor advice to their advisee, which represents situations where advisors' interests may sometimes be contrary to judges' interests, unbeknownst to the judges. In the second scenario,

advice is simply noisier than the judge's own initial estimate, either because the judge is less competent at the task, less willing to exert the required effort for the task, or because the advice is communicated imperfectly. In the third scenario, agents belong to either a 'cautious' or a 'confident' group in how they express and interpret advice, which is a simple analogue of the observation that people's expressions of confidence are idiosyncratic (**aisIndividualConsistencyAccuracy2016**; **navajasIdiosyncraticNatureConfidence2017**). In each of these three scenarios, it is hypothesised that some level of egocentric discounting will emerge as the dominant strategy, i.e., the mean population weighting for initial estimates versus advice will be greater than .50. ## General method

Agent-based computational models of an evolutionary process were programmed in R (**rcoreteamLanguageEnvironmentStatistical2018**) and run variously on a home computer and the Oxford Advanced Research Computing cluster (**richardsUniversityOxford**
The code is available at `https://github.com/oxacclab/EvoEgoBias`, and the specific data presented below are archived at !TODO.

The models reported here use 1000 generations of 1000 agents which each make 30 decisions/generation on which they receive the advice of another agent. Decisions are either point estimation (Scenarios [1](#models-scenario-1) and 2) or categorical decision with confidence (Scenario 3). Each agent combines their own initial estimate with the advice of another agent, with the relative weights of the initial estimate and advice set by the agent's egocentric bias parameter, to produce a final decision. Final decisions are evaluated by comparison with the objective answer, and an agent's fitness is the sum of its performance over the 30 decisions of its lifetime.

8.0.1 Initial estimates

The agents perform a value estimation (category estimation in Scenario 3) task. Agent i's initial estimate t is the true value (v_t), plus some noise drawn from a normal distribution with mean 0 and standard deviation equal to the agent's insensitivity parameter (s^i, which is itself drawn from a positive-clamped normal distribution with mean and standard deviation 10 when the agent is created).

An agent's initial estimate (e_t^i) is thus:

$$e_t^i = v_t + N(0, s^i) \tag{8.1}$$

8.0.2 Advice

Each agent receives advice from another agent which it combines with its initial estimate to reach a final decision. The advice has a probability of being mutated in some fashion. The mutation depends upon the scenario and is described separately for each.

8.0.3 Final decisions

In the basic model from which other models inherit their decision procedure, agent i produces a final decision t as the average of the agent's initial estimate (e_t^i) and another agent's advice (a_t^i), weighted by the agent's egocentric bias (b^i). The models typically change the value of a_t^i, which is typically a function of some other agent j's initial estimate e_t^j.

An agent's final decision (d_t^i) is thus:

$$d_t^i = \frac{e_t^i \cdot b^i + a_t^i \cdot (1 - b^i)}{2} \tag{8.2}$$

The final decisions in Scenario 3 are more complex, but follow a similar structure.

8.0.4 Reproduction

Roulette wheel selection is used to bias reproduction in favour of agents performing best on the decisions. Performance is determined by a fitness function which differs slightly between categorical and continuous decisions. For scenarios 1 and 2, which use continuous decisions, this fitness is obtained by subtracting the absolute difference between the final decision and the true value for each decision:

$$u^i = -\sum_{t=1}^{30} |v_t - d_t^i| \tag{8.3}$$

The selection algorithm proceeds as follows: The worst performance is subtracted from each agent's fitness and 1 added to put fitness scores in a positive range. These scores are then continually multiplied by 10 until the lowest score is at least 10 to improve resolution. Each agent is then given a probability to reproduce equal to their share of the sum of all fitness scores:

$$r^i = \frac{u^i}{\sum_{j=1}^{n} u^j} \tag{8.4}$$

where n is the number of agents and u has undergone the transformations described above.

Reproducing agents pass on their egocentric bias to their offspring. Other agent features, e.g. decision-making accuracy, are randomised when they are created. In the present simulations, agents receive no feedback on decisions, and cannot learn about or discriminate between their advisors. The key outcome of interest in each simulation is whether the population evolves towards egocentric discounting as the dominant adaptive strategy.

8.1 Scenario 1: misleading advice

In scenario 1, agents sometimes choose to offer misleading advice to their advisee.

8.1.1 Method

The true value (v_t) is fixed at 50 in this scenario. The agents do not learn about the true value over time, so a fixed and arbitrary value does not alter the results of the simulation. Advice in this scenario is either the advising agent's initial estimate (e_t^j), or an extreme answer in the opposite direction to the advising agent's initial estimate (i.e. lower than 50 if e_t^j was above 50, and vice-versa). !TODO[describe what we actually do with this work; it has moved on a bit since transfer report]

8.1.2 Results

Where advice is always genuine, egocentric discounting does not emerge. Where advice is sometimes misleading, however, egocentric discounting emerges rapidly and remains stable throughout the rest of the stimulation.

8.1.3 Discussion

As one might expect, where there is the potential for exploiting trust it is safer to trust less, even if this reduces some of the benefits which would be gained from trusting. Egocentric discounting may be a viable strategy, even if there is no difference in ability between advice-seekers and advice-givers, given social contexts where the interests of advice-seekers and advice-givers are not perfectly aligned.

8.2 Scenario 2: noisy advice

In this scenario, agents offer advice which is of slightly lower quality on average than their own intial decisions. This could reflect a difference in effort or in expertise.

8.2.1 Method

As with Scenario 1, the true value (v_t) is fixed at 50.

Advice in this scenario has additional noise added:

$$a_t^i = e_t^j + N(0, x) \tag{8.5}$$

Where $x = 0$ in the no-noise condition and 5 in the noise condition.

Models were also run where a new decision was used as the basis for advice, rather than an initial estimate that some other agent would be using as a component in its own final decision:

$$a_t^i = v_t + N(0, s^j + x) \tag{8.6}$$

but the results were the same.

8.2.2 Results

As before, egocentric discounting emerges rapidly in the active condition and remains stable throughout the rest of the simulation.

8.2.3 Discussion

Where the advice is of worse quality on average, encoded here as additional variation, egocentric discounting tailors the relative weights of the estimates according to their average quality. As with situations in which an advice-seeker is more competent at making the relevant decision than their advisor, some measure of egocentric discounting is warranted in this scenario. It is worth noting that different competencies are not the only reason why advice may be systematically less valuable than initial decisions: difficulties in communicating the advice or different levels of conscientiousness in decision-making may also produce this effect.

8.3 Scenario 3: confidence confusion

While lackadaisical, incompetent, deliberately bad, or poorly communicated advice produces an obvious adaptive advantage for egocentric discounting, it is plausible that scenarios may exist where equally competent, wholly well-intentioned advice may still favour egocentric discounting. A common feature of advice is the communication of confidence, and this improves outcomes (Bahrami et al. 2010). Notably, there are large and consistent individual differences in people's expressions of confidence (**aisIndividualConsistencyAccuracy2016**; **songRelatingInterindividualDifferences20**
In this scenario agents are equally competent at the task, and do their best to assist one another, but may be hampered by expressing their estimates and advice with different confidences.

8.3.1 Method

The true value was drawn from a normal distribution around 50:

$$v_t = N(50, 1) \tag{8.7}$$

This allowed categorical answers which identified whether or not v_t was greater than 50. Agents were equiprobably assigned a personal confidence factor (c) of 1 or 10. This was used to scale the difference between the advising agent's intial estimate and the category boundary to produce the advice:

$$a_t^i = (e_t^j - 50 \cdot c^j) + 50 \tag{8.8}$$

Each agent then used the reciprocal of this process to translate advice back into its own confidence scale before integrating it with the initial estimate and arriving at a final decision:

$$d_t^i = \frac{e_t^i \cdot b^i + (\frac{1}{c^i}(a_t^i - 50) + 50) \cdot (1 - b^i)}{2} \tag{8.9}$$

This process amounts to the outgoing advice being translated into the advising agent's confidence language, and incoming advice being translated into the advised agent's confidence language. Where these languages are the same, the resulting advice is understood equivalently by both agents, but where there are differences the advice will be of greater or lesser magnitude compared to the initial estimate.

8.3.2 Results

Egocentric discounting again emerges in the condition where agents can have different confidence factors. The adaptiveness of egocentric discounting in this scenario arises because advice from a mismatched partner (e.g. one using a confidence scaling of 10 when the judge uses 1, or vice-versa) requires a different aggregation approach than advice from a matched partner. The application of an inappropriate aggregation approach results in misleading advice, so there is a pressure for a middle-of-the-road strategy whereby advice is counted, but not too much.

8.3.3 Discussion

Even where there is no intent to mislead, and no difference in basic ability, it is possible for egocentric discounting to emerge as an optimal strategy, purely because of differences in how agents communicate and understand estimates.

8.4 General discussion

The computational models show that egocentric discounting is an adaptive strategy in an array of plausible advice contexts: misleading advice, noisy advice, and different interpretations of confidence. While these models are necessarily limited in applicability to real life, they do demonstrate that egocentric discounting, while irrational for simple estimation problems with an objective answer and with advice that is not systematically better or worse than an individual's own judgement, may be beneficial for many of the kinds of decision for which we have sought and used advice in everyday life throughout our evolution. This argument invites attention to the advice-taking task as much as to the properties of the advisor: it predicts that egocentric discounting will be attenuated where the outcome of decisions affects judges as well as advisors (Gino 2008, observed this effect but attributed it to judges falling prey to the sunk costs fallacy); where decisions rely more on objective than on subjective criteria (Van Swol 2011); where advisors and judges have opportunities to calibrate their confidence judgements with one another by completing training trials where they have to produce a shared decision with a shared confidence; and where incentives for judges and advisors are more closely aligned (Gino, Brooks, et al. 2012; Sniezek, Schrah, et al. 2004).

Notably, the utility of these heuristics does not depend on malice, mistake, or miscommunication: inconsistency in the usage of confidence terminology can produce adaptive pressure for egocentric discounting. More generally, the results indicate that properties of the advice-giving milieu can influence advice-taking strategies.

These models establish that it is plausible that people have deeply ingrained hyperpriors towards discounting advice. It is also possible, however, that people

can flexibly respond to contexts, modulating their advice-taking appropriately. The voluminous experiments in advice taking discussed in the introduction to this section !TODO[that discussion?] can be seen as eliciting exactly this behaviour. In the next chapter I present new behavioural experiments which suggest that people do indeed modulate their advice-taking to the specifics of the context in which they are in.

9

Behavioural responses to advice contexts

9.1 Benevolence of the advisor population

!TODO[Make sure to report in brief/as post mortem the results for the in/out group studies and the early direct benevolence experiments, and point readers to the full details.]

The evolutionary models discussed in the last chapter demonstrated that optimal advice-taking strategies depend in part upon the advice one receives being a genuine effort to help. Difference in benevolence, or the extent to which the interests of the advisor and the judge overlap, have been shown to affect levels of advice-taking. !TODO[CITE] showed that, where advisors were paid contingent upon the quality of the final decision, advice-taking was higher than where advisors were paid a flat fee for providing advice. !TODO[lit review stuff - inc. Behrens/Hunt? social decision making]

Advice-taking can be contingent on the properties of the advice, or on the properties of the advisor. In order to maximise the value of advice while minimising the potential exposure to exploitation, advice-taking should be contingent on a combination of these factors. Where advice is plausible it should be weighted relatively equally, whether it comes from an advisor who is sometimes misleading or not, but where advice is more implausible it should only be trusted when

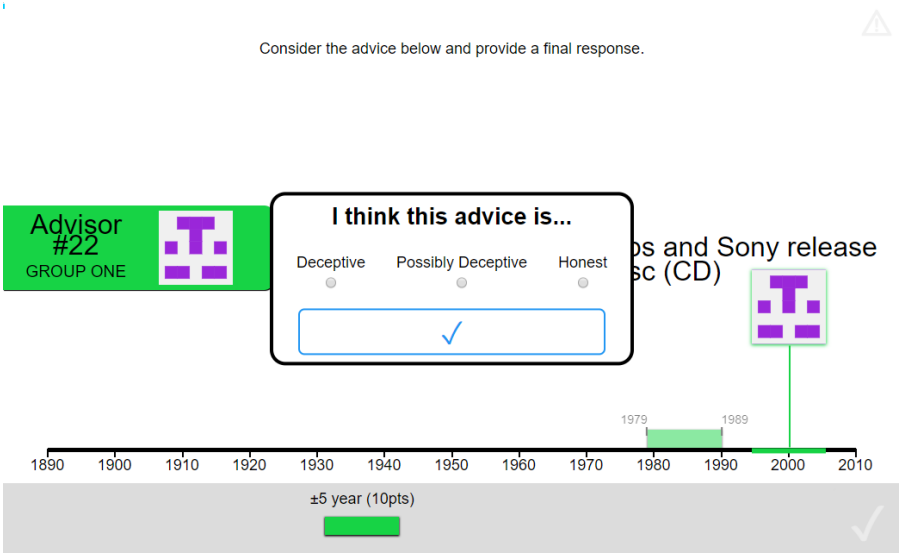


Figure 9.1: Advice honesty rating.
Participants rated advice on a three-point scale according to whether they thought the advice was deceptive or honest.

it comes from an advisor who is never misleading. To explore whether people's behaviour matches this pattern, participants were recruited for a series of behavioural experiments in which they were given advice on a date estimation task from advisors who were described as either always helpful or occasionally misleading. Results indicated that participants' advice-taking depended upon both the plausibility of the advice and the benevolence of the advisor, although the specific pattern differed from the one described above.

9.1.1 Method

Procedure

The procedure for these experiments follows the dates estimation task, described previously. Additionally, in this version, participants had to provide an assessment of the honesty of each advisory estimate they received before they could integrate it into their final decision.

Manipulation

Participants performed two blocks of experimental trials. In one block, participants received advice from an advisor who they were told would 'always try to help', while in the other they received advice from an advisor who 'may sometimes try to mislead'. The order of advisors was counterbalanced between participants.

The advice offered by the advisors was drawn from equivalent distributions for both advisors, meaning that there were no systematic differences in the quality of the advice. Both advisors offered advice sampled from a roughly normal distribution around the correct answer with a standard deviation of 11 years.

Experiments

There were a series of experiments in this topic during which the experimental design was tweaked in order to find an effective manipulation and operationalisation of the key concepts. The experimental code for previous versions can be found on the commit history on GitHub for the main experimental file, as well as its dependencies. To give a brief overview of the major changes:

- Version 2.0.0
  - introduced a much clearer manipulation
    - reminded participants of advisor description in a message they had to acknowledge
    - kept participant's group visible throughout
    - included a single trial where the misleading advisor did actually mislead the participant
- Version 2.1.0
  - removed actual differences in advice and adjusted advisor descriptions to match

- Version 3.0.0
  - added a question probing the perceived honesty of the advice between receiving advice and providing a final decision
- Version 3.0.1
  - pre-registered replication of version 3.0.0
  - Deviations from pre-registration:
    * Added a new exclusion rule for those people who use translation software.
- Added exclusions for participants with NA values in the in- vs out-group t-test (where e.g. no outgroup advice was rated as 'honest').

```
* Included frequentist stats in the trustworthiness questionnaire item t-tes


* Fixed some labels on graphs


* Exploratory analyses expanded to include analysis of trials with WoA > .05
```

Below, only the results from **version 3.0.1** are reported, but the results of this version are compatible with findings of previous versions. Data for all versions are accessible !TODO[archive data].

9.1.2 Results

!TODO[Restructure this code to use the esmData package. Should be easy...]

!TODO[consistent advisor naming, including in figures]

Exclusions

Participants (total $n = 20$) could be excluded for a number of reasons: failing attention checks, having fewer than 11 trials which took less than 1 minute to complete (one participant), providing final decisions which were the same as the initial estimate on more than 90% of trials, or using non-English labels for the honesty questionnaire (one participant). The latter exclusion was added after data were collected because it was not anticipated that participants would use translation software in the task.

The final participant list consists of 18 participants who completed an average of 11.83 trials each.

Task performance

Participants performed as expected, decreasing the error between the midpoint of their answer and the true answer from the initial estimate to the final decision, which suggests that they incorporated the advice, which was indicative of the correct answer.

```
## `summarise()` regrouping output by 'pid' (override with `.groups` argument)
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

The advice participants received was subject to variation, but there did not appear to be overall systematic differences between the advisors either in terms of error (distance between the centre of the advice estimate and the correct answer) or distance (distance between the centre of the advice estimate and the centre of the initial estimate). !TODO[error and distance graphs]

Hypotheses

The hypotheses tested were that * weight on advice would be higher for advice rated as 'honest' versus advice rated as 'deceptive' * weight on advice would be higher for advisors who were described as 'never misleading', even for advice rated as 'honest'

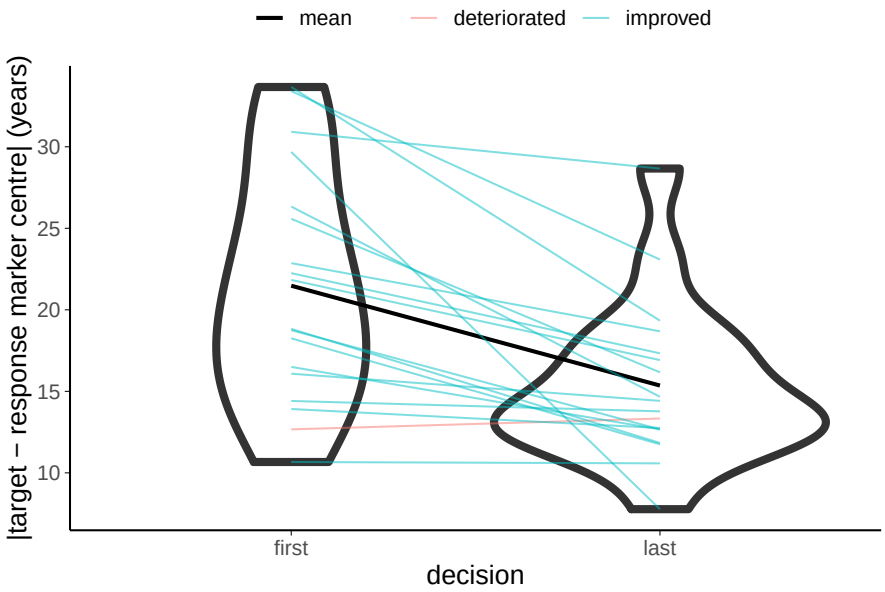


Figure 9.2: Mean answer error for initial estimates and final decisions.
Faint lines show means for individual participants collapsed across trials, while boxplots, violins, and the heavy dashed line show aggregate participant means.

Effect of advice

```
## `summarise()` regrouping output by 'pid' (override with `.groups` argument)
```

As expected, participants were substantially more influenced by advice they rated as 'honest' compared to advice they rated as 'deceptive' (!TODO[fix ttest]).

Effect of advisor

```
## `summarise()` regrouping output by 'pid' (override with `.groups` argument)
```

Participants were also more influenced by 'honest' advice from the advisor who was described as 'never misleading' (!TODO[fix ttest]).

!TODO[plot of woa by advice rating and advisor]

Results plot The details of the data are best presented in a combined figure.

The figure shows that there is a different pattern in the responses to advice across advice ratings depending upon the advisor supplying the advice. Advice perceived as not being honest is discounted relatively equally from the advisor described as 'always helpful', and is still fairly influential in the final decision. The advice perceived as 'honest' from this advisor is the most influential of all

advice. Likewise, the 'honest' advice from the advisor described as 'sometimes-misleading' is more influential than the 'possibly deceptive' advice, which is in turn more influential than the 'deceptive' advice. The 'deceptive' advice from the 'sometimes-misleading' advisor is almost completely ignored. Within this broader pattern there is some variability between participants, including some participants whose responses violate the pattern quite dramatically - note the two participants with large decreases in influence for 'honest' advice from the 'sometimes-misleading' advisor as compared to the 'possibly deceptive' advice from the same. This may be a consequence of participants' values in each cell in the design being comprised of a low number of trials. #### Exploration

!TODO[below]

- were participants' criteria for judging advice as dishonest different between advisors?
- Type 1 SDT might be able to do this if a sensible definition of objectively-deceptive advice can be produced

  * E.g.

  whether averaging with the advice would reduce error

  * There will be all sorts of problems with this, noteably it'll classify 'ne

  – May be able to investigate this with probabilities of classification for several distance bins - ROC type curve

    * what to do with 'possibly deceptive'?

9.1.3 Discussion

- Both the source and the plausibility of advice matter, and they seem to interact in a complex way.

- Flexible adaptation to advice context, both in terms of categorisation and influence.
- Support for the premise of the evolutionary simulations.
  - Does not prove people actually do have these hyperpriors

Limitations

- N
- No gender breakdown
  - simplicity/privacy vs data gap
- Difficult task (high levels of advice-taking by default)

9.2 Noise in the advice

The second scenario explored in the evolutionary models added noise to the advice agents received and demonstrated that this provided an evolutionary pressure towards egocentric discounting. This scenario is not explored in behavioural experiments because its conclusions are well supported by existing literature. Specifically, the literature demonstrates that the normativity of discounting where the judge outperforms the advisor, that advice-taking is sensitive to advisor expertise, and that people are likely to consider themselves superior to the average advisor.

The addition of noise in a point-value estimation task lowers the relative performance, and thus this scenario was essentially a manipulation of advisor expertise. Normative models support egocentric discounting where the judge is systematically more accurate than the advisor !TODO[PAR model paper, early advice-taking theory papers]. The decrease in advice-taking for novice as opposed to expert advisors is a robust result in the literature: !TODO[expertise manipulation experiment lit review]. The evidence above makes an empirical investigation on this point somewhat redundant, provided it can be shown that people have a naive

assumption of superiority on a given task compared to the average other they are likely to encounter. This point is also well supported in the literature on self-serving biases !TODO[describe driving self-assessment, naive realism?, etc.]. It is expected that this is somewhat dependent upon the difficulty of the task presented; people faced with a difficult task will under- rather than overestimate their ability relative to others !TODO[hard-easy effect of advice]. Taken together, the above arguments demonstrate that, on average, people are likely to consider themselves more able on a given task than an arbitrary advisor, and consequently that they are likely to downweight advice relative to their own initial estimate. This behaviour is supported by normative mathematical models which show biasing towards the better estimator (in this case the judge) is the optimal strategy. As discussed in the scenario, the belief that one is better at a task than the average advisor may not be misguided: advisors may not dedicate the same amount of time, concentration, or thought to producing advice as judges do for initial estimates. Judges have to live with the consequences of their decisions, while advisors do not. ## Confidence mapping

The third scenario explored in the evolutionary models assigned each agent a confidence mapping, and demonstrated that discounting emerged as an appropriate response where the advisor's confidence mapping was unknown. The key difficulty in conducting behavioural experiments to test the effects of known vs unknown confidence mapping is finding a manipulation of confidence mapping knowledge which is not confounded by familiarity with an advisor or the amount of information provided by an advisor. We train people in two contexts, in both of which they receive advice with confidence from one of two advisors and see feedback on their final answers. In the first context, all the advice is well calibrated, and both advisors use either high or low confidence distributions. In the second context, one advisor uses a high and the other a low confidence distribution. Participants are then tested with a brief exposure to an advisor whose confidence distribution is drawn from the middle of the scale. Where participants are unprepared for different confidence mappings (the first context), they will treat the new advisor like the old advisors - considering the advice as high- or low-confidence (and weighting it accordingly)

depending on whether it is high or low for the advisors' distributions. Where participants are prepared for different confidence mappings (the second context), they will treat the new advice as medium confidence, and weight it accordingly. There will be a difference in the weight on advice in the probe trials between contexts: WoA will be highest where both training advisors have a low-confidence distribution, middling where they have different distributions, and lowest where they both have a high-confidence distribution.

9.2.1 Individuality as a cue to confidence

Method

Results

9.2.2 Identifiability of advice

Method

Results

9.3 General discussion

The results of the direct benevolence experiments show that people are sensitive to the motivations of their advisors, and exercise appropriate caution where those motivations may mean the advice is misleading. They also demonstrate that, even where advice is considered trustworthy, it is discounted when coming from a less trusted advisor. Together with existing literature on the effects of advisor expertise, this supports the idea from the models that advice-taking can be fleixbly adjusted according to the context. The results of the confidence mapping experiments were less conclusive. We were unable to produce a manipulation which was sufficiently clear and strong to produce observable effects. It is plausible that people are sensitive to their knowledge of an advisor's confidence mapping, but it is also plausible that this degree of flexibility is beyond most people, and that simple heuristics (e.g. relying on cultural norms about the meanings of metacognitive terms !TODO[cite Bang, Heyes, et al. on cultural metacognition]) are used instead of complex calculations. The lack of flexibility on the timescale of a behavioural

experiment does not, of course, negate the possibility of flexibility on the timescale of a human interpersonal relationship (perhaps there are a certain number of people whose confidence mappings we can track, a la Dunbar's number). Nor does it negate the possibility of an evolutionary or cultural-evolutionary mechanism baking in advice discounting as a protection against unknown confidence mapping, although it is questionable whether advice has been occuring in human societies for long enough to allow the former to take place. We have a plausbile explanation of egocentric which only relies on rational responses to environmental effects. These effects are ubiquitous, and thus a ubiquitous egocentric bias prior to engaging with the specifics of a situation makes sense. It is entirely plausible that experiments showing egocentric discounting behaviour fail to overcome the hyperpriors held by participants that taking advice is risky for various reasons.

We saw in the previous section that people are sensitive to the quality of advice and advisors, and will curate their information environments based on that sensitivity. We have seen in this section how the properties of the information environment can change people's sensitivity to advice. !TODO[If we have time, it'd be neat to see if we can get people's selection rates for advisors to recover after the advisors have been lying to them, showing that that how we curate information environment (source selection) also depends on the existing information environment properties.]

!TODO[This theory really doesn't explain combining multiple estimates with one's own. How important is that failure? Do other theories explain that feature?]

Part IV

Interaction

10

Interaction of psychological processes across minds

!TODO[extend the simulation/agreement-effects-analysis stuff a bit and it will fill this chapter out well.]

- Even when biases don't change (Niccolo covered that), source selection on the basis of agreement is inherently pathological (unless judgements are uncorrelated).
  - This is true whether or not confidence weights updating.
- Why should we care if it's not changing biases (i.e. opinions?)
  - because it reinforces final decisions through agreement instead of attenuating them through disagreement.
- Our psychological mechanisms from Section II thus produce bad network effects when feedback is absent.
- Can we relate the networks back to the advice-giving environments from Section III?

Models say this will go horribly wrong.

11

Network effects of interaction

My models also suggest horrible things.

12
Real-world network effects

My models may not be good models.

Part V

Conclusion

13
Conclusion

I haven't wasted 3 years of my life and Nick's time.

- I've demonstrated that the structure of information networks may be partially dependent upon psychological mechanisms of advisor evaluation.

- I've provided some evidence for a model of advisor evaluation in the absence of feedback and demonstrated how that model produces adverse network effects.

- I've shown how the information environment can affect advice taking, and I've provided and reviewed empirical evidence that humans can alter advice-taking flexibly depending on these contextual factors.

- I've modelled how the psychological mechanisms create an information network structure, and how that information network structure might use the flexibility to context to change advisor evaluation processes.

- I've demonstrated how those changes might attenuate or exacerbate the formation of pathological network structures.

13.1 Open questions

- People's social networks constrain whom they can receive advice from. Within these constraints, source selection processes operate to evaluate the relative (and absolute?) trustworthiness of each (or the most salient?) members. Is trustworthiness (or at least the ability component) modelled discretely for different domains, or is it unitary?
  - Perhaps a matter of association with a particular concept?
  - Quite possible associative processes drive source selection rather than direct comparisons.
- Relative magnitude of advice from selected individuals in a social network vs. adverts/web searches/consulting experts.
- Frequency of advice-seeking as characterised in the thesis.
- (Why) do people lump group advice together before discounting - this seems idiotic even with the adapativeness of discounting...

!TODO[Chronological account of advisor choice experiments. Perhaps in that chapter, perhaps appendix. Be nice + transparent to present stuff in best scientific order in thesis but have a record of the actual narrative.]

Appendices

A

The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

And here's another one from the same chapter, i.e. Chapter ??:

B

The Second Appendix, for Fun

Works Cited

Bahrami, Bahador et al. (2010). "Optimally Interacting Minds". In: *Science* 329.5995, pp. 1081–1085.

Bonner, Bryan L. and Brian D. Cadman (Dec. 2014). "Group Judgment and Advice-Taking: The Social Context Underlying CEO Compensation Decisions". English. In: *Group Dynamics-Theory Research and Practice* 18.4, pp. 302–317.

Ernst, Marc O. and Martin S. Banks (Jan. 2002). "Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion". en. In: *Nature* 415.6870, pp. 429–433.

Fetsch, Christopher R. et al. (2012). "Neural Correlates of Reliability-Based Cue Weighting during Multisensory Integration". en. In: *Nature Neuroscience* 15.1, pp. 146–154.

Gino, Francesca (Nov. 2008). "Do We Listen to Advice Just Because We Paid for It? The Impact of Advice Cost on Its Use". en. In: *Organizational Behavior and Human Decision Processes* 107.2, pp. 234–245.

Gino, Francesca, Alison Wood Brooks, and Maurice E Schweitzer (2012). "Anxiety, Advice, and the Ability to Discem: Feeling Anxious Motivates Individuals to Seek and Use Advice". en. In: *Journal of Personality and Social Psychology* 102.3, pp. 497–512.

Gino, Francesca and Don A. Moore (Jan. 2007). "Effects of Task Difficulty on Use of Advice". en. In: *Journal of Behavioral Decision Making* 20.1, pp. 21–35.

Harvey, Nigel and Ilan Fischer (May 1997). "Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility". en. In: *Organizational Behavior and Human Decision Processes* 70.2, pp. 117–133.

Harvey, Nigel and Clare Harries (July 2004). "Effects of Judges' Forecasting on Their Later Combination of Forecasts for the Same Outcomes". en. In: *International Journal of Forecasting* 20.3, pp. 391–409.

Hütter, Mandy and Fabian Ache (2016). "Seeking Advice: A Sampling Approach to Advice Taking". en. In: *Judgment and Decision Making* 11.4, p. 16.

Jacowitz, Karen E. and Daniel Kahneman (Nov. 1995). "Measures of Anchoring in Estimation Tasks:" en. In: *Personality and Social Psychology Bulletin*.

Körding, Konrad P. et al. (Sept. 2007). "Causal Inference in Multisensory Perception". en. In: *PLOS ONE* 2.9, e943.

Liberman, Varda et al. (Mar. 2012). "Naïve Realism and Capturing the "Wisdom of Dyads"". en. In: *Journal of Experimental Social Psychology* 48.2, pp. 507–512.

Mahmoodi, Ali et al. (Mar. 2015). "Equality Bias Impairs Collective Decision-Making across Cultures". en. In: *Proceedings of the National Academy of Sciences* 112.12, pp. 3835–3840.

Minson, Julia A., Varda Liberman, and Lee Ross (Oct. 2011). "Two to Tango: Effects of Collaboration and Disagreement on Dyadic Judgment". en. In: *Personality and Social Psychology Bulletin* 37.10, pp. 1325–1338.

Moussaïd, Mehdi et al. (Nov. 2013). "Social Influence and the Collective Dynamics of Opinion Formation". en. In: *PLOS ONE* 8.11, e78433.

Önkal, Dilek, Sinan M. Gönül, et al. (Jan. 2017). "Evaluating Expert Advice in Forecasting: Users' Reactions to Presumed vs. Experienced Credibility". en. In: *International Journal of Forecasting* 33.1, pp. 280–297.

Önkal, Dilek, Paul Goodwin, et al. (Oct. 2009). "The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments". English. In: *Journal of Behavioral Decision Making* 22.4, pp. 390–409.

Rakoczy, Hannes et al. (Oct. 2015). "Young Children Heed Advice Selectively". en. In: *Journal of Experimental Child Psychology* 138, pp. 71–87.

Rollwage, Max et al. (May 2020). "Confidence Drives a Neural Confirmation Bias". en. In: *Nature Communications* 11.1, p. 2634.

Ronayne, David and Daniel Sgroi (2018). "Ignoring Good Advice". In:

Schultze, Thomas, Andreas Mojzisch, and Stefan Schulz-Hardt (May 2017). "On the Inability to Ignore Useless Advice: A Case for Anchoring in the Judge-Advisor-System". en. In: *Experimental Psychology* 64.3, pp. 170–183.

Schultze, Thomas, Anne-Fernandine Rakotoarisoa, and Stefan Schulz-Hardt (2015). "Effects of Distance between Initial Estimates and Advice on Advice Utilization". en. In: *Judgment and Decision Making* 10.2, p. 28.

See, Kelly E. et al. (Nov. 2011). "The Detrimental Effects of Power on Confidence, Advice Taking, and Accuracy". In: *Organizational Behavior and Human Decision Processes* 116.2, pp. 272–285.

Sniezek, Janet A., Gunnar E. Schrah, and Reeshad S. Dalal (July 2004). "Improving Judgement with Prepaid Expert Advice". en. In: *Journal of Behavioral Decision Making* 17.3, pp. 173–190.

Sniezek, Janet A. and Lyn M. Van Swol (Mar. 2001). "Trust, Confidence, and Expertise in a Judge-Advisor System". In: *Organizational Behavior and Human Decision Processes* 84.2, pp. 288–307.

Soll, Jack B. and Richard P. Larrick (2009). "Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions." en. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.3, pp. 780–805.

Soll, Jack B. and Albert E. Mannes (Jan. 2011). "Judgmental Aggregation Strategies Depend on Whether the Self Is Involved". en. In: *International Journal of Forecasting* 27.1, pp. 81–102.

Svenson, Ola (Feb. 1981). "Are We All Less Risky and More Skillful than Our Fellow Drivers?" en. In: *Acta Psychologica* 47.2, pp. 143–148.

Swol, Lyn M. and Janet A. Sniezek (Sept. 2005). "Factors Affecting the Acceptance of Expert Advice". en. In: *British Journal of Social Psychology* 44.3, pp. 443–461.

Tost, Leigh Plunkett, Francesca Gino, and Richard P. Larrick (Jan. 2012). "Power, Competitiveness, and Advice Taking: Why the Powerful Don't Listen". English. In: *Organizational Behavior and Human Decision Processes* 117.1, pp. 53–65.

Trouche, Emmanuel et al. (Jan. 2018). "Vigilant Conservatism in Evaluating Communicated Information". en. In: *PLOS ONE* 13.1. Ed. by Alexander N. Sokolov, e0188825.

Van Swol, Lyn M. (Jan. 2011). "Forecasting Another's Enjoyment versus Giving the Right Answer: Trust, Shared Values, Task Effects, and Confidence in Improving the Acceptance of Advice". en. In: *International Journal of Forecasting* 27.1, pp. 103–120.

Wang, Xiuxin and Xiufang Du (Nov. 2018). "Why Does Advice Discounting Occur? The Combined Roles of Confidence and Trust". English. In: *Frontiers in Psychology* 9, p. 2381.

Yaniv, I. (Jan. 2004). "Receiving Other People's Advice: Influence and Benefit". English. In: *Organizational Behavior and Human Decision Processes* 93.1, pp. 1–13.

Yaniv, Ilan and Shoham Choshen-Hillel (Dec. 2012). "Exploiting the Wisdom of Others to Make Better Decisions: Suspending Judgment Reduces Egocentrism and Increases Accuracy". en. In: *Journal of Behavioral Decision Making* 25.5, pp. 427–434.

Yaniv, Ilan and Eli Kleinberger (Nov. 2000). "Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation". en. In: *Organizational Behavior and Human Decision Processes* 83.2, pp. 260–281.

Yaniv, Ilan and Maxim Milyavsky (May 2007). "Using Advice from Multiple Sources to Revise and Improve Judgments". English. In: *Organizational Behavior and Human Decision Processes* 103.1, pp. 104–120.