Constructive biases in clinical judgment

Bartosz W. Wojciechowski[1], Bernadetta Izydorczyk[1], Pawel Blasiak[2], James M. Yearsley[3], Lee C. White[3], & Emmanuel M. Pothos[3]

1. Faculty of Management and Social Communication, Institute of Applied Psychology, Jagiellonian University, Kraków, Poland.

2. Institute of Nuclear Physics, Polish Academy of Sciences, Kraków, Poland.

3. Department of Psychology, City, University of London, London, UK.

*Correspondence: Bartosz W. Wojciechowski, b.wojciechowski@uj.edu.pl

Running text word count, including abstract and acknowledgments, but not references and Online Supplementary Material: 7187 words

Abstract

With a pair of oppositely valenced stimuli, rating the first one sometimes leads to a more extreme evaluation for the second (e.g., if the second is negatively valenced, rating the first stimulus would lead to a more negative rating for the second; White et al., 2014). We considered an evaluation bias in the case of clinical diagnosis relating to eating disorders. A population sample which included experienced clinical psychologists and psychiatrists, showed partial evidence of an evaluation bias, when judging descriptions of individuals designed to be consistent with eating disorders or not. Quantum probability theory, the probability rules from quantum mechanics without any of the physics, is particularly well-suited to modeling the evaluation bias (and constructive influences generally), because a measurement (or judgement) can change the state of the system. We applied a previous quantum model to the present result, an extension of the model embodying noisy processes, and Hogarth and Einhorn's (1992) belief adjustment model. We discuss how model fits inform an examination of rationality in the observed behavior.

key words: evaluation bias, constructive influences, quantum probability theory, clinical decision making, eating disorders

1. Introduction

1.1 How does the process of judgment affect the outcome?

Consider a clinical psychologist evaluating different pieces of information about a patient. After each piece of evidence, the clinician might offer a judgment. It would be unsurprising to observe that these judgments develop in a way consistent with the presented information. What would be surprising is if the process of making a judgment affects the clinician's perception that the patient is ill or not. From a folk perspective, surely the rational expectation is that the clinician determines the patient to be ill or not on the basis of the available evidence, not on whether any intermediate judgments are made or not.

In fact, it has been well-known for a while that intermediate judgments can impact on an eventual conclusion. Hogarth and Einhorn (1992) systematized a large body of research showing that, when evaluating a sequence of pieces of evidence, step-by-step (SbS; an intermediate judgment is made after each piece of evidence) vs. end-of-sequence (EoS; a single judgment is made after all pieces of evidence) modes of evaluation sometimes result in different overall conclusions. Hogarth and Einhorn (1992) considered long series of pieces of evidence (the 'short' series corresponded to sequences between 2 and 16) and it is possible that the impact of intermediate judgments in their analysis simply relates to e.g. memory reinforcement effects. However, there is evidence that just a single judgment can alter corresponding beliefs (Ariely & Norton, 2008; Brehm, 1956; Sharot, Velasquez, & Dolan, 2010). Such findings can be called 'constructive influences', because the judgment appears to alter (construct) the relevant mental state.

We focus on the evaluation bias (White, Pothos, and Busemeyer, 2014). In a typical experiment, participants would be presented with pairs of stimuli having opposite affective valence, positive (P) or negative (N). The second stimulus in each pair would always be rated.

The first stimulus would be sometimes be rated, sometimes offered just for observation. White et al. (2014) found that rating the first stimulus would lead to a more extreme rating for the second stimulus (i.e., if the second stimulus was N, it would be rated as more negative etc.). Thus, for the same two stimuli presented in the same order, judging the first stimulus would alter participants' perception for the second one in a systematic way. However, White et al.'s (2014) work, as well as most related research, concerns low stakes decision situations, with college undergraduate participants (an exception is White et al., 2020). It is possible that judgment biases, such as the evaluation bias, disappear when the judgment situation sufficiently engages thoughtful, considerate decision processes (Kahneman, 2001).

1.2 Rationality and biases in clinical diagnosis

We expect clinical diagnosis to be as rational as possible, and this includes expectations concerning the extent of information about the patient and expertise of the clinician. However, there is a more fundamental requirement for rationality, if one recognises that clinical diagnosis is probabilistic inference on uncertain premises. Probabilistic inference is based on rules for how to combine different pieces of information and update the probability of a target conclusion (e.g., whether a patient is ill or not). The established rational standard for doing so is Bayesian probability theory (Oaksford & Chater, 2009; Tenenbaum et al., 2011). Using Bayesian principles, a reasoner benefits from powerful convergence and consistency/ coherence arguments.

Bayesian theory is not the only probability theory. Quantum theory is the probability rules from quantum mechanics, without any of the physics. It has been applied in behavioral modeling, when baseline Bayesian models appear problematic (Busemeyer & Bruza, 2012; Pothos & Busemeyer, 2013; Wang et al., 2014). One of the key rational justifications for Bayesian theory applies to quantum theory as well (Pothos et al., 2017). But, if both Bayesian

theory and quantum theory can benefit from the same rational justification, and sometimes diverge in their prediction, which theory do we choose to judge some behavior as rational vs. not? The answer to this question depends on which probability theory better describes the situation at hand.

We introduce the distinction between compatible and incompatible questions. Compatible questions are ones which can be resolved concurrently. For incompatible question, this is not possible, e.g., because the meaning of one question might alter the meaning of the other or because just asking one question might change or disturb the system of interest. An example is the pair of questions ‘how is your exams revising going’ and ‘what will you do Friday night’ for a teenager. Regardless of the answer, just asking the first question might make the teenager anxious, thus altering response probabilities for the second question. That is, the two questions are incompatible because in this case asking one question changes/ disturbs the relevant system (the teenager).

Bayesian theory is rational for compatible questions and quantum theory for incompatible questions. Can we simply consider the consistency of any judgment with Bayesian principles (for compatible questions) or quantum principles (for incompatible questions) and so make a determination regarding its rational status? There are three complications. First, a technical issue is that quantum theory works for compatible questions too, but in that case it mostly reproduces Bayesian theory; and Bayesian theory can accommodate some features of incompatibility (e.g., with conditionalizing), but quantum theory is better suited to inference with incompatible questions. Second, Bayesian and quantum theories are complex mathematical frameworks and it is difficult to disprove them (cf. Jones & Love, 2011). Quantum theory, Bayesian theory and any general framework for computational modeling are general enough to be flexed to cover any results. But, of course, not all models would be convincing to the scientific community for other reasons, for

example, concerning the degree to which model assumptions are reasonable and well-motivated. Third, there have been proposals for how to reconcile Bayesian theory with apparent inconsistencies, if, for example, resource limitations (Lieder & Griffiths, 2019) or conversational implicatures (Goodman & Frank, 2016) are taken into account.

Notwithstanding these issues, we can examine whether 'reasonable' Bayesian vs. quantum models might be more appropriate in an experimental situation and whether human behavior follows the relevant principles (cf. Shepard, 1992). Our focus is the presence or not of an evaluation bias in clinical diagnosis of eating disorders, with participants including highly experienced clinical psychologists and psychiatrists.

Clinical diagnosis is a demanding task. Several studies indicate that the use of diagnosis and treatment manuals does not enhance diagnostic accuracy (Huppert et al., 2001; Norcross, 2002), but also that few professionals actually follow the available recommendations and guidelines (Currin et al., 2007; Lilienfeld et al., 2013). The relation between experience and diagnostic validity has also been investigated. Perhaps surprisingly, there is evidence that professionals are sometimes just as accurate as graduate students (Hermann et al., 1999; Muller & Davids, 1999). Studies on illusory correlations, which occur when a person believes that events are correlated, even though they really are not, further suggest it can be difficult for clinicians to learn from clinical experience (Garb, 1998).

It is well-known that clinical judgment is not immune to apparent cognitive biases (Garb, 2003). For example, research on covariation misestimation suggests that mental health professionals are more likely to remember instances in which a test indicator and symptom are present than those in which a test indicator is absent and a symptom is either present or absent (Kayne & Alloy, 1988). Another example concerns reports that the act of making a diagnosis can influence how a mental health professional remembers a client's symptoms (Arkes & Harkness, 1980). That is, mental health professionals may forget that a client has

particular symptom because the symptom is not typical of those associated with the client's diagnosis.

Evidence for apparent biases in clinical judgment contrasts with recommendations for best practice. Currently there is an emphasis on using evidence-based practice in psychology (EBPP; Cierpiałkowska & Sęk, 2016; APA, 2006). The purpose of EBPP is to promote effective psychological practice by applying empirically supported principles of psychological assessment, case formulation, therapeutic relationship, and intervention (Bauer, 2007). Some of the components of EBPP are (a) assessment and diagnostic judgment; (b) clinical decision-making; (c) appropriate evaluation and use of research evidence. At the diagnostic stage, EBPP focuses on searching for answers that would enable correct diagnosis and estimating the probability that a patient has a particular disorder. The EBPP recommends scientific methods and independently verifiable facts as much as possible, as opposed to e.g. personal intuitions, but it falls short of providing tools for clinical inference using Bayesian theory (Straus et al., 2011; Youngstrom, 2013).

The present focus on eating disorders concerns our (BW, BI) experience in the area, the availability of case studies, and access to professional populations. Eating disorders can be serious and debilitating mental illnesses, characterized by preoccupation with one's body weight and shape. The incidence of eating disorders is on the increase, especially anorexia and bulimia nervosa for younger women in Europe (Garner, 2004), including in Poland, which is where the present study was carried out (Izydorczyk, 2011a, 2013). The differentiation between the various types of eating disorders (e.g., bulimia, anorexia nervosa, or compulsive overeating/binge eating) is performed using criteria corresponding to medical classifications of diseases and behavioral disorders (Cierpiałkowska & Sęk, 2016).

1.3 An evaluation bias in the diagnosis of eating disorders

The finding of Arkes and Harkness (1980), that making a diagnosis can influence the memory for particular symptoms, shows that making a judgment vs. not can impact on subsequent judgments, even when the presented information is identical. We can understand such results as constructive influences, when a judgment can change the underlying mental state (Gloekner et al., 2009; Hogarth & Einhorn, 1992; Sharot et al., 2010).

Baseline Bayesian theory is fairly uninformative concerning constructive influences. We can write $Prob(Disorder, second\ judgment) \neq Prob(Disorder, second\ judgment|first\ judgment)$, but how does knowledge of the first judgment impact on the second one? It could reduce, increase, or not affect the probability of the second judgment. An analogous argument can be offered when considering whether Bayesian theory provides an accurate description of situations when judgments or measurements alter the system (as for the example of the revising, anxious teenager). In quantum theory, constructive influences arise fairly naturally. To explain how this occurs, we briefly introduce quantum theory.

The starting point in a quantum model is a multidimensional space, where all possible questions are represented as subspaces. For example, in Figure 1, we have a two-dimensional overall space corresponding to all the questions a clinician could be asking about a patient. In that space we represent one question: does the patient suffer from an eating disorder vs. not. The mental state of the clinician is represented by a normalized (i.e., length 1) vector, ψ. Probability for a question outcome is the squared length of the 'projection' of the mental state vector, to the subspace corresponding to the question outcome. For example, to compute the probability that the clinician considers the patient having the disorder, we project (lay down) the vector ψ onto the disorder subspace and compute the square of the length of this projection. So, higher overlap between the mental state vector and different subspaces indicates higher probability.

What happens if the clinician considers the evidence and then decides that the patient suffers from an eating disorder? In quantum theory, the mental state vector has to identify with the judgment outcome, so the new mental state vector has to be a normalized vector along the disorder ray, denoted as ψ'. This is a very constrained prediction for what happens to the mental state following a judgment (and concerns a fundamental theorem in quantum theory). For example, outside quantum theory, there is no reason why the judgment could not produce a mental state that is halfway between the judgment outcome and the original state (e.g., the green state vector Z in Figure 1 is not allowed) or indeed leave the original mental state unaffected. This is what motivates the use of quantum theory in the present work. Quantum theory could describe fairly naturally previous demonstrations of the evaluation bias (White et al., 2014, 2015, 2020).

We briefly introduce two more elements of quantum theory. First, belief updating can be represented by rotating the state vector. For example, suppose that the clinician receives some evidence that the patient is healthy. Then, the clinician's mental state vector is changed by rotating it towards the healthy subspace (to ψ'). Second, given some evidence, the clinician might think the patient is healthy, but respond that she has an eating disorder. That is, there may be a mismatch between the judgment outcome and the change in mental state. In quantum theory, there are technical tools which allow noise in projection.

Putting everything together, in the present work we examine evidence for constructive influences (evaluation biases) in clinical judgment (concerning eating disorders). This is an interesting application, because of the high expectations for rational judgment in clinical diagnosis and because of the apparent complexity of such judgments. We focus our rational analysis on a specific issue: by applying quantum models to the results, we can examine evidence of constructive processes in our participants' judgments. But it should be clear that in this case such constructive processes are not rational. In the case of a clinician evaluating a

patient, if the clinician makes an intermediate judgment, this has no impact on the probabilities that the patient is actually suffering from an eating disorder or not. In such a case, an evaluation bias is irrational. The application of quantum theory can then inform whether such a clinician mistakenly adopts quantum processes when this is not appropriate vs. whether he/she employs some heuristic or bias outside quantum (or Bayesian) theory. The latter possibility is taken into account by applying a plausible heuristics model, the belief adjustment model (Hogarth & Einhorn 1992). Arguably, this informs the extent of irrationality (or noise; or maybe our inability to describe these results) in the clinician's judgment, but we defer further discussion for later.

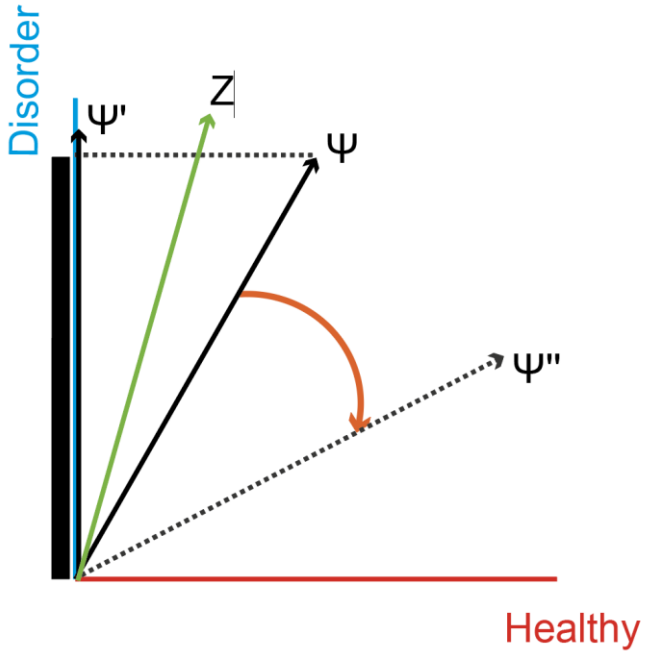


Figure 1. The mental state of a clinician initially thinking that a patient is likely to have an eating disorder is ψ. Following a judgment, if the clinician indeed decides that there is a disorder, the mental state has to become ψ', according to quantum theory; alternatives, such as Z, are not allowed. Finally, the presentation of some evidence that the patient is healthy (without a judgment) requires a rotation of the mental state towards the healthy subspace.

2. Material and methods

2.1 Participants

Three groups of participants took part in the experiment: 37 clinical psychologists (33 women, aged between 28 and 58 years, M=41.97; SD=7.23; with professional experience from two to 34 years, M=17.24; SD=6.81); 36 psychiatrists (29 women; aged between 30 and 59 years, M=42.28; SD=7.42; with professional experience raging from five to 34 years, M=15.86; SD = 7.41) and 33 participants without clinical background, students of psychology (29 women, aged between 21 and 40 years M=23.45; SD=3.49; their non-clinical professional experience ranged from zero to 14 years M=.82, SD=3.23). Participants were recruited in Poland. One of us (BI) used her professional network to identify participants willing to take part in the study and sample sizes were opportunistic (we recruited as many participants in the first two categories as was practical; the sample size for the students was then matched to the first two). Note, it is possible that this limits generalizability of the present results. Participants were informed that the aim of the research was to explore clinical decision making and the distinctive characteristics of psychiatrists, psychologists and students, for judgments relating to eating disorders. Participants received no compensation for taking part in the research. Ethics approval was granted by the Ethics Committee of the Faculty of Pedagogics and Psychology of the University of Silesia in Katowice, Poland.

2.2 Design

The dependent variable corresponded to the probability of disorder for the second piece of information, out of the two pieces of information which comprised each hypothetical patient case. There were three main factors. First, whether the first piece of information was rated or

not (single vs. double). Second, the valences of the two pieces of information in each patient case in the order processed by the participants (ordering condition, with four levels: health-health, disorder-disorder, health-disorder, and disorder-health). Finally, the rater profession (student, clinical psychologist, psychiatrist).

2.3 Materials

Between 2007-2014 a study was conducted on a population of 121 females with eating disorders (clinical group) and 92 healthy women (control group); the two groups were matched for age (Izydorczyk, 2011b). Participants in the clinical group suffered from a variety of eating disorders, including anorexia nervosa, bulimia nervosa, or binge eating disorder. Both the clinical and the control groups were comprised of females of similar age, namely early adulthood. Selection criteria for this earlier study were the presence (clinical group) or absence (control) of an eating disorder, as demonstrated by pre-existing medical diagnosis (based on the International Classification of Disorders, ICD, issued by the World Health Organization): anorexia nervosa, bulimia nervosa (F50) or a binge eating disorder (ICD10 F50.4). The individuals in the clinical group were being treated in Polish centers for eating disorders.

Out of the 213 participant cases as above, we first selected 36 cases aiming for an approximate balance between healthy and disordered cases. We then assessed whether corresponding case summaries correctly revealed the health or disorder of the patient. This was established by having four independent, competent raters – two experienced clinical psychologists and two psychiatrists experienced in diagnosis and treatment of eating disorders – examine each summary and determine whether it led to a conclusion of patient health or eating disorder. Only clinical case summaries of confirmed valence during this preliminary

stage, for which Kendall's concordance coefficient was over .70, were used in the main experiment. The above criteria allowed us to identify eight case summaries, each consisting of two parts: two cases in which both halves indicated eating disorders, two cases in which both halves indicated patient health, and four cases in which one half indicated health and the other eating disorder. Descriptions of the patients and their behavior in the "disorder" parts included symptoms of various eating disorders, whereas in the "healthy" parts neutral biographical data or symptoms of disorders different than eating disorders (Appendix 1).

2.4 Procedure

Participants received eight clinical case summaries, each composed of two parts. They could choose where and when to complete the task. Participants were instructed to read the descriptions carefully and to assign a probability that the patient is suffering from an eating disorder on a 10-point scale, with anchors: 1 - definitely healthy (does not suffer from eating disorders), to 10 - definitely disordered (suffers from an eating disorder). For each case, participants either rated the probability of the patient's disorder after reading the first and then after reading the second part of the summary (a rating for the first part, followed by a rating for both parts; double rating condition) or they provided a single rating for both parts (single rating condition; Figure 2). Presentation order of the two parts in each case was sometimes reversed between participants, so that for the same patient, some participants were shown the relevant information in one order and other participants in the reverse order. Note, there is no guarantee that participants would process the relevant information in the intended sequential way and some participants might go back to the first part, after having read the second part. Such flexibility in assessing the relevant information is closer to clinical practice.

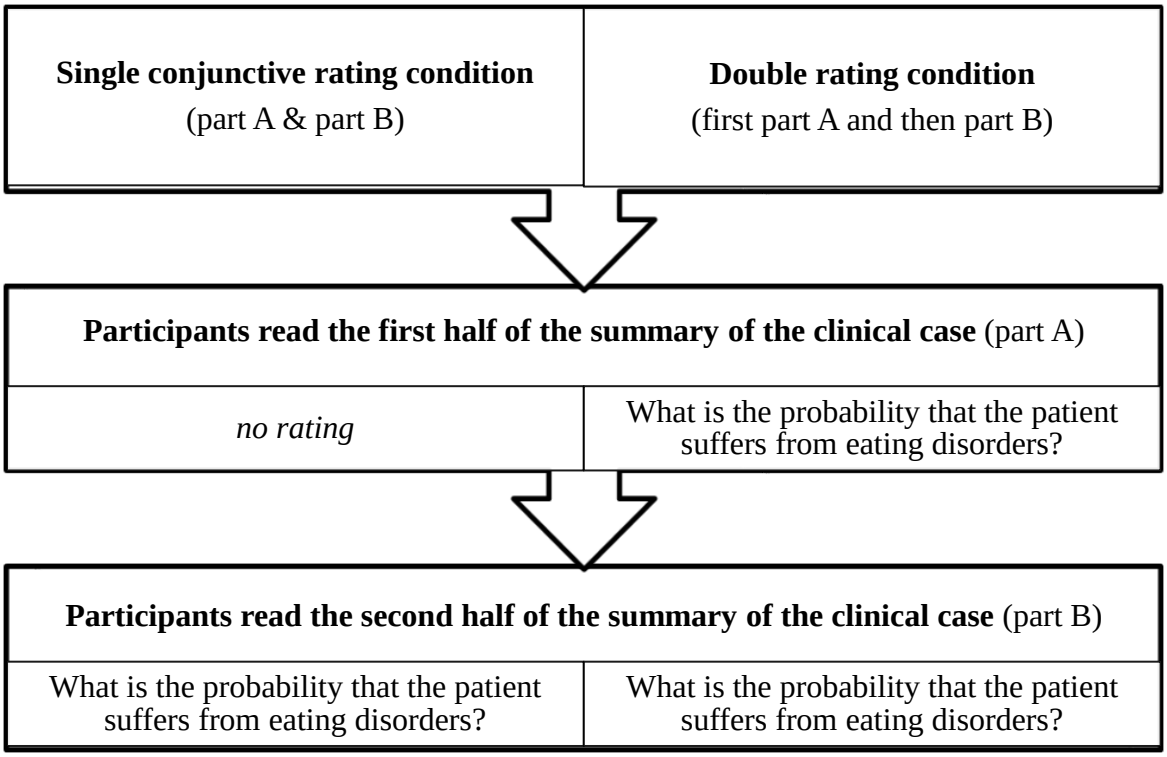


Figure 2. The main parts of the procedure.

3. Results

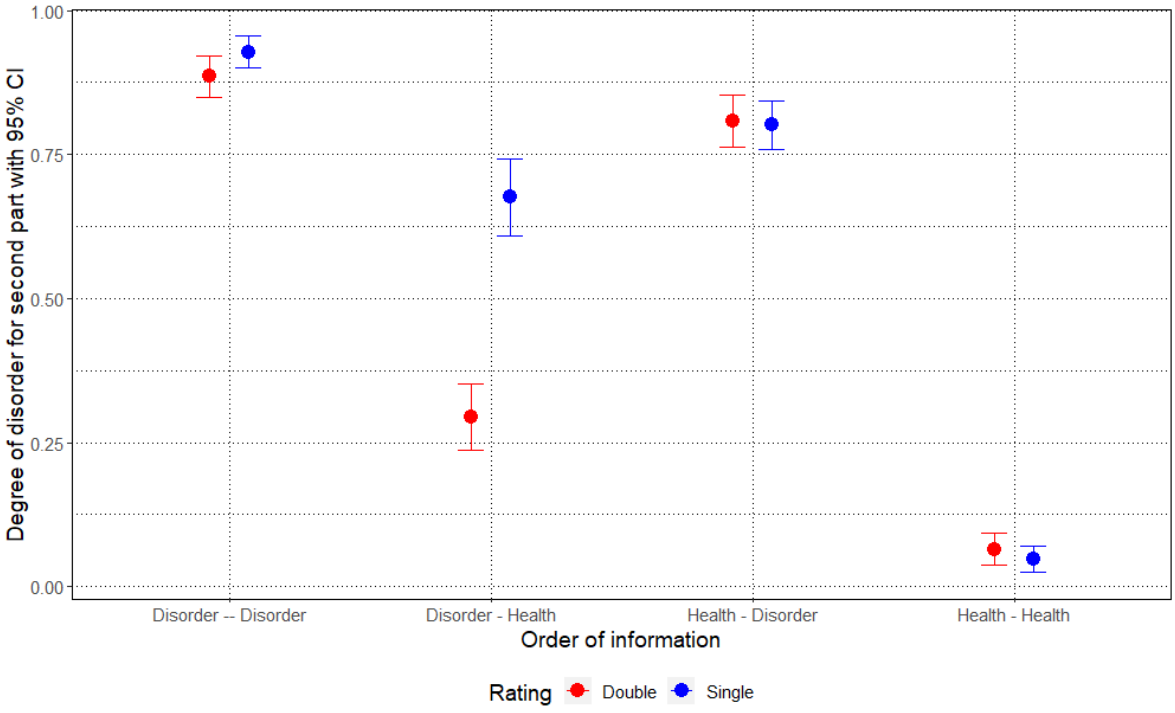


Figure 3. Summary of the results averaged across rater professions. The vertical axis indicates degree of disorder rescaled onto a [0,1] range.

Table 1. Comparison of the ratings from the pilot study with ratings in the experiment, for the eight cases which were eventually selected.

		Part A	
	mean/ SD	mean/ SD	Bayes factor
Alicja	1/ 0	1.96/ 1.2	1.04
Anna	8.75/ 0.96	9.1/ 1.11	2.05
Celina	1/ 0	2.22/ 1.54	1.02
Joanna	2.5/ 1.73	3.83/ 2.43	1.58
Malgorzata	2/ 0.82	2.92/ 1.85	1.7
Maria	1.75/ 0.96	2.7/ 1.54	1.42
Marta	2/ 0.82	3.24/ 2.11	1.49
Oliwia	9.5/ 1	9.22/ 1.13	2.16

		Part B	
	mean/ SD	mean/ SD	Bayes factor
Alicja	1/ 0	2.18/ 1.68	1.22
Anna	8.25/ 2.87	9.35/ 1.25	1.18
Celina	1/ 0	2.25/ 1.67	1.14
Joanna	9.5/ 0.58	9.24/ 0.79	2.03
Malgorzata	6.5/ 3.7	6.56/ 2.81	2.33
Maria	9.5/ 1	8.84/ 1.32	1.7
Marta	7.75/ 0.96	7.73/ 1.99	2.33
Oliwia	8.5/ 1.73	7.56/ 2.02	1.76

Note. The ‘Pilot’ column indicates ratings in the pilot study (N=4) and the ‘Experiment’ column ratings in the main experiment (N=848, corresponding to separate judgments). Each case is labeled by the name of a hypothetical person in the case.

We first considered the valence of the various pieces of information in the case summaries, by comparing the ratings for each part with the ratings of the four expert raters who were employed in the design stage of the study. It can be seen in Table 1 that there is high consistency between the preliminary ratings and averages from the main experiment (inferential tests are only indicative because of the small sample sizes). That is, case parts assumed to reflect eating disorders vs. health were perceived so by participants in the main experiment.

The main purpose of the behavioral analyses is to examine whether there is an evaluation bias and whether this depends on the profession (and so relevant expertise) of the participants. In the health—disorder condition, the second part indicates disorder; we therefore expect mean ratings in the double rating condition to be higher than in the single rating one (recalling that the dependent variable corresponds to degree of disorder). In the disorder—health condition, the second part indicates health, and so the evaluation bias prediction is that the mean in the double rating condition would be lower (indicating a stronger rating for health) than in the single rating one. The two health—health and disorder—disorder conditions were control ones, for which we assume equality in the mean rating for the first and second parts. Figure 3 shows the overall results, that is, mean disorder ratings, depending on different combinations of health and disorder for the two parts.

We ran a mixed effects model (multilevel linear model in SPSS with both random and fixed effects) with the probability of disorder for the second part (which we will abbreviate as *Prob*(*disorder*, *second part*)) as the dependent variable, three fixed effects, and two random effects. The fixed effects were whether there were judgments just on the second part vs. on both parts vs. (single vs. double), whether the parts were ordered according to health—health, disorder – disorder, health—disorder, disorder – health (ordering condition), and the rater profession (psychologist, psychiatrist, or student). The random effects corresponded to the hypothetical patient cases (case) and the participant. The dependent variable was rescaled onto the [0, 1] range, to resemble probabilities for disorder for the judgment following the second part, but we treated the variable as unbounded (that is, not restricted to the [0, 1] range).

Detailed results are shown in Appendix 2. The best model for *Prob*(*disorder*, *second part*) included the fixed effects and their two-way interactions and the random effects were modeled with both intercepts and slopes for the ordering condition (diagonal covariance

matrix). The two-way interaction between the single vs. double and ordering fixed effects was significant ($F(3,848)=36.9$, $p<.0005$), which is a prerequisite condition for the evaluation bias. The lack of three-way interaction involving rater profession precludes the possibility that the evaluation bias depends on this factor.

Further analyses are needed to establish whether the two-way interaction between single vs. double and the ordering condition reflects an evaluation bias. We next ran separate mixed effects models for each ordering condition, with fixed effects rater profession, single vs. double, and the interaction between the two variables; random effects (case, participant) were modelled only with intercepts. These separate models test the differences in *Prob*(*disorder*, *second part*) between single vs. double conditions, for each of the four groups in Figure 3. For the health—disorder, disorder—health, health—health, and disorder—disorder, we observed respectively $F(1, 227)=.07$, p=ns; $F(1, 197)=72.8$, $p<.0005$; $F(1, 211)=.99$, p=ns; $F(1, 213)=2.92$, $p=.09$.

To conclude, regarding the evaluation bias, results indicate a difference consistent with expectation in the disorder—health case, but not the health—disorder one. We can provisionally suggest that there is enough evidence for constructive effects to examine them with formal models. Note, it is puzzling that there is a trend for non-equality in the disorder—disorder case. The lack of interaction effects with rater profession also undermines the expectation that more experienced professionals might be less susceptible to this judgment bias in their diagnosis.

4. Modeling examination

The partial evidence for an evaluation bias suggests that judgments were influenced by constructive influences in the eating disorders diagnostic task. This motivates fitting White et al.'s (2014, 2020) quantum model. The main characteristic of this model is that a judgment

entails the change of the mental state in a specific way (Figure 1). However, the complexity of clinical judgments makes it likely that representations and processes focused just on the presence or not of eating disorders do not capture all relevant behavior. Therefore, it is possible that a quantum constructive influence is relevant, but the basic quantum processes in White et al.'s (2014, 2020) model only offer a noisy approximation to behavior. In a quantum framework, one way to introduce noise is as a mismatch between a judgment outcome and the change in the mental state, and this is the approach we pursue. A final, reasonable possibility is that there is a constructive influence, but this influence is different from the one assumed in quantum theory. This possibility is captured well by Hogarth and Einhorn's (1992) belief adjustment model.

We introduce the notation FSDR (first stimulus, double rating condition) = the probability that the person in a case study suffers from an eating disorder, after the first piece of information in the double rating condition; SSDR (second stimulus, double rating condition) = likewise, but after the second piece of information in the double rating condition; and SSSR (second stimulus, single rating condition) = likewise, but after the second piece of information in the single rating condition.

The basic quantum model for the evaluation bias is presented in detail in White et al. (2020) and we summarize its key components here. We assume a two-dimensional subspace, with one-dimensional subspaces for (eating) disorder vs. health. Consider the disorder-health case. A participant would first be exposed to information indicating disorder, so the initial mental state vector, ψ, is placed near the subspace for disorder (Figure 3). The angle between ψ and the disorder ray is labeled as *rating*. Introducing the health information results in a rotation, angle *n*, towards the health ray. In the single rating condition, the new mental state is ψ_{single}. In the double rating condition, after the initial disorder information, the judgment should first result in a projection of the state vector along the disorder ray. Then, introducing

the health information would rotate the state vector along angle *n*, giving us mental state ψ_{double}. In both cases, we compute the probability of health by projecting along the healthy ray, and it can be seen that $Prob(healthy, \psi_{double}) > Prob(healthy, \psi_{single})$, which is the evaluation bias (Appendix 3 offers the equations). The parameters of the quantum model are *rating* and *n* (both angles in an unrestricted range).

The noisy quantum model is specified analogously, but for an assumption of noise in the projection process, so that there is a probability, ϵ, that the judgment outcome is, e.g., disorder, but the mental state changes to align with health. The motivation for this elaboration is that we are trying to model a complex clinical judgment, using simple representations (one-dimensional subspaces) and simple dynamical processes. Therefore, noise is one way to represent the lack of correspondence between model assumptions and behavioral processes. The noisy quantum model is a novel proposal, not previously explored. The representations in the noisy quantum model are set up in a way analogous to that for the basic quantum model and the parameters *rating* and *n* are still in place. Additionally, there is a third noise parameter (ϵ, [0,1] range).

Hogarth and Einhorn's (1992) model is included in the present comparison because it offers an alternative perspective of constructive influences and the evaluation bias compared to the quantum models. We adopted White et al.'s (2020) formulation, who also applied the model to evaluation bias data (in Appendix 3 we summarize the equations). The model assumes an initial belief state. When the evaluation of the available information is SbS, the belief state is updated after each piece of evidence, in a way that reflects a compromise between the current state and the presented evidence. In the EoS case, there is a single revision in the belief state, once all evidence has been presented. In both cases, a change in the belief state can be interpreted as a constructive influence. However, the nature of the constructive influence is more flexible than in the case of the quantum models. Even adopting

several restrictions to the belief adjustment model (to make it possible to apply it to the present data, see Appendix 3), the change in the belief state is parametrically determined. The model has two parameters analogous to those of the basic quantum model.

As noted, regarding rationality, a constructive influence in the case of eating disorder diagnosis would be irrational either way. So, the question we are addressing by fitting the three models is whether participants are adopting an approach which *could* be rational (even if it is not in the present case; cf. the anxious teenager discussion) or one that has no obvious rational justification.

We conducted fits separately for data points conforming to disorder-health vs. points conforming to health-disorder, based on whether the first judgment was more consistent with an impression of disorder vs. health. The details of how the models were fitted are shown in Appendix 4. The essential point is that the data comprised of 12 triplets of SSSR, FSDR, and SSDR judgments (one for each patient case) and each model had two main parameters, with the noisy quantum model having an additional noise parameter. For each triplet separately, we used the SSSR and FSDR judgments to determine the two model parameters. Given these parameters, each model produces a prediction for SSDR. Applying the models in this way meant that we examined whether they predicted correctly the degree of constructive influence in each of the 12 cases. Model fit was assessed in terms of the closeness of correspondence between model predictions and observed data for the 12 SSDR judgments.

Table 2 shows the summary fit statistics, corresponding to Bayesian Information Criterion (BIC) values. BIC is a fit statistic penalizing for model complexity; lower values indicate better fit. Note that differences in BIC between 2 and 6 constitute ‘positive’ evidence for the model with the lower BIC value against the one with the higher BIC value (e.g., Kass & Raftery, 1995). Table 3 presents the 12 SSDR datapoints against which model performance was assessed. In the disorder-health case, it looks like all three models were unable to produce

constructive influences as large as those which were observed. In the health-disorder case, the noisy quantum model produced good fit, but the other two models produced constructive influences different to those which were observed.

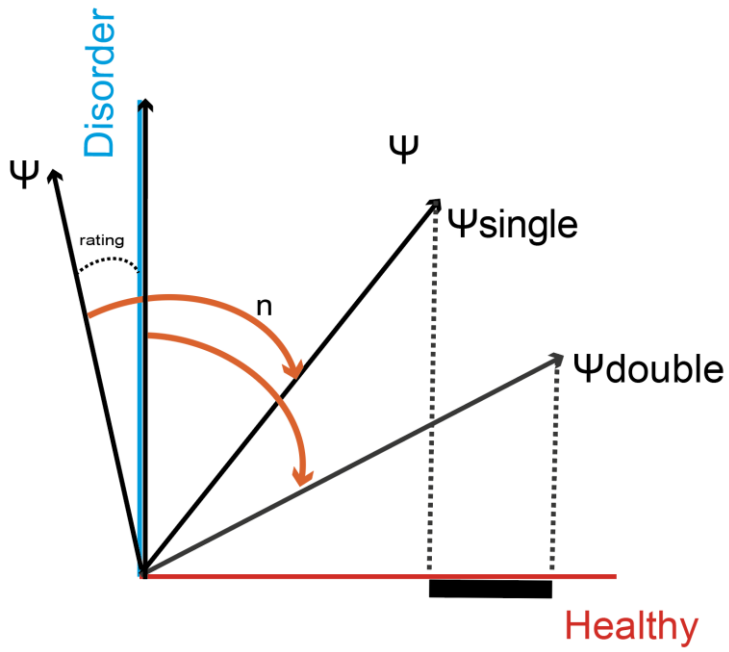


Figure 4. An illustration of the basic quantum approach (see text for explanation). The thick line along the healthy ray is a measure of the evaluation bias.

Table 2. BIC values for the three models.

	BIC value for disorder-health	BIC value for health-disorder
Model		
basic quantum	-16.56	-19.76
noisy quantum	-13.22 ($\epsilon = 0.49$)	-28.31 ($\epsilon = 0.1$)
belief adjustment	-14	-22.60

Note. BIC statistics for the three models examined in this work.

Table 3. Correspondence between observed and predicted FSDR values, for the basic quantum model, the noisy quantum model, and the belief adjustment model.

Condition	Observed SSDR	Predicted FSDR		
		basic quantum	noisy quantum	belief adjust.
Oliwia DD	.83	.47	.66	.88
Anna DD	.94	1	.72	.93
Marta DH	.26	.08	.61	.62
Malgorzata DH	.24	.21	.62	.68
Maria DH	.26	.27	.62	.69
Joanna DH	.42	.88	.6	.70
Celina HH	.07	0	.10	.03
Alicja HH	.06	0	.11	.06
Marta HD	.76	.99	.80	.82
Malgorzata HD	.6	.99	.77	.83
Maria HD	.9	1	.89	.84
Joanna HD	.93	.97	.87	.65

Note. Case names can be referenced against the materials in Appendix 1. The shaded cases corresponded to the disorder-health datapoints (see main text) and the unshared ones to the health-disorder ones. DD, HH, DH, and HD refer respectively to disorder-disorder, health-health, disorder-health, and health-disorder.

5. Discussion – in search of rationality

It is worth highlighting that understanding the present judgments was not expected to be straightforward. Each stimulus was a description of a patient case, which was realistic, fairly extensive, and with many unique characteristics. Participants included highly experienced professionals with relevant clinical or psychiatric experience. Attempting to understand such judgments with simple models, embodying no more than three parameters, has been ambitious and goes beyond the more standard, highly controlled experimental paradigms in cognitive science. Indeed, we would have weak expectations concerning the generalizability of the empirical data, beyond the conclusion that clinical judgements of this kind *can* show evaluation biases, as the empirical analyses and model fits showed. Regarding the latter, by setting model parameters using part of the data and examining model predictions

for the rest, we believe we have more direct tests of the key assumptions of the models, that beliefs appear to change when there is a judgment after the first piece of information.

Is there evidence that participants behave rationally in the present tasks? Constructive influences specifically consistent with the quantum model would be rational, under certain circumstances, because of the normative justification for quantum theory (Pothos et al., 2017; i.e., a constructive influence inconsistent with quantum theory would not have this rational justification). So, if participants are following quantum principles in their behavior, concerning constructive influences, they are in principle following rational principles (there was good fit for the noisy quantum model, at least in the health-disorder condition). However, quantum theory provides rational prescription only when it matches well the situation at hand. Concerning constructive influences, it would be rational to follow quantum principles, only when the system which is the focus of probabilistic inference is subject to constructive influences (perhaps there is disturbance from measurement). In this case, this is not so! That is, there is a mismatch between constructive influences in the minds of the participants (the empirical evidence shows such influences) and the world (where there is no mechanism for constructive influences to occur). Therefore, it looks like participants are sometimes applying the rational principles from quantum theory concerning constructive influences, when this is not warranted by the world.

Regarding the disorder-health condition no model produced superior fits, so it is may be that participants followed quantum principles (albeit imperfectly) or it may be that they adopted more idiosyncratic heuristics. The heuristics model we adopted was the belief adjustment model. It is hard to see how the assumption of a belief adjustment process can be justified in rational terms, without some argument concerning the adaptive value of such principles. However, it is doubtful that the belief adjustment model (as presently applied) could capture all heuristic influences in the present results. Recall, we observed a small, not

statistically reliable evaluation bias in the health-disorder condition and a strong evaluation bias in the disorder-health condition. So, if the initial information indicated health, whether a judgment was made or not seemed to have little effect. However, if the initial information indicated disorder, making an initial judgment vs. not would have a (more) major impact on the subsequent judgment based on information indicating health. A possible explanation is that clinically oriented participants might be anxious about providing a false disorder diagnosis.

With these points in mind, we can ask what we learned from the present work. Theoretically, the quantum models formalize the idea that constructive influences in judgments can be rational, e.g., when a judgment disturbs the relevant system in a certain way. Quantum theory formalizes the way we can describe this disturbance and how a sense of rational inference can emerge from judgments which appear, classically, erroneous. This argument hinges on the consistency of quantum theory with particular requirements for normative probabilistic inference (Pothos, Busemeyer, Shiffrin, & Yearsley, 2017).

Empirically, it is clear that neither the quantum model nor the belief adjustment model offer fully satisfactory fits to the data. This is a significant conclusion, because both models are constructed on the basis of reasonable processes (given our overall understanding of decision making) and both models have previously received considerable support. It is possible that suitable revisions to the models might accommodate the data or that the data reflects strategies too idiosyncratic to the particular sample (psychologists and psychiatrists in a particular nation) for description with general principles. Without further empirical work, it would be premature to further revise the models. Either way, the conclusions concerning the models and the specific empirical patterns should be considered particular to the specific population we targeted and not necessarily generalizable to the general population (Simons et al., 2017).

Methodologically, it may seem unsatisfactory that model tests were carried out on 12 data points. The reason for this is that a within participants approach would have required each participant to conduct both single rating and double rating judgments (Appendix 4). This would introduce task demands, since experienced clinicians would be unlikely to offer differing judgments for the same cases across the two conditions. Therefore, we had to focus on a within items approach, whereby we were limited by the number of suitable patient cases we could identify and the time it would take for participants to go through multiple cases. However, we contend that these limitations are not severe. Averaged data across participants are sometimes used to test decision models (e.g., this is how models for prisoner's dilemma are often tested, e.g., Pothos & Busemeyer, 2009) and one advantage of employing averaged data is that they offer insight onto response biases (fairly) independently of idiosyncratic strategies.

This work allows some recommendations for how to further study the rationality of behavior in realistic, clinical situations. There is some evidence that participants might be misapplying rational principles concerning constructive influences. Is this because, in their experience, there are situations when such rational principles would be appropriate? It would be interesting to extend the present paradigm when the clinician's judgment can change the relevant system. For example, consider clinical judgments concerning e.g. anxiety or depression, offered directly to participants. An anxious participant being told that some information about his/her profile indicates anxiety might plausibly affect the participant's anxiety levels (thus altering or disturbing the 'system'). Clearly, it would not be possible to conduct such a study in the laboratory, but maybe relevant data can be obtained through the observation of practicing clinicians.

It is also unsurprising that there is room to elaborate the modeling approaches. Within a quantum framework, an asymmetry between the health-disorder and the disorder-health

conditions can naturally arise if we postulate that the concept (and so subspace) for health is larger than for disorder, in the sense that there are more ways in which a person can be healthy than suffering from a specific disorder (Pothos, Busemeyer, & Trueblood, 2013). But it is unclear whether this would suffice to accommodate the very large constructive influence in the disorder-health condition. More generally, there are several promising possibilities, and we outline some of them below, based on recent work with Bayesian theory and the adaptive toolbox.

A probabilistic language of thought approach follows the general principles of hierarchical Bayesian modelling, according to which when we encounter a stimulus we infer some information both about the stimulus and the broader category the stimulus belongs to (a hypothesis and an over-hypothesis; Kemp et al., 2007). So, in the e.g. disorder-health direction, when encountering the first stimulus, a participant first infers that the clinical case indicates disorder. Then, the subsequent health information is evaluated from the point of view of a probability distribution for disorder information. The more narrow this distribution about disorder, the more extreme the subsequent impression regarding health. Under what circumstances would we expect a narrow, specific distribution for disorder information from the first stimulus? An overarching assumption in communication and pragmatics is that, in general, speakers are helpful (Goodman & Frank, 2016). So, from a participant's point of view, the request to rate vs. not rate the first piece of information may create a covert bias to process this information in a deeper vs. more superficial way. Overall, this picture can accommodate a prediction of judgments indicating more disorder, when the first disorder information is rated, because the resulting distribution would be sharper and less noisy.

According to the adaptive toolbox framework, one of the key tenets of rational behavior is that an available heuristic is chosen to match the properties of the task and environment at hand (Gigerenzer and Selten, 2001). A fast-and-frugal heuristic that ignores

information, or integrates information in a simpler fashion, often performs better than a complex model by trading off more bias with less variance (Gigerenzer and Brighton, 2009). This is particularly advantageous for small and/or uncertain data samples—in such a case, heuristics can outperform complex models (the less-is-more effect, Gigerenzer et al, 1999). Consequently, the appropriate heuristic to choose from the adaptive toolbox depends on properties of the information set. A decision maker called upon to provide a rating only after viewing the whole informational sequence may use a different heuristic than when providing two separate ratings after each part. Furthermore, the temporal order and partitioning of information may signal the relative importance or validity of the information. Such a picture could explain a dependence of judgments on the exact presentation characteristics of the stimuli. Note, the adaptive heuristics toolbox may appear to present a view of rationality distinct from that of Bayesian theory or quantum theory, but in both cases there have been discussions attempting to converge the different frameworks (for the former see Lieder & Griffiths, 2019; for the latter see Kvam & Pleskac, 2017).

Finally, from the perspective of resource-rational analysis (Icard & Goodman, 2005), people may neglect the first stimulus unless the task makes it worthwhile to consider it. When the previous stimulus was not rated, then people use the simplest possible causal model that includes only two variables: the second stimulus (cause) and their current experience (effect). But when people previously rated the first stimulus, then their causal model includes three variables: the first stimulus (cause 1), the second stimulus (cause 2), and their current experience (effect). This might be resource-rational because when people have not rated the first stimulus, incorporating its effect into the judgment for the second one would require more effort. As a consequence, the contrast between the first and second stimulus becomes more pronounced when people rate the first stimulus. Moreover, the model of utility-weighted learning (Lieder, Griffiths, & Hsu, 2018) offers a perspective on why the evaluation bias

appears for disorder-health stimuli, but not for health-disorder ones. According to the model, extreme events come to mind more readily than neutral events because 1) they are more important to consider, and 2) it is more valuable to consolidate their memories. This line of reasoning suggests that, in contrast to the emotionally salient information about the disease, the initial information about health does not contaminate the subsequent judgment because it is not important enough to warrant further attention.

All these perspectives could reveal, to varying degrees, rational facets of the evaluation bias. However, equally, some of these theoretical elaborations require higher model parameterizations and so more complex experimental paradigms. A tension between the simplicity of the experimental paradigm (and so the intuitive compellingness of an empirical result) and the sufficiency of the dataset to discern subtle modeling assumptions is hardly new, of course, and one of the most central challenges in developing our understanding of rationality in human behavior. We hope the present results provide enough groundwork for the further elaboration of the paradigm and relevant theory, with a view to provide a more accurate evaluation of the rational status of constructive influences.

Acknowledgements

BW, PW, and EMP were supported by ONRG grant N62909-19-1-2000. We would like to thank the colleagues who helped us with the way the alternative frameworks could cover the key findings in this paper: MH Tessler for probabilistic language of thought, Leonidas Spiliopoulos for the adaptive toolbox, and Falk Lieder for resource-rationality. We would like also to thank Joyce Wang and two anonymous reviewers for their helpful comments.

References

American Psychological Association Presidential Task Force on Evidence-Based Practice (2006). Evidence - Based Practice in Psychology. American Psychologist, 61(4), 271-285.

Ariely, D., & Norton, M.I. (2008). How actions create – not just reveal – preferences. Trends in Cognitive Sciences, 12, 13-16.

Arkes, H.R. & Harkness, A.R. (1980). Effect of making a diagnosis on subsequent recognition of symptoms. Journal of Experimental Psychology: Human Learning and Memory, 6, 568–575.

Bauer, R.M. (2007). Evidence - Based Practice in Psychology: Implications for Research and Research Training. Journal of Clinical Psychology, 63, 685 - 694.

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Oppliger, R. A. (1998). Clinical diagnosis and order information. Medical Decision Making, 18, 412-417.

Brehm, J.W. (1956). Post-decision changes in the desirability of choice alternatives. Journal of Abnormal and Social Psychology, 52, 384–389.

Broekaert, J. B., Busemeyer, J. R., & Pothos, E. M. (2020). The disjunction effect in two-stage simulated gambles. An experimental study and comparison of a heuristic logistic, Markov and quantum-like model. Cognitive Psychology, 117, 101262.

Busemeyer, J. R., & Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge, UK: Cambridge University Press.

Busemeyer, J. R., Pothos, E., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment 'errors'. Psychological Review, 118(2), 193-218.

Cierpiałkowska, L. & Sęk, H.(2016). Psychologia Kliniczna. Wydawnictwo Naukowe PWN, Warszawa.

Currin, L., Walter, G., Treasure, J., Nodder J., Stone, C., & Yeomans, M. (2007). The use of guidelines for dissemination of 'best practice' in primary care of patients with eating disorders. International Journal of Eating Disorders, 40, 476 - 479.

Field, A. (2017). Discovering statistics using IBM SPSS Statistics. Sage Publications Ltd.

Garb, H.N. (1998). Studying the clinician: Judgment research and psychological assessment. American Psychological Association, Washington DC.

Garner, D. (2004). EDI 3 Eating Disorder Inwentory-3. Psychological Assessment Recources,Inc., Florida Avenue-Lutz.

Gigerenzer, G., Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences Topics in Cognitive Science 1(1), 107-143. https://dx.doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., Todd, P., and the ABC Group (1999). Simple heuristics that make us smart Oxford University Press.

Gigerenzer, G., Selten, R. (2001). Bounded rationality: The adaptive toolbox MIT Press.

Gloeckner, A., Betsch, T., & Schindler, N. (2009). Coherence shifts in probabilistic inference tasks. Journal of Behavioral Decision Making, 23, 439 – 462.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in cognitive sciences, 20, 818-829.

Hermann, R.C., Ettner, S. L., Dorwart, R. A., Langman-Dorwart, N., & Kleinman, S. (1999). Diagnoses of patients treated with ECT: A comparison of evidence-based standards with reported use. Psychiatric Services, 50, 1059–1065.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: the belief-adjustment model. Cognitive Psychology, 24, 1–55.

Huppert, J.D., Fufka, L.F., Barlow, D.H., Gorman, J.M., Shear, M.K., Woods, S.W. (2001). Therapist, therapist variables and cognitive - behavioral therapy outcome in a multicenter trial for panic disorder. Journal of Consulting and Clinical Psychology, Vol. 69, pp. 747 - 755.

Icard, T., & Goodman, N. D. (2015). A Resource-Rational Approach to the Causal Frame Problem. In CogSci 2015.

Izydorczyk, B. (2011a). A psychological diagnosis of the structure of the body self in a group of selected young Polish females without eating or the mental disorders. Archives of Psychiatry and Psychotherapy, Vol. 2, pp 21 - 36.

Izydorczyk, B. (2011b). A psychological profile of the bodily self characteristics in women suffering from bulimia nervosa [in:] Hay P (ed.), New insights into the prevention and treatment of bulimia nervosa. Croatia: Intech Open Access Publisher, 2011, pp. 147-167.

Izydorczyk B. (2013). Selected psychological traits and body image characteristics in females suffering from binge eating disorder. *Archives of Psychiatry and Psychotherapy*, Vol. 1, pp. 19–33.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences, 34, 169, 231.

Kahneman, D. (2001). Thinking fast and slow. Penguin: London, UK.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773-795.

Kayne, N.T. & Alloy, L.B. (1988). Clinician and patient as aberrant actuaries: Expectation-based distortions in assessment of covariation. In , L.Y. Abramson (Ed.) *Social cognition and clinical psychology: A synthesis*, Guilford Press, New York, pp. 295–365.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental science, 10(3), 307-321.

Kvam, P. D., & Pleskac, T. J. (2017). A quantum information architecture for cue-based heuristics. Decision, 4, 197.

Lieder, F. & Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behavioral and Brain Sciences, 43, 1-85.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. Psychological review, 125, 1-32.

Lilienfeld, S.O., Ritschel, L.O., Lynn, S.J., Brown, A.P., Cautin, R.L., & Latzman, R.D. (2013). The research - practice gap: Bridging the schism between eating disorder researchers and practitioners. International Journal of Eating Disorders, 46, 386 - 394.

Muller, M.J. & Davids, E. (1999). Relationship of psychiatric experience and inter-rater reliability in assessment of negative symptoms. Journal of Nervous and Mental Diseases, 187, 316–318.

Norcross, J.C. (2002). Psychotherapy relationships that work: Therapist contributions and responsiveness to patient needs. Oxford University Press, New York.

Oaksford, M. & Chater, N. (2009). Précis of Bayesian rationality: the probabilistic approach to human reasoning. Behavioral and Brain Sciences, 32, 69-120.

Pothos, E. M. & Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. Proceedings of the Royal Society B, 276, 2171-2178.

Pothos, E.M. & Busemeyer, J.R. (2013). Can quantum probability provide a new direction for cognitive modeling? Behavioral & Brain Sciences, 36, 255-327.

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. Psychological Review, 120, 679-696.

Pothos, E.M., Busemeyer, J.R., Shiffrin, R.M., & Yearsley, J.M. (2017). The rational status of quantum cognition. Journal of Experimental Psychology: General, 146, 968-987.

Sharot, T., Velasquez, C.M., & Dolan, R.J. (2010). Do decisions shape preference?: Evidence from blind choice. Psychological Science, 21, 1231–1235.

Shepard, R. N. (1992). The Perceptual Organization of Colors: An Adaptation to Regularities of the Terrestial World? In Barkow, J. H., Cosmides, L., Tooby, J. (Eds) The adapted mind, Oxford: Oxford University Press.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6), 1123-1128.

Straus, S.E., Glasziou, P., Richardson, W.S., & Haynes, R.B. (2011). Evidence - based medicine: How to practice and teach EBM. Churchill Livingstone, New York.

Tenenbaum, J. B, Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. Science, 331, 1279-1285.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjuctive fallacy in probability judgment. Psychological Review, 90, 293-315.

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. Proceedings of the National Academy of Sciences, 111, 9431–9436.

White, L.C., Pothos, E.M., & Busemeyer, J.R. (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. Cognition, 133, 48-64.

White, L. C., Barque-Duran, A., & Pothos, E. M. (2015). An investigation of a quantum probability model for the constructive effect of affective evaluation. Philosophical Transactions of the Royal Society A, 374, 20150142.

White, L.C., Pothos E.M., & Jarrett, M. (2020). The cost of asking: how evaluations bias subsequent judgments. Decision, 7, 259-286.

Youngstrom, E.A. (2013). Future Directions in Psychological Assessment: Combining Evidence - Based Medicine Innovations with Psychology's Historical Strengths to Enhance Utility. Journal of Clinical Child and Adolescent Psychology, 42, 139 - 159.

All appendices are intended to be online supplementary materials and so not included in the word count.

Appendix 1

Additional information regarding the eight cases of individuals with or without eating disorders employed in the experiment.

We provide summaries of the clinical case descriptions used in the study, arranged so that the first column indicates part A and the second part B. Recall that during the experiment the order of sequential presentation for the two parts was counterbalanced. The Health and Disordered characterizations after each story concern just eating disorders, as relevant to the design of the study.

Patient's name	Part 1 of the description	Part 2 of the description

Alicja	Alicja searched for professional support because she has been experiencing fear of sudden death or illness, loss of job and social anxieties, especially in situations related to professional work. Alicja reported lack of satisfaction with her life (also at work), the presence of a depressed mood and negative thinking about herself. Lately she had been suffering from sadness, anxiety, dissatisfaction, feelings of loneliness and helplessness, as well as pessimism. Alicja also complained about memory losses and attention deficits. *Healthy, Polish version used in the study had 107 words)*	Alice describes herself as an ambitious, sociable person, but also presents submission and excessive dependence on the approval of the others. She experiences fear of risk, hypersensitivity to criticism and poor control of expression of the emotions, constant feelings of inferiority. While describing social relationships from childhood to adulthood (including the work environment) one gets an impression of low self – esteem and difficulties with building relations. *(Healthy, Polish version used in the study had 120 words)*
Anna	Anna does not have internal motivation to start the treatment. During the consultation she denies the significance of weight loss noticed by her parents. At the time of admittance Anna weighs 47 kg and is 167 cm tall. She has not menstruated for several months. Due to cardiac arrhythmias she has been seen by a cardiologist. Anna's mother observed persistent restrictive slimming and physical exercise, avoiding meals with the family and visible weight loss. *s (Disordered, Polish version used in the study had 292 words)*	Anna undertakes restrictive slimming (including fasting) and intense physical exercise. Anna is afraid of weight gain, controls meals, and is alienated from her peers. Anna likes her current body shape, but can't imagine getting fat. She has increasingly reacted with anger at attempts to make her increase the amount of food consumed every day. She feels fat and unattractive. In addition to the growing dysphoric mood in adolescence she demonstrates emotional instability, perfectionism and desire for full control. *(Disordered, Polish version used in the study had 344 words)*

Celina	Celina seeks clinical help for the first time in her life because of a condition she has been feeling for several months described as a "severe nervous breakdown." When asked what this means, she answers: "I can't sleep, I have nightmares, I'm afraid, I don't want anything, I have felt bad all the time, I don't feel good, my partner wanted me to leave, so I moved out". *(Healthy, Polish version used in the study had 128 words)*	When Celina went to study away from her home town, she experienced anxiety. For several months she has also reported a feeling of being overwhelmed by negative thoughts, apathy and aversion to everyday activities, including professional activities. Till now Celina has not been hospitalized. *(Healthy, Polish version used in the study had 110 words)*
Joanna	Joanna searches for medical help for the first time in her life, saying that “she is under stress and has trouble eating and no appetite” because of recent death of her grandmother. She has had problems eating food for the last 6 months, she describes it as" growing food in my mouth ". The results of medical tests did not confirm presence of gastrointestinal diseases or other somatic diseases. Joanna reports weakness, apathy and frequent trouble falling asleep since her grandmother died. The onset of these symptoms appeared six months ago. *(Healthy, Polish version used in the study had 205 words)*	Joanna admits she eats a lot during the day and then uses various laxatives to get rid of food because she is afraid to gain weight. She also gets up at nights to eat something. She describes own body as "monstrous, ugly, sagging". Joanna is afraid to gain weight, she tries to exercise every day, but she can't be consistent She has been controlling calories for some time; she is often hungry, but she also eats more often. Initially she would eat 1500 calories, now she limits food to 600 calories or "throws herself on every food in the fridge" and eats everything she can find. Joanna weighed 60 kg. and had a height of 165 cm. *(Disordered, Polish version used in the study had 338 words)*

Malgorzata	After breaking up with a partner, Malgorzata experiences "nervous breakdown, sadness, depression (…) I couldn't get out of bed, the world collapsed". Since then, she has lost the appetite and is unwilling to eat anything. She has alienated herself from her college environment, stopped contact with her friends, but has been attending classes and passing her exams. Marta has always been a perfect exemplary student. *(Healthy, Polish version used in the study had 187 words)*	She has difficulties eating and notices weight loss. For several years, she has been consulted by doctors because of recurring fainting, mood changes, anxiety, and apathy. Malgorzata weights 57 kg with a height of 170 cm. She is unhappy with her appearance. Marta has been using a variety of diets and is intensely exercising at the. She often feels frustrated, sad, and complains about mood swings. She occasionally reaches for amphetamines. *(Disordered, Polish version used in the study had 221 words)*
Maria	She visits a doctor due to stomach aches and diarrhea getting worse, especially during stressful exam sessions. Medical tests do not confirm the presence of gastrointestinal or other somatic diseases. At the time of going to the doctor, she reported a weight loss of 5 kilos over the last two months, an increase in apathy, unwillingness to act, and problems with professional activity, which are associated with frequent periods of depressed mood. *(Healthy, Polish version used in the study had 221 words)*	After losing 7 kilograms she noticed a change in the appearance, began limiting food (ate less, introduced reduction of fats, carbohydrates and other nutrients, and started counting calories). She is unhappy with her body, often feels fat and unattractive, and has low self-esteem. She often experiences headaches, fatigue and has trouble concentrating. She exercises intensively every dayShe is on ses a diet of no more than 1000 calories per day. *(Disordered, Polish version used in the study had 274 words)*

Marta	Marta attends medical studies in a city located a few hundred kilometers from her family home. Since the beginning of the academic year she has been experiencing lack of appetite and unwillingness to eat anything. Marta alienates from her college environment, breaks her contacts with her friend, but attends classes and passes her exams. Marta has always been an exemplary student. *(Healthy, Polish version used in the study had 177 words)*	In an interview with a psychiatrist, she says that she has difficulties eating and noticed weight loss. She has been consulted by doctors several times because of recurring fainting, anxiety and apathy. She weights 57 kilograms and is 170 cm tall. She has not menstruated for months. She is unhappy with her appearance, keeps controlling the calories, and weighs herself daily. As a teenager, she had followed diets and regularly exercised at the gym. *(Disordered, Polish version used in the study had 210 words)*
Oliwia	Olivia can't cope with herself, stress and binge eating. She says that she can't think and focus on anything except for food. She reaches for food even when she doesn't feel hungry. Twice a day she vomits after meals, but she is not using laxatives. Medical tests did not show any malfunctions of the digestive system or other somatic disorders. Sometimes she practices fasting and exercise. *(Disordered, Polish version used in the study had 400 words)*	Olivia had been on medication for two years due to periodically increasing mood swings and periods of severe depression. She also admits to acts of verbal aggression. In addition to the growing dysphoric mood in adolescence, Olivia has observed a pattern of emotionally-based responses and a tendency for compulsiveness and emotional instability. She does not accept own appearance. *(Disordered, Polish version used in the study had 207 words)*

Appendix 2

We first consider the mixed effects models, which were run to establish the presence of an evaluation bias. To briefly review the structure of the models, we employed three fixed effects, and two random effects. The fixed effects were whether there were judgments on both parts vs. on a single part (single vs. double), whether the parts were ordered according to health—health, disorder – disorder, health—disorder, disorder – health (ordering condition), and the rater profession (psychologist, psychiatrist, or student). The random effects corresponded to the hypothetical patient case (case) and the participant. The analyses employed a Maximum Likelihood Estimation (MLE) method. Model fits are expressed by minus twice log likelihoods (-2LL) and nested models are compared using chi squared distributions for the difference in model fits, with degrees of freedom corresponding to the difference in model parameters. The significance level is taken to be .05.

As is typical in mixed effects analyses, a series of nested models has to be run in order to establish, first, which interaction terms can be retained and, second, whether random effects have to be modeled with just intercepts or with intercepts and slopes and the structure of the covariance matrix (if slopes are employed). Clearly, situations which would be fairly straightforward in the case of traditional ANOVA designs can entail a large number of possibilities/ models in mixed effects designs. We proceeded in a standard way closely following the recommendations of Field (2017).

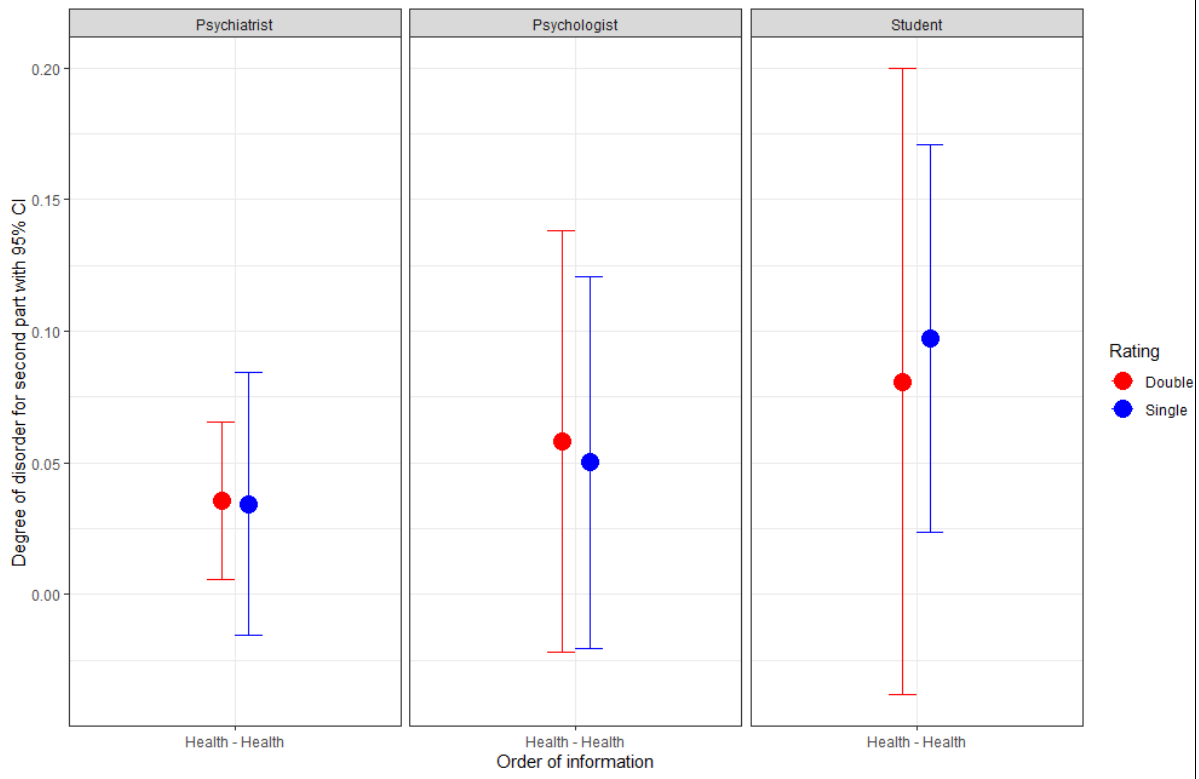
As noted, we first consider which interaction terms can be retained. For the models with only main effects, two-way interaction terms, and the three-way interaction term we observed respectively -2LL values and parameters of -67.6, 9; -184.6, 20; and -186.9, 26 (lower -2LL values indicate better fit). The model with two-way interactions is clearly superior to the model with just main effects, $\chi^2(11) = 117$, p<.0005, but the model with a three-way interaction did not constitute a significant improvement, $\chi^2(6) = 2.3$, p=ns. This

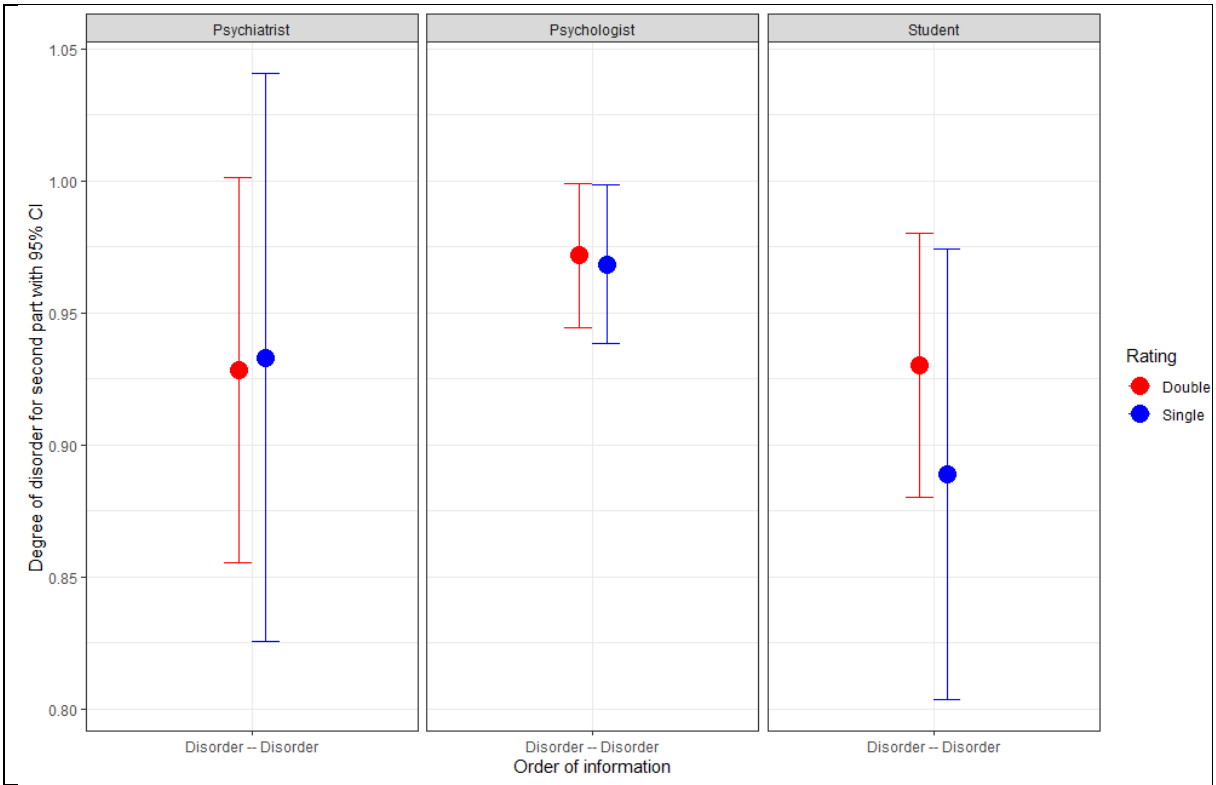
first part of the analyses allows us to reach two conclusions. First, rater profession does not interact with the presence or not of the evaluation bias. The evaluation bias requires an interaction between the single vs. double factor and the ordering factor. So, if the rater profession impacted on the evaluation bias, then this would require a corresponding three-way interaction. As this was not observed, we can dismiss the rater profession factor. Second, the model with two-way interactions showed the crucial single vs. double, ordering condition interaction to be significant ($F(3,848)=36.9$, $p<.0005$), thus offering preliminary evidence for an evaluation bias of some kind (full evidence for an evaluation bias is not possible unless the interaction is examined more carefully).

Subsequent analyses include all main fixed effects and their two-way interactions (-2LL and parameters: -184.6, 20). We then considered whether the modelling of the random effect could benefit from slopes for both the ordering condition and the rater profession fixed effects (it is less meaningful to employ slopes for the single vs. double fixed effect too). We initially restricted the covariance matrix to the simplest possible option of variance components. The models with slopes for just the ordering condition, just the rater profession, and for both fixed effects had -2LL values and parameters as, respectively, -184.6, 21; -184.6, 21; -184.6, 22. It is immediately clear that these models do not offer an improvement compared to the model with intercepts only for the random effect. However, it is possible that the variance components covariance matrix fits the present data poorly. We therefore repeated these analyses with diagonal covariance matrices, observing -353.2, 24; -191.0, 23; -344.2, 27, and with unstructured covariance matrices, observing -319.1, 34; -191.0, 29; -263.2, 55. Overall, the best model is the one with slopes only for the ordering condition and a diagonal covariance matrix, which offers significant improvement over both the model without any slopes ($\chi^2(24-20) = 353.2 - 184.6 \Leftrightarrow \chi^2(4) = 168.6$, $p<.0005$) and the model with

slopes for the ordering factor, but with a variance components covariance matrix ($\chi^2(24-21) = 353.2 - 184.6 \Leftrightarrow \chi^2(3) = 168.6$, p<.0005).

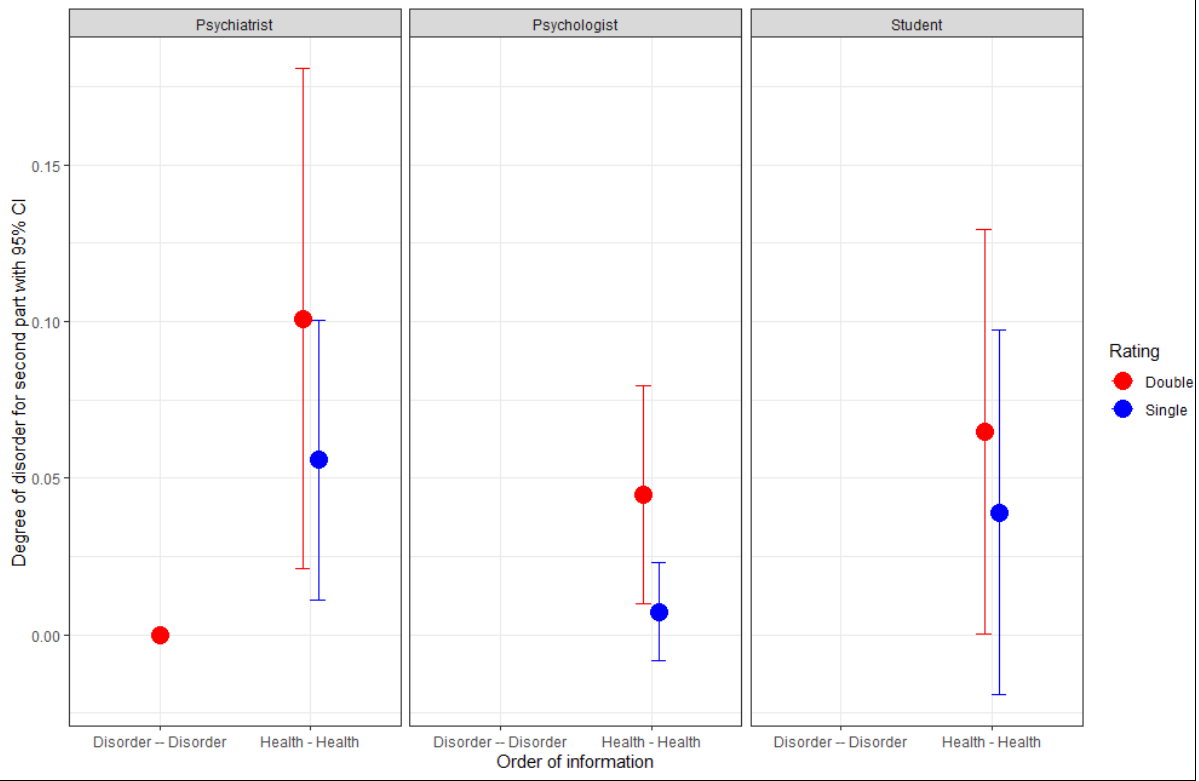
Finally, for descriptive purposes, we present figures analogous to Figure 3, but split across profession and case. One observes a pattern consistent with that in Figure 3 across most cases. As these comparisons are offered for illustration, for simplicity, we just ran ordinary independent samples t-tests, with a single fixed effect corresponding to single vs. double ratings (singe vs. double). For each cell we indicate whether the finding is consistent with the evaluation bias or not.

Students	Psychologists	Psychiatrists
Alicja 		
mean$_{\text{double rating}}$= .03 *mean*$_{\text{single rating}}$ =.03 t(34)=.04, p=.97	*mean*$_{\text{double rating}}$= .06 *mean*$_{\text{single rating l}}$= .05 t(35)=.16, p=.88	*mean*$_{\text{double rating}}$= .08 *mean*$_{\text{single rating}}$ =.10 t(31)=-.26, p=.80
Anna		



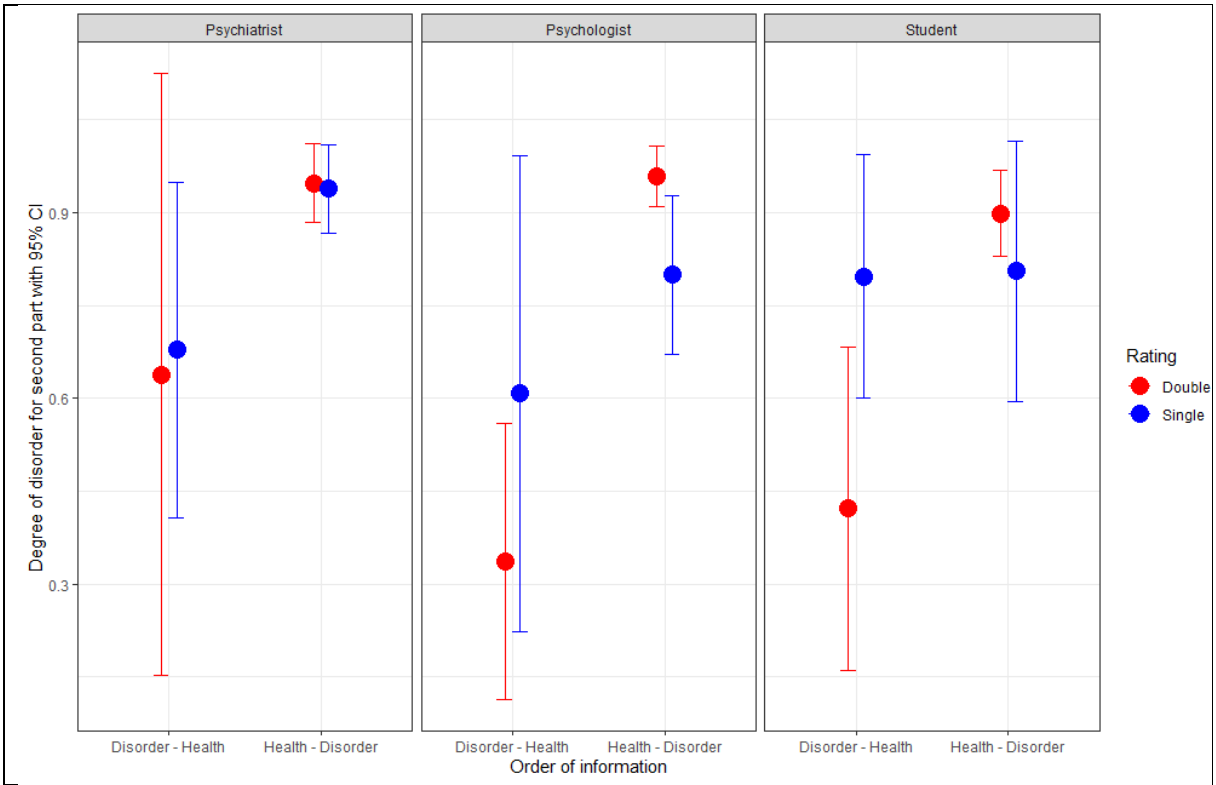
$mean_{\text{double rating}}$=.93 $mean_{\text{single rating}}$ =0.89 t(31)=.94, p=.35	$mean_{\text{double rating}}$=.97 $mean_{\text{single rating}}$ =.97 t(35)=.17, p=.86	$mean_{\text{double rating}}$=.93 $mean_{\text{single rating}}$ =.93 t(34)=-.08, p=.94

Celina

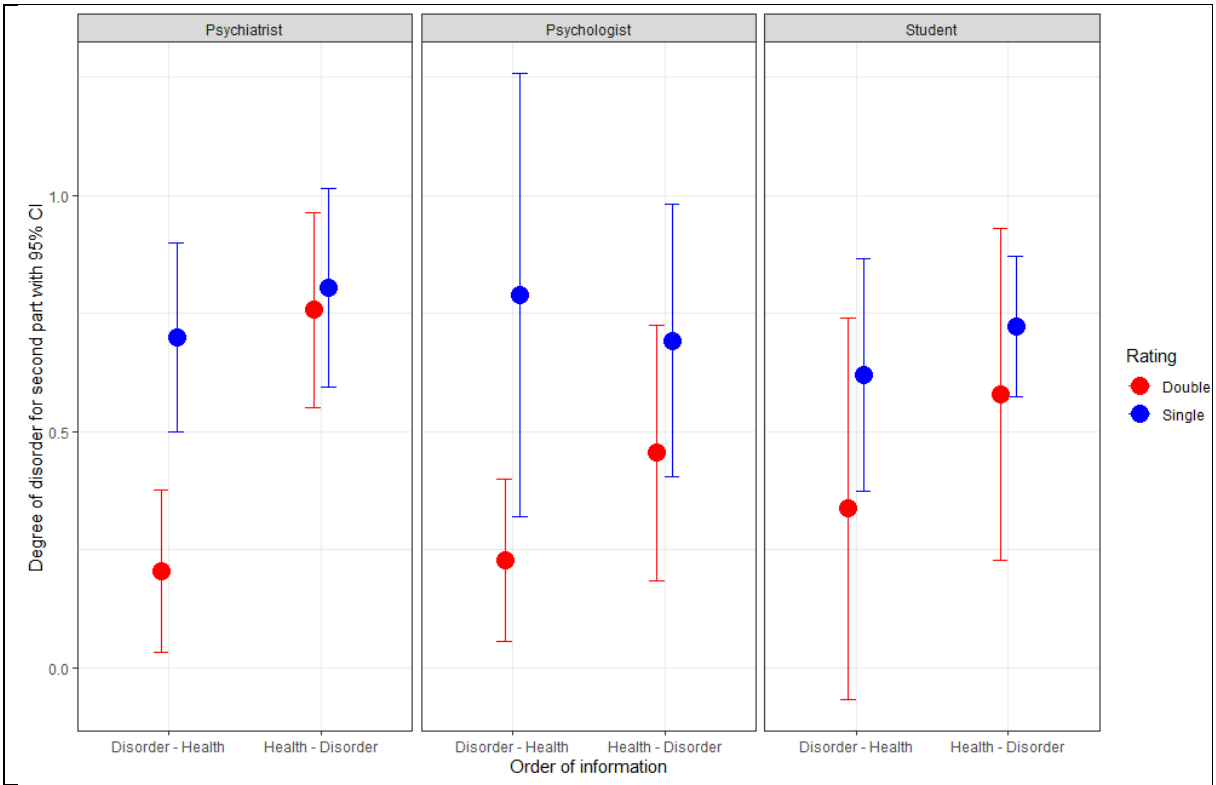


$mean_{\text{double rating}}$=.14 $mean_{\text{single rating}}$ =.00 (t(33)=.81, p=.42)	$mean_{\text{double rating}}$=.04 $mean_{\text{single rating}}$ =.01 t(35)=1.93, p=.06	$mean_{\text{double rating}}$=.06 $mean_{\text{single rating}}$ =.04 t(31)=.63, p=.54

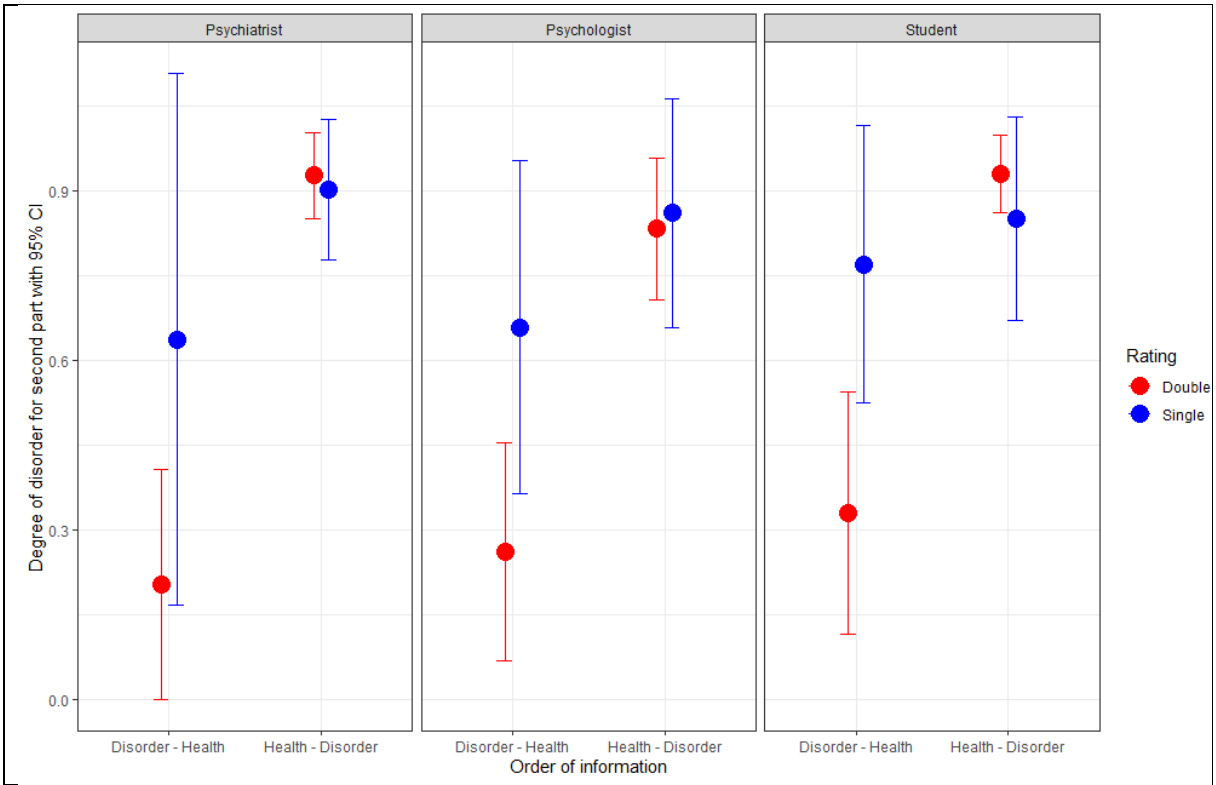
Joanna



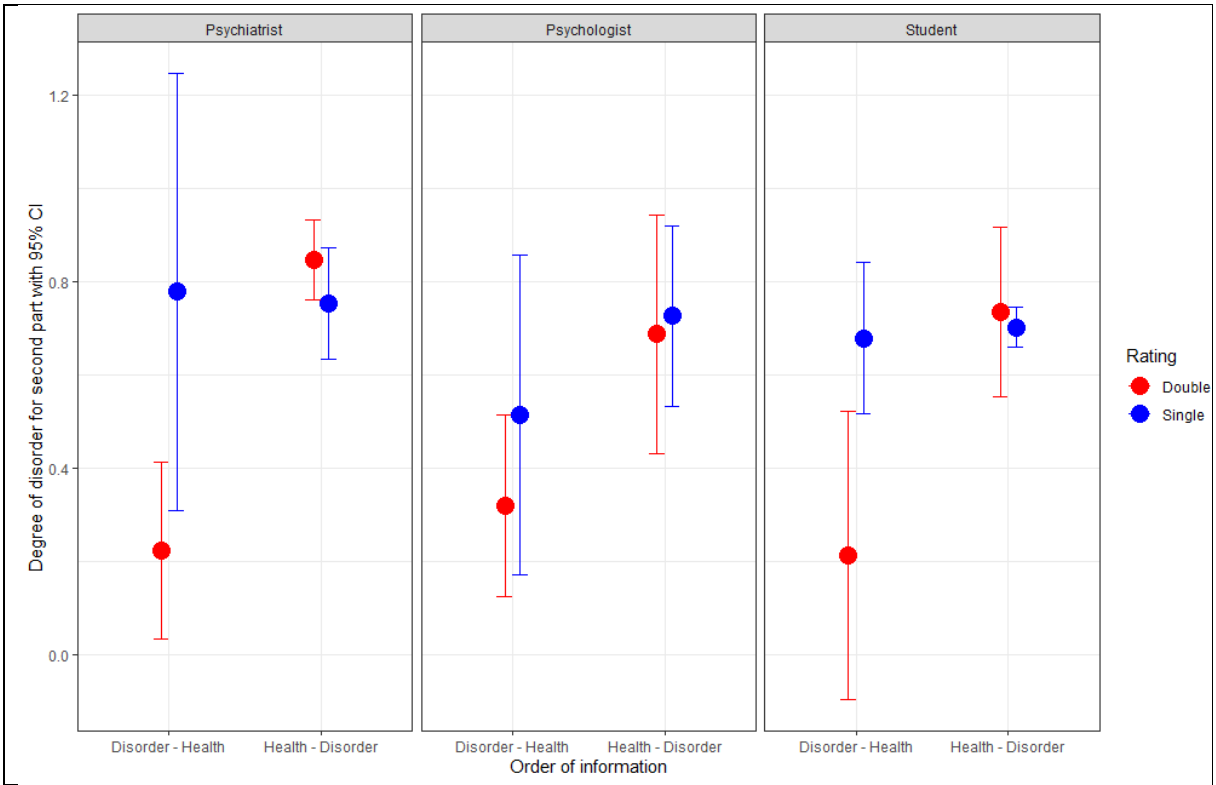
$mean_{doubleratingHD}$=.95 $mean_{single\ rating\ HD}$=.94 $mean_{doubleratingDH}$=.64 $mean_{single\ rating\ DH}$=.68 Inconsistent with prediction in DH condition, intermediate (red) judgment does not produce more health (t(13)=-.19, p=.85). Inconsistent with prediction in HD condition, intermediate (red) judgment does not produce more disorder (t(19)=.21, p=.84).	$mean_{doubleratingHD}$=.98 $mean_{single\ rating\ HD}$=.80 $mean_{doubleratingDH}$=.34 $mean_{single\ rating\ DH}$=.61 Inconsistent with prediction in DH condition, intermediate (red) judgment does not produce more health (t(16)=-1.52, p=.15); in HD conditions differences consistent with predictions, intermediate judgment produces more disorder (t(17)=2.51, p<.05)	$mean_{doubleratingHD}$=.90 $mean_{single\ rating\ HD}$=.80 $mean_{doubleratingDH}$=.42 $mean_{single\ rating\ DH}$=.80 Consistent with predictions in DH condition, intermediate (red) judgment produces more health (t(12)=-2.59, p<.05); in HD condition ns differences, inconsistent with predictions (t(17)=1.11, p=.28)
Malgorzata		



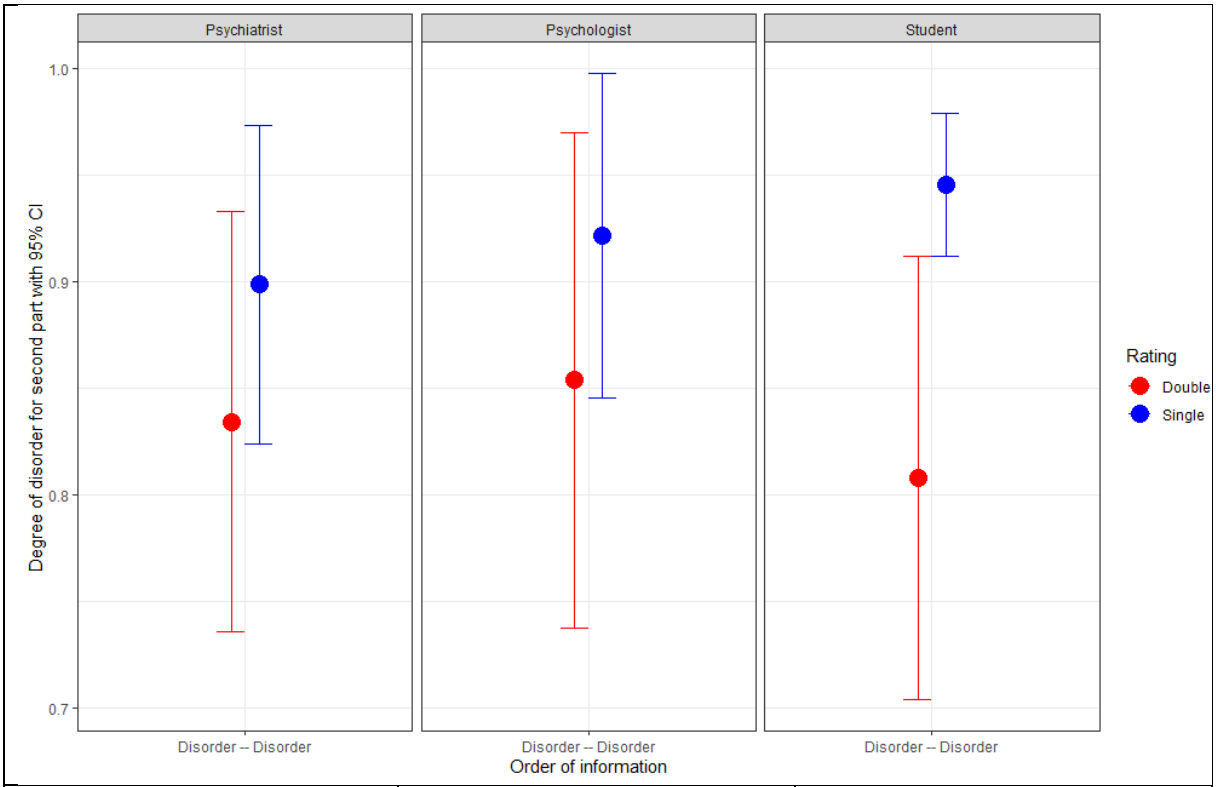
$mean_{doubleratingHD}$=.76 $mean_{single\ rating\ HD}$=.81 $mean_{doubleratingDH}$=.20 $mean_{single\ rating\ DH}$=.70 Consistent with predictions in DH condition, intermediate judgment produces more health(t(18)=-4.25, p<.001); In HD condition ns differences inconsistent with predictions (t(14)=-.38, p=.71)	$mean_{doubleratingHD}$=.45 $mean_{single\ rating\ HD}$=.69 $mean_{doubleratingDH}$=.23 $mean_{single\ rating\ DH}$=.79 Consistent with predictions in DH condition, intermediate judgment produces more health (16)=-3.25, p<.01); In HD condition ns differences Inconsistent with predictions (t(16)=-1.39, p=.18)	$mean_{doubleratingHD}$=.58 $mean_{single\ rating\ HD}$=.72 $mean_{doubleratingDH}$=.34 $mean_{single\ rating\ DH}$=.62 ns differences inconsistent with predictions in HD condition t(16)=-1.02, p=.32; in DH condition t(13)=-1.55, p=.15
Maria		



mean$_{doubleratingHD}$=.93 *mean*$_{single\ rating\ HD}$=.90 *mean*$_{doubleratingDH}$=.20 *mean*$_{single\ rating\ DH}$=.64 Consistent with predictions in DH condition, intermediate judgment produces more health (t(12)=-2.54, p<.05); In HD condition ns differences(t(20)=.41, p=.68) inconsistent with predictions	*mean*$_{doubleratingHD}$=.83 *mean*$_{single\ rating\ HD}$=.86 *mean*$_{doubleratingDH}$=.26 *mean*$_{single\ rating\ DH}$=.66 Consistent with predictions in DH condition, intermediate judgment produces more health t(14)=-2.46, p<.05); In HD condition ns differences inconsistent with predictions (t(19)=-.27, p=.79	*mean*$_{doubleratingHD}$=.93 *mean*$_{single\ rating\ HD}$=.85 *mean*$_{doubleratingDH}$=.33 *mean*$_{single\ rating\ DH}$=.77 Consistent with predictions in DH condition, intermediate judgment produces more health t(13)=-3.18, p<.01; In HD condition ns differences inconsistent with predictions (t(16)=1.15, p=.27
Marta		



$mean_{doubleratingHD}$=.85 $mean_{single\ rating\ HD}$=.75 $mean_{doubleratingDH}$=.22 $mean_{single\ rating\ DH}$=.78 Consistent with predictions in DH condition, intermediate judgment produces more health(t(13)=-3.33, p<.01); In HD condition ns differences inconsistent with predictions (t(19)=1.48, p=.15)	$mean_{doubleratingHD}$=.69 $mean_{single\ rating\ HD}$=.73 $mean_{doubleratingDH}$=.32 $mean_{single\ rating\ DH}$=.51 ns differences inconsistent with predictions $t_{DH}(14)$=-1.20, p=.24; $t_{HD}(14)$=-.27, p=.79;	$mean_{doubleratingHD}$=.73 $mean_{single\ rating\ HD}$=.70 $mean_{doubleratingDH}$=.21 $mean_{single\ rating\ DH}$=.68 Consistent with predictions in DH condition, intermediate judgment produces more health (t(13)=-3.57, p<.01; In HD condition ns differences inconsistent with predictions (t(16)=.27, p=.79)
Oliwia		



$mean_{\text{double rating}}$=.83 $mean_{\text{single rating}}$ =.90 t(34)=-.98, p=.34	$mean_{\text{double rating}}$=.85 $mean_{\text{single rating}}$ =.92 t(35)=-1.07, p=.29	$mean_{\text{double rating}}$=.81 $mean_{\text{single rating}}$ =.94 t(31)=-3,19, p<.01

Appendix 3

We summarize the mathematical formulation for the basic quantum model, the noisy quantum model, and the belief adjustment model, for the evaluation bias. The basic quantum model and the belief adjudgment model for the evaluation bias are presented in detail in White et al. (2020), so here we offer only a description of the main technical tools and the equations. The noisy quantum model is a novel proposal and we show the algebra in more detail. We do not offer a detailed tutorial on quantum methods, but instead refer readers to Busemeyer and Bruza (2012) or Pothos and Busemeyer (2013) and references therein.

Regarding the basic quantum model, recall from main text and Figure 4 that there is an initial state vector, a rotation, and a projection. Projections are modeled using so-called projection operators, P_x, which are linear operators. Then, the projection of state ψ onto subspace X is given by $P_x\psi$, which is just a vector along subspace X. Also, probabilities are computed by squaring the length of these projections, e.g., $Prob(X;\psi) = |P_x\psi|^2$ (the vertical bars indicate length). Note, this rule relating probabilities to subspaces is a key assumption distinguishing quantum theory from projective linear algebra. In the real spaces we employ here, rotations are implemented with just rotation operators, e.g., a clockwise rotation of vector ψ by angle θ is given by $U(\theta)\cdot\psi$, where $U(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$. That is, $U(\theta)\cdot\psi$ is a vector rotated by θ radians from ψ, in a clockwise direction. In quantum theory, the operator $U(\theta)$ is an example of a so-called unitary operator, whose main properties are $\langle \widehat{U}\psi | \widehat{U}\varphi \rangle = \langle \psi | \varphi \rangle$, so that $\widehat{U}^{-1} = \widehat{U}^{\dagger}$ (the 'dagger' indicates the complex conjugate of the transpose). Given these simple tools, it is straightforward to specify the basic quantum model.

Consider the disorder-health condition. Define a normalized vector along the disorder ray, $Perfect_D$. Recall that in the disorder-health condition the initial mental state vector is assumed to be close to the disorder ray (since the initial information is for disorder the mental

state is set in a way to indicate a high probability for disorder). Note, the initial mental state vector could be placed either on the left or the right of the disorder ray. The particular placement can be justified in terms of what happens if we consider the impact of a series of successive judgments (White et al., 2020). Then, we can write $\psi_D = U(-rating) \cdot Perfect_D$, which just tells us that ψ_D can be identified as a vector rotated in an anticlockwise way, by *rating* degrees, from $Perfect_D$ (this is just a technical convenience in how to define the initial state, that is, as a rotation from a state of known position). In main text, we noted that introducing the health information would lead to a rotation towards the health ray and, in Figure 4, this rotation was shown as a rotation in a clockwise direction. The direction of rotation is simply the same as the direction of a projection, if a judgment were to be made (White et al., 2020). For example, in disorder-health cases, given that the state vector is placed in the top left quadrant, a projection would require collapse in a clockwise direction. The equations for the basic quantum model essentially follow:

$$Prob(FSDR; \psi_D) = |P_{Disorder} \cdot \psi_D|^2 \quad (1)$$

$$Prob(SSSR; \psi_D) = |P_{Disorder} \cdot U_H(n) \cdot \psi_D|^2 \quad (2)$$

$$Prob(SSDR; \psi_D) = |P_{Disorder} \cdot U_H(n) \cdot Perfect_D|^2 \quad (3)$$

The subscript H in the case of $U_H(n)$ indicates that the rotation corresponds to the introduction of information that the patient is healthy. These equations have two free parameters, *n* and *rating*, the latter required for the specification of $\psi_D = U(-rating) \cdot Perfect_D$. A few notes are necessary. First, these equations assume that in the disorder-health condition, the first judgment, if asked for, would indicate disorder. Second, if the information is health-disorder, then in Equation (2) $n \rightarrow -n$ and in Equation (3) the state is $Perfect_H$, since we now assume that the first decision indicated health. Third, the reason why in all cases the final projector is $P_{Disorder}$ is that all the participant data correspond to how likely it

is that the subject is suffering from an eating disorder. Therefore, the model likewise produces a probability for an eating disorder. Fourth, the impact of the second piece of information on the mental state is the same, regardless of whether the mental state is the initial one (ψ_D) or the one after a decision for the first stimulus (if the first decision indicates disorder, then this would be $Perfect_D$).

Regarding the noisy quantum model, even though conceptually the introduction of noise (in the form of a mismatch between judgment and change in mental state vector) is straightforward, there are various technical elaborations, which are less straightforward. First, instead of employing a mental vector state ψ, we have to employ a so-called density operator ρ. A density operator ρ can be set up to closely correspond to a particular mental state vector, and this is the approach in the present work, that is, e.g. ρ_D is specified in a way closely analogous to ψ_D. Second, the probability rule is different, instead of $Prob(X;\psi) = |P_x\psi|^2$, we have $Prob(X;\rho) = Tr(P_x\rho)$, where Tr indicates the Trace operator, which sums matrix elements along the diagonal. For example, the trace of matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is just *a*+*d*. Third, instead of projectors P_x, we employ positive operator valued measures (POVMs). A POVM is just like a projector, but for the fact that there is a small probability that there will be a mismatch between answer and response. Note that a POVM can be decomposed into two measurement operators, so that $P_{POVM} = M_P \cdot M_P^\dagger$ (the † indicates conjugate transpose of a matrix; in the present model $M_P = M_P^\dagger$). The measurement operators are necessary for specifying the state post-measurement.

Is it not possible to have a POVM model with mental state vectors, as for the basic quantum model? It is possible. However, depending on the POVM, an initial so-called pure state (that is, one that can be represented by a state vector), might not stay that way under measurements. That is, measuring with a POVM might turn a pure state into one that has to be

represented with a density operator (a mixed state). So, when using POVMs, it is more appropriate to employ density operators.

The first two equations for the noisy quantum model follow:

$$Prob(FSDR;\rho_D) = Tr\left(P_{Disorder,POVM}\cdot\rho_D\right)\text{.....(4)}$$

$$Prob(SSSR;\rho_D) = Tr\left(P_{Disorder,POVM}\cdot U(n)\cdot\rho_D\cdot U^{\dagger}(n)\right)\text{.....(5)}$$

For SSDR, we first need to identify the state following the judgment for the first piece of information. The collapse postulate in quantum theory works analogously for state vectors and density operators. Assuming the answer following the first piece of information is disorder, we have:

$$\rho_{FSDR,D\ answer} = \frac{M_{Disorder,POVM}\cdot\rho_D\cdot M^{\dagger}_{Disorder,POVM}}{Tr\left(P_{Disorder,POVM}\cdot\rho_D\right)} = \frac{M_{Disorder,POVM}\cdot\rho_D\cdot M_{Disorder,POVM}}{Prob(FSDR,D;\rho_D)}\text{.....(6)}$$

$$Prob\left(SSDR;\rho_{FSDR,D\ answer}\right) = Tr\left(P_{Disorder,POVM}\cdot U(n)\cdot\rho_{FSDR,D\ answer}\cdot U^{\dagger}(n)\right)\text{.....(7)}$$

To make equations 4-7 intelligible, we need to present in more detail the various elements. The main components of the model are as follows:

$\rho_{initial} = \begin{pmatrix} rating & 0 \\ 0 & 1-rating \end{pmatrix}$, $P_{Disorder} = \begin{pmatrix} \epsilon & 0 \\ 0 & 1-\epsilon \end{pmatrix}$, $P_{Health} = \begin{pmatrix} 1-\epsilon & 0 \\ 0 & \epsilon \end{pmatrix}$ and the corresponding measurement operators are $M_{Disorder} = M^{\dagger}_{Disorder} = \begin{pmatrix} \sqrt{\epsilon} & 0 \\ 0 & \sqrt{1-\epsilon} \end{pmatrix}$ and $M_{Health} = M^{\dagger}_{Health} = \begin{pmatrix} \sqrt{1-\epsilon} & 0 \\ 0 & \sqrt{\epsilon} \end{pmatrix}$. Note, $\rho_{initial}$ can be ρ_D if the first stimulus indicates disorder or ρ_H if health. Also, we will skip noting POVM in projector names, and just assume this in what follows. The parameter ϵ is the error rate for mismatching answers and projections. Then, for each of the three possible ratings in the experiment, we can write:

FSDR

$$Prob(FSDR = Disorder; \rho_{initial}) = Tr(P_{Disorder}\rho_{initial}) =$$

$$Tr\left(\begin{pmatrix}\epsilon & 0\\ 0 & 1-\epsilon\end{pmatrix}\begin{pmatrix}rating & 0\\ 0 & 1-rating\end{pmatrix}\right) = Tr\begin{pmatrix}\epsilon\cdot rating & 0\\ 0 & (1-\epsilon)\cdot(1-rating)\end{pmatrix} = \epsilon\cdot$$

$$rating + 1 - rating - \epsilon + \epsilon\cdot rating = 1 - (rating - 2\cdot\epsilon\cdot rating + \epsilon) = 1 -$$

$$Prob(FSDR = Health; \rho_{initial}).$$

SSSR

For the case of disorder-health, we have:

$$Prob(SSSR, P; \rho_{initial}) = Tr\left(P_P U(n)\rho_{initial}U^{\dagger}(n)\right) =$$

$$= Tr\left(\begin{pmatrix}\epsilon & 0\\ 0 & 1-\epsilon\end{pmatrix}\begin{pmatrix}\cos n & \sin n\\ -\sin n & \cos n\end{pmatrix}\begin{pmatrix}rating & 0\\ 0 & 1-rating\end{pmatrix}\begin{pmatrix}\cos n & -\sin n\\ \sin n & \cos n\end{pmatrix}\right)$$

$$= Tr\left(\begin{pmatrix}1-\epsilon & 0\\ 0 & \epsilon\end{pmatrix}\begin{pmatrix}\cos n\cdot rating & \sin n\cdot(1-rating)\\ -\sin n\cdot rating & \cos n\cdot(1-rating)\end{pmatrix}\begin{pmatrix}\cos n & -\sin n\\ \sin n & \cos n\end{pmatrix}\right) =$$

$$= Tr\left(\begin{pmatrix}(1-\epsilon)\cdot\cos n\cdot rating & (1-\epsilon)\cdot\sin n\cdot(1-rating)\\ -\epsilon\cdot\sin n\cdot rating & \epsilon\cdot\cos n\cdot(1-rating)\end{pmatrix}\begin{pmatrix}\cos n & -\sin n\\ \sin n & \cos n\end{pmatrix}\right)$$

$$= Tr\begin{pmatrix}(1-\epsilon)\cdot\left(\cos^2 n\cdot rating + \sin^2 n\cdot(1-rating)\right) & \dots\\ \dots & \epsilon\cdot\left(\sin^2 n\cdot rating + \cos^2 n\cdot(1-rating)\right)\end{pmatrix}$$

$$= (1-\epsilon)\cdot\left(\cos^2 n\cdot rating + \sin^2 n\cdot(1-rating)\right) + \epsilon$$

$$\cdot\left(\sin^2 n\cdot rating + \cos^2 n\cdot(1-rating)\right)$$

That is, for the case of health-disorder, we need only assume a rotation of -*n*, instead of *n*.

SSDR

We first need to identify the state after the first stimulus rating is in the disorder-health condition:

$$\rho_{FSDR,D\ answer} = \frac{M_D\rho_{initial}M_D^{\dagger}}{Tr(P_D\rho_{initial})} = \frac{M_D\rho_{initial}M_D}{Prob(FSDR, D; \rho_{initial})}$$

$$M_D \rho_{initial} M_D = \begin{pmatrix} \sqrt{\epsilon} & 0 \\ 0 & \sqrt{1-\epsilon} \end{pmatrix} \begin{pmatrix} rating & 0 \\ 0 & 1-rating \end{pmatrix} \begin{pmatrix} \sqrt{\epsilon} & 0 \\ 0 & \sqrt{1-\epsilon} \end{pmatrix}$$

$$= \begin{pmatrix} rating \cdot \epsilon & 0 \\ 0 & (1-rating) \cdot (1-\epsilon) \end{pmatrix}$$

So, $\rho_{FSDR,D\ answer} = \begin{pmatrix} rating \cdot \epsilon & 0 \\ 0 & (1-rating) \cdot (1-\epsilon) \end{pmatrix} \cdot \frac{1}{Prob(FSDR,D;\rho_{initial})}$

Similar algebra shows that:

$$\rho_{FSDR,H\ answer} = \begin{pmatrix} rating \cdot (1-\epsilon) & 0 \\ 0 & (1-rating) \cdot \epsilon \end{pmatrix} \cdot \frac{1}{Prob(FSDR,H;\rho_{initial})}$$

To summarize so far, if FSDR indicates a disorder answer, then the state is $\rho_{FSDR,D\ answer}$. If FSDR indicates a health answer, then the state is $\rho_{FSDR,H\ answer}$. These states take into account the possibility of mismatch between response and projection. Then,

$$Prob(SSDR; \rho_{FSDR,D\ answer}) = \frac{1}{Prob(FSDR,D;\rho_{initial})} Tr\left(P_P U(n) \begin{pmatrix} rating \cdot \epsilon & 0 \\ 0 & (1-rating) \cdot (1-\epsilon) \end{pmatrix} \cdot U^{\dagger}(n) \right) =$$

$$\frac{1}{Prob(FSDR,DP;\rho_{initial})} Tr\left(\begin{pmatrix} \epsilon & 0 \\ 0 & 1-\epsilon \end{pmatrix} \begin{pmatrix} \cos n & \sin n \\ -\sin n & \cos n \end{pmatrix} \begin{pmatrix} rating \cdot \epsilon & 0 \\ 0 & (1-rating) \cdot (1-\epsilon) \end{pmatrix} \begin{pmatrix} \cos n & -\sin n \\ \sin n & \cos n \end{pmatrix} \right) =$$

$$\frac{1}{Prob(FSDR,DP;\rho_{initial})} Tr\left(\begin{pmatrix} \cos n \cdot \epsilon & \sin n \cdot \epsilon \\ -\sin n\,(1-\epsilon) & \cos n\,(1-\epsilon) \end{pmatrix} \begin{pmatrix} rating \cdot \epsilon \cos n & rating \cdot \epsilon(-\sin n) \\ (1-rating) \cdot (1-\epsilon) \sin n & (1-rating) \cdot (1-\epsilon) \cos n \end{pmatrix} \right) =$$

$$\frac{1}{Prob(FSDR,D;\rho_{initial})} Tr\begin{pmatrix} \cos^2 n \cdot \epsilon^2 \cdot rating + (1-\epsilon)\epsilon \sin^2 n\,(1-rating) & \dots \\ \dots & \sin^2 n\, rating \cdot \epsilon(1-\epsilon) + \cos^2 n\,(1-rating) \cdot (1-\epsilon)^2 \end{pmatrix}$$

$$= \frac{\cos^2 n \cdot \epsilon^2 \cdot rating + (1-\epsilon)\epsilon \sin^2 n(1-rating) + \sin^2 n\, rating \cdot \epsilon(1-\epsilon) + \cos^2 n(1-rating) \cdot (1-\epsilon)^2}{Prob(FSDR,D;\rho_{initial})}$$

Similar algebra shows that

$$Prob(SSDR; \rho_{FSDR,H\ answer})$$

$$= \frac{(1-\epsilon)^2 \cos^2 n\, rating + (1-rating) \cdot \epsilon(1-\epsilon) \sin^2 n + (1-\epsilon)\sin^2 n\, rating \cdot \epsilon + \epsilon(1-\epsilon) \cos^2 n\,(1-rating)}{Prob(FSDR,H;\rho_{initial})}$$

This completes the presentation of the noisy quantum model. The superficially complex algebra should not obscure the fact that essentially all we have done is replace state vectors with density operators and projector operators with POVMs.

We next consider the application of Hogarth and Einhorn's (1992) belief adjustment model to the present data. The model is a powerful framework for describing the way an overall judgment is formulated as a result of a sequence of pieces of evidence. However, it involves several parameters and choices in equations which make it difficult to apply it to the simple case of the evaluation bias paradigm. As noted in text, we follow the approach of White et al. (2020), who offered a reasonable series of restrictions to the model, so it can be applied more or less on equal footing to the quantum models.

When considering a set of *k* statements, parameter S_k ($0 \leq S_k \leq 1$) corresponds to the current evaluation after the k^{th} piece of evidence, parameter $s(x_k)$ to the evaluation for just the k^{th} piece of evidence, and $s(x_1, \ldots, x_k)$ to the evaluation of statements 1 to *k* together; in this case, $0 \leq s(x_k) \leq 1$. Each piece of evidence is assessed against a reference point R and there is a weight which determines how the assessment of the k^{th} piece of evidence informs the current evaluation. These parameters are related by the main equation in the model:

$$S_k = S_{k-1} + w_k[s(x_k) - R] \quad (8)$$

For an SbS process, Hogarth and Einhorn (1992) proposed different versions of the model, depending on e.g. whether the rating scale is unipolar or bipolar. Given the present paradigm adopted unipolar rating scales, the reference parameter R can be set as $R = S_{k-1}$. The updating weight, w_k, depends on whether the current piece of evidence is evaluated more highly relative to the reference point or less highly, and two separate sensitivity parameters, which in the present case can be, fairly safely, set to 1. Putting all these assumptions together (see White et al., 2020, for details), the SbS process for the evaluation bias paradigm involves the equations:

$S_{DH} = (1 - S_D)S_D + S_D s(x_H)$.....(9a)
$S_D = S_0^2 + (1 - S_0)s(x_D)$.....(9b)
and

$S_{HD} = S_H^2 + (1 - S_H)s(x_D)$.....(10a)
$S_H = (1 - S_0)S_0 + S_0 s(x_H)$.....(10b)

Some explanatory remarks are necessary. The subscripts follow the convention in the rest of this paper, so that H indicates health and D disorder. S_0 is the initial evaluation of the participant, following the presentation of the first piece of information; this is a free parameter of the model, analogous to the *rating* parameter in the quantum models. S_D and S_H are the participant evaluations after the first piece of information. $s(x_H)$ and $s(x_D)$ are the participant evaluations after observing information indicating health and disorder, respectively. So, for example, Equation (9b) tells us that the participant's initial state S_0 is adjusted by the presence of disorder information, $s(x_D)$, and leads to new state S_D -- the subscript D in S_D indicates our expectation that this belief state would indicate disorder. We set $s(x_P) = 1 - s(x_N)$ so that $s(x_P)$ is the second parameter of the SbS part of Hogarth and Einhorn's (1992) model.

The EoS part is analogously specified, but instead of there being multiple evaluation steps, there is a single one, $S_k = S_0 + w_k[s(x_1, \ldots, x_k) - R]$. We can set parameters R and w_k using the same kind of reasoning as for the SbS process. The end result is summarized in the following equations:

$S_{DH} = S_0 + S_0[s(x_P, x_N) - S_0]$.....(11a)

$S_{HD} = S_0 + (1 - S_0)[s(x_H, x_D) - S_0]$.....(11b)

Parameters $s(x_D, x_H)$, $s(x_H, x_D)$ are the overall evaluations of two pieces of information provided, in the disorder, health order (first parameter) or the health, disorder order (second parameter). Because we wanted to fit the models separately for each triplet of data points, FSDR, SSDR, SSSR, as things stand Hogarth and Einhorn's (1992) model is overparameterized; there are four parameters, S_0, $s(x_D)$, $s(x_D, x_H)$, $s(x_H, x_D)$. One approach to

avoid this problem is to set $s(x_D, x_H) = s(x_H)$ and analogously for $s(x_H, x_D)$, as White et al. (2020) did, which introduces a bias for recency in the model. In the present dataset, this is a reasonable assumption. This completes the summary of the belief adjustment model.

Appendix 4

In this section we consider the details of fitting the three models. Because the paradigm involved highly distinctive case studies of eating disorder patients and because it was impractical to create too many different cases, each participant provided a handful of judgments and in no instance did we have a triplet of FSDR, SSDR, and SSSR judgments, for the same patient case, from the same participant. Therefore, we fitted the models using item-based analyses. There were four triplets corresponding to disorder-health cases and two to disorder-disorder cases, which were considered together (we will refer to these as disorder-health); there were four triplets corresponding to health-disorder cases or health-health cases, which were also considered together (referred to as health-disorder).

As noted in main text, each triplet is three datapoints {FSDR, SSDR, SSSR}. For each triplet separately, for the basic quantum model, we used the FSDR and SSSR judgments to determine the two parameters of *rating* and *n*. Given these two parameters, the model's prediction for SSDR follows. In the case of the basic quantum model, the two parameters were extracted using the FindRoot function in Mathematica – one parameter was extracted using the FSDR rating and the other using the SSSR rating (in both cases, one question for one unknown). Given the two parameters, we used the equation for SSDR to obtain the prediction from the quantum model. In the case of the noisy quantum model, parameter extraction was the same as for the basic quantum model, but we employed a simple grid search regarding the noise parameter; for each possible value of the noise parameter (in steps of 0.1), we computed best parameters and fit, and retained the solution with best fit. In the case of the belief adjustment model, we extracted the two model parameters using the FindMinimum function in Mathematica and a cost term based on sums of squares – even though here too we have two equations with two unknowns, we could not always identify

parameter values exactly satisfying them (as for the quantum model), which is why FindMinimum had to be employed. As for the other models, given the two parameter values, a prediction for SSDR follows from the belief adjustment model. So, all three models were assessed in a similar way. Note, given the small number of data points (12 triplets) no cross-validation was carried out.

To summarize so far, the success or not of any of the models is determined by comparing the 12 predicted SSDR values against the 12 observed SSDR values. To reiterate, predicted SSDR values correspond to probabilities for disorder and observed SSDR values to suitably scaled ratings for disorder.

For each model, the 12 predicted and observed SSDR values were compared using residual sum of squares (RSS) and $BIC = N \cdot \ln\left(\frac{RSS}{N}\right) + p \cdot \ln(N)$ with N=6 and p=0 vs. 1; lower values of BIC indicate better fit. Note, for the disorder-health condition the noisy quantum model produces a few equivalent solutions (to two decimal places for residual sums of squares) in the vicinity of $\epsilon = 0.5$; we show a solution just off the value $\epsilon = 0.5$, since otherwise all predicted probabilities are at 0.5.

Finally, in main text, Table 3 shows exact values for the 12 predicted and observed SSDR values. Here, we offer a plot which shows the correspondence between predicted and observed SSDR values, Figure A3.1.

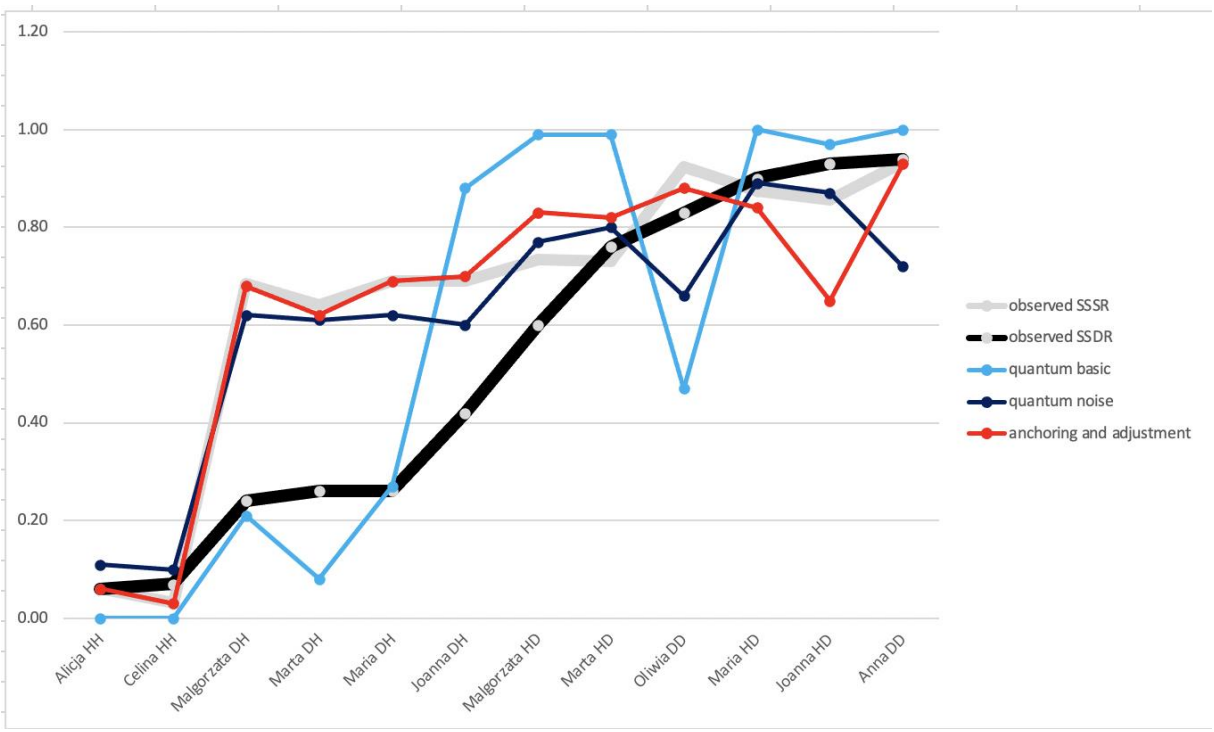


Figure A3.1. The 12 SSDR values are presented in ascending order. Next to each case name we note whether the FSDR judgment is from an HH (health-health), DH (disorder-health), HD (health-disorder), or DD (disorder-disorder) pair.