Primacy, Congruence and Confidence in Diagnostic Decision-Making

Thomas Frotvedt[1], Øystein Keilegavlen Bondevik[1], Vanessa Tamara Seeligmann, Bjørn Sætrevik
Department of psychosocial science, University of Bergen
1: These authors contributed equally to the manuscript

Abstract
Some heuristics and biases are assumed to be universal for human decision-making, and may thus be expected to appear consistently and need to be considered when planning for real-life decision-making. Yet results are mixed when exploring the biases in applied settings, and few studies have attempted to robustly measure the combined impact of various biases during a decision-making process. We performed three pre-registered classroom experiments in which trained medical students read case descriptions and explored follow-up information in order to reach and adjust mental health diagnoses ($\sum N = 224$). We tested whether the order of presenting the symptoms led to a primacy effect, whether there was a congruence bias in selecting follow-up questions, and whether confidence increased during the decision process. Our results showed increased confidence for participants that did not change their decision or sought disconfirming information. There was some indication of a weak congruence bias in selecting follow-up questions. There was no indication of a stronger congruence bias when confidence was high, and there was no support for a primacy effect of the order of symptom presentation. We conclude that the biases are difficult to demonstrate in pre-registered analyses of complex decision-making processes in applied settings.

Highlights:
- Three pre-registered classroom experiments on trained expert decision-makers
- Congruence bias was weakly indicated
- No support for primacy effect or indication that confidence drives the congruence bias

Keywords: Congruence bias, primacy effect, confidence, diagnostics, decision-making

1.1: Decision-making in medical diagnosis

Diagnosing patients involves making decisions under uncertainty. Decision-making is a complex mental process which is vulnerable to errors. Such errors may have grave consequences, as diagnostic decisions inform the treatment, and therefore directly impact the health and well-being of the patients as well as the efficiency of the health-care systems. Studies have indicated (Berner & Graber, 2008; Graber, 2013) that medical professionals make the wrong diagnosis in about 10-15% of the cases. Mental health diagnoses may in particular be "wicked problems" (Hannigan & Coffey, 2011), as the problem space is poorly defined, decision feedback provides poor grounds for learning as treatments give ambivalent feedback, and ground truth may be difficult to establish (Hogarth, Lejarraga, & Soyer, 2015).

Cognitive psychology may be relevant for understanding and preventing errors caused by the clinician's thinking process. Graber, Franklin and Gordon (2005; see also Graber, Gordon, & Franklin, 2002) developed a taxonomy of medical errors which distinguished between no-fault errors, systematic errors and cognitive errors. The latter category has been found to be most frequent (Croskerry, 2009a; Graber, 2005, 2011), and covers faults in data gathering, knowledge, reasoning, and verification. Errors in reasoning due to cognitive biases are an important subcategory.

Heuristics are mental shortcuts that allow clinicians to make quick decisions that tend to be sufficiently accurate when thorough analyses of all possibilities and probabilities are inexpedient or simply impossible (Gigerenzer, 2008; Graber et al., 2002; Ludolph & Schulz, 2018; Todd & Gigerenzer, 2001). Such shortcuts can be necessary and useful in dynamic medical settings. However, heuristics entail systematic biases, that may lead to misinterpretation and oversimplification of information and may limit information search and evaluation (see Crowley et al., 2013; Tversky & Kahneman, 1974). Heuristics in diagnostic reasoning may be understood within a dual process theory (Morewedge & Kahneman, 2010; Pelaccia, Tardif, Triby, & Charlin, 2011), where fast and automatic "system 1" processing of diagnostic information is subject to a number of biases, while a more effortful, analytic "system 2" reasoning is needed to avoid them (see Croskerry, 2009a; see also Croskerry, 2009b; Payne, 2011; van den Berge & Mamede, 2013). Overlearning, dysrationalia override, fatigue, distractions and time constraints may lead to "system 1" being dominant for clinicians (see also Croskerry, 2009b; Norman, 2009; Norman & Eva, 2010). While the degree of uncertainty and errors in diagnostic decisions may vary across medical fields, errors may occur within any specialty (Croskerry, 2003). Decades of research have contributed to understanding how heuristics can lead to errors, and may provide information to how the errors may be reduced. There is sparse research on the use of heuristics in mental health diagnosis, but some studies have indicated some of the same phenomena as in general decision-making. Among the biases that have been found to influence decision-making in medical research are the primacy effect, congruence bias and excessive confidence.

1.1.3: Primacy bias. Research on decision making (Curley, Young, Kingry, & Yates, 1988; Hogarth & Einhorn, 1992; Lund, 1925) has shown that the information presented first tend to have a larger influence on the judgement than information presented later. A sub-type of this mechanisms may be Tversky and Kahneman's (1974) *anchoring bias,* where a belief is established based on initial information, and that belief is not adequately adjusted by subsequent information. Crowley and colleagues (2013) argued that primacy could influence medical diagnoses if the clinician "locks onto" salient symptoms presented early in the diagnostic process, leading to an initial diagnosis that influences the final diagnosis. The exploration of primacy bias in mental health diagnosis has shown inconsistent findings between similar studies (Ellis, Robbins, Schult, Ladany, & Banker, 1990; Friedlander & Stockman, 1983; Richards & Wierzbicki, 1990). Nevertheless, several studies where participants assess written descriptions of mental health cases have shown that diagnoses tend to be compatible with the symptoms presented first (Cunnington, Turnbull, Regher, Marriott, & Norman, 1997; Richards & Wierzbicki, 1990). For instance, Parmley (2006) manipulated the order of symptom presentation in case descriptions, and showed that a third of clinicians failed to alter their initial diagnosis when presented with disconfirming evidence. While this was not Parmley's focus, she predicted (2006, p. 84) that participants would "fail to alter their initial diagnosis even when new information presented at time two has disconfirming evidence". This corresponds with how primacy has been operationalized in studies of general decision-making (Friedlander & Stockman, 1983; Tversky & Kahneman, 1974).

1.1.2: Congruence bias. We tend to seek or interpret information in ways that can corroborate or support our current beliefs, expectations or a hypothesis at hand, while the information that is inconsistent with our beliefs may be ignored or deemphasized (Baron, Beattie, & Hershey, 1988). A sub-type of *congruence bias* may be the traditional confirmation bias (Nickerson, 1998) where new information is interpreted as supporting one's current beliefs. The terms have been used interchangeably in much of the empirical literature (Beattie & Baron, 1988; Xiao, Lan, Piara, & Feldman, 2020). In a diagnostic setting, congruence bias could lead to closing the exploratory phase prematurely; accepting a particular diagnosis before it has been fully verified, or neglecting plausible alternatives (Croskerry, 2002, 2009a; Eva, 2001; Parmley, 2006). Studies of mental health diagnoses have shown congruence bias in selecting which additional information to gather (Martin, 2001; Mendel et al., 2011) and how it is interpreted (Croskerry, 2002; Eva, 2001; Oskamp, 1965; Parmley, 2006). Two of these studies (Martin, 2001; Oskamp, 1965) presented inconclusive symptom information and found the degree of confidence to be associated with the diagnostic judgment. Congruence bias may thus be closely associated with the primacy effect, and may compound the errors that a congruence error causes (Croskerry, 2002, 2003): Clinicians may "lock onto" salient symptoms early in the diagnostic process, which leads them towards a preliminary diagnosis (primacy). Subsequent processes of seeking and interpreting additional information may be biased towards this initial hypothesis, while alternative plausible explanations may be ignored (congruence). Most previous studies of congruence bias in diagnostics start out by indicating an incorrect answer, and examine whether participants are able to find the correct diagnosis when provided with additional information (Mendel et al., 2011). While this approach may make it easier to establish a congruence bias, an approach where participant are allowed to arrive at an initial diagnosis based on ambiguous information may have higher ecological validity and can be used to examine how congruence influences the ongoing decision process.

1.1.3: Overconfidence. Clinicians' confidence in their decision may be affected by the perceived qualitative and quantitative aspects of available information, and on the existence of plausible alternatives (see Eva, 2001; Martin, 2001; Oskamp, 1965). Croskerry (2002) defined *overconfidence* bias as a universal tendency to believe that we know more than we do or place too much faith in opinions rather than evidence. He specified that overconfidence may be augmented by primacy. "Locking onto" salient information early in an encounter with a patient may make the clinician confident that this information is particularly important. In turn this may affect the formation and rigidness of a preliminary diagnosis. Further, overconfidence may in itself lead to diagnostic errors, for instance by leading to unsuited heuristics (Berner & Graber, 2008). Clinicians that feel confident about their diagnostic decision may be more biased in their search and interpretation of additional information (see Martin, 2001; Oskamp, 1965). Confidence may thus influence a clinician's decision process, and may act as both an effect and a cause for the cognitive biases. This may make it difficult to say when the confidence in a decision is "excessive" or "unfounded". In the current approach we will say that an increase in confidence without being provided with additional information that is conclusive for or against a diagnosis indicates "overconfidence".

1.2: Research needs

Cognitive biases like primacy, congruence and overconfidence have been argued to be relevant for diagnostic decisions (Graber et al., 2002), but attempts to examine how biases influence the mental health diagnostics have shown inconsistent results (Ellis et al., 1990; Friedlander & Stockman, 1983; Oskamp, 1965; Parmley, 2006; Richards & Wierzbicki, 1990). There is a need for experiments that control for some of the issues these studies had, like the balance between the severity of the cases and the amount of information provided simultaneously (Richards & Wierzbicki, 1990). Further, few studies have examined how the biases relate to both seeking and interpreting information (with some exceptions, see Martin, 2001; Mendel et al., 2011), and their effects on diagnostic processes. However, there appears to be no controlled investigation of how all three biases may influence a given mental health decision. It would be of value to combine the testing of both primacy and congruence bias in one experimental design, and to model decision-making as a sequential process where information is gathered over time and confidence in a decision is established. Finally, as only a minority of studies use medically trained personnel (34%; Blumenthal-Barby & Krieger, 2015), it would be of value to establish similar effects in a reasonably realistic problem field in which the participants had relevant training.

1.3: Current study

1.3.1: Aims for the current study. The current study seeks to develop an experimental paradigm for testing the interaction of primacy, congruence and confidence on seeking information, evaluating information, and making diagnostic decisions. To achieve this, we designed a basic experimental procedure that measures information gathering, choice of and confidence in decisions. The experiment can measure (1) whether participants prefer diagnoses that match the symptoms presented first, (2) whether participants seek information to corroborate the assumption they currently hold, (3) whether levels of confidence predict corroborating information seeking, and (4) whether diagnosis and confidence changes across the diagnostic process.

1.3.2: Hypotheses. Based on previous research on primacy, we expected that the order in which symptoms are presented will affect the choice of a preliminary diagnosis. Our first hypothesis was thus that (H1) participants will be more likely to select the preliminary diagnosis congruent with the symptoms presented first in a case vignette, rather than selecting the diagnosis congruent with symptoms presented later.

Based on previous research on congruence, we expected that participants would primarily seek out information that appeared to confirm their existing diagnostic beliefs. Our second hypothesis was thus that (H2) participants would request more information related to the diagnosis they already preferred (indicated on the preceding question about diagnostic preference), rather than seeking out information that may support an alternate diagnosis.

Furthermore, we expected such confirmatory styles of information gathering to correspond to higher levels of confidence in one's existing diagnostic beliefs. The third hypothesis was thus that (H3) requests for confirmatory information would be preceded by higher levels of confidence than when dissenting information was requested.

We expected participants who did not change their mind or explore other alternatives would end up with more confidence in their diagnostic decision. The final hypothesis was thus that (H4) participants who prefer the same diagnosis throughout the case exploration and only request confirming information should increase in confidence between their first and final diagnostic decision. This hypothesis can only be tested on participants that show this response pattern. As this applied to only 12, 20 and 21 participants in the three experiments, it is underpowered to be tested individually, but was tested across all three experiments pooled.

1.3.3: Procedure and pre-registration. All four hypotheses were explored in three experiments that were pre-registered ahead of each data collection (https://osf.io/dn4rv/registrations). The experiment procedure was tested in classroom experiments in order to efficiently collect data from participants with medical training. Each experiment was done consecutively, building on the analysis of the preceding experiments to allow for iterative improvements in experiment design, to better control for competing hypotheses, and to further investigate earlier findings. This led to removing some details in case descriptions and making a few changes in the materials for Experiment 2 and 3 to make the manipulation more effective (see more details about changes between experiments below). For further information see the pre-registrations and experiment details at Open Science Framework (https://osf.io/dn4rv/).

2: Methods

2.1: Participants

Participants in the study were advanced medical students with extensive education in of somatic and mental health issues, and were drafted from a university hospital in Norway. Most of the students that were present in three lectures participated, constituting 71, 56 and 91 students respectively for the three experiments. In accordance with the pre-registration, Experiment 3 was completed in two separate data collection sessions, as the initial session provided a lower turnout than required.

Demographic variables were not collected in any of the experiments to preserve a sense of anonymity. However, we judged the student population from which the sample was drawn to be predominantly female (about 75%), and in their mid-twenties. A lottery was conducted immediately after each data collection, in which about 5% of the participants won a gift card for a lunch meal at a campus cafe.

2.2: Experiment overview

Each experiment was conducted in an auditorium during a break in the lecture that the participants attended. The lecturer introduced the experimenters to the classes, and described the project as an investigation of decision-making under uncertainty. The experimenters informed the students that participation was voluntary and anonymous, and that they could withdraw from the investigation at any time without consequences. As the experiment came in the form of an online survey, participation was possible through laptop computers, tablets and smartphones. No personal information such as names or email addresses were collected. The online questionnaires (in Google Forms) were accessed through the university internet connection, causing the IP address to be the same for all participants.

All three experiments had the same overall structure, where two fictitious patient cases were evaluated (see Table 1 for an overview of the steps in evaluating each case). Completion of the experiment took about 10 minutes.

Table 1: Experiment procedure for all tree experiments. All but the last step was presented online.

#	Experiment procedure step
1.	Introduction and description of the experiment procedure
2.	Introduction of the two diagnostic options (A and B), and presentation of the ICD-10 diagnosis criteria for both options
3.	First half of initial case description, randomized to present symptoms either in support of A or B
4.	T1a choice (Experiment 3 only): State confidence in one of the two diagnoses: A or B
5.	Second half of initial case description, presenting the symptom information not presented in step 3 (supporting either B or A)
6.	T1b choice: State confidence in one of the two diagnoses: A or B
7.	T1 request: Select one of four options to further explore one of the diagnoses: A1, A2, B1, or B2
8.	T2 choice: State confidence in one of the two diagnoses: A or B
9.	T2 request: Select one of four options to further explore one of the diagnoses: A1, A2, B1, or B2
10.	T3 choice: State final confidence in one of the two diagnoses: A or B
11.	End of the first case, repeat 2-10 for the second case where A and B are replaced with C and D
12.	Four follow-up questions about what the participant assumed the research hypotheses to be (in Experiment 3 only)
13.	Brief oral debrief at the end of experiment

2.3: Materials and experiment procedure

We developed text descriptions of two hypothetical patient cases, both featuring symptom information that could be read to support one of two mental health diagnoses as listed in the ICD-10 diagnosis manual. The cases were developed similarly as those used by Parmley (2006), though considerably shorter. The materials were verified as clinically relevant for mental health diagnosis by a clinical psychologist, and the experiment procedure was evaluated as a relevant test of a diagnostic approach by a medical doctor in charge of the medical training. More detailed descriptions of the experiment procedures and text descriptions of the cases in the original Norwegian and translated to English are available in the pre-registrations for each of the experiments (https://osf.io/dn4rv/registrations). The first case presented a choice between (A) dementia and (B) depressive episode, while the second case presented a choice between (C) bipolar mood disorder and (D) borderline personality disorder. The participants were instructed to base their decisions on these criteria, rather than any prior knowledge they had about the diagnoses in question.

Except for a few minor changes, the material remained unchanged for all three experiments, and all participants received the same cases in the same order (see Table 1 above). The two diagnostic options and the corresponding ICD criteria were presented before each of the case descriptions (stage 2 in Table 1). Each case consisted of an initial vignette describing a hypothetical patient. This vignette first (stage 3) presented all the symptoms that supported one of the diagnoses, and then (stage 5) presented all the symptoms supporting the opposite diagnosis.[1] In addition, "neutral" symptoms that are compatible with both diagnoses were included to avoid the contrast between the other pieces of information becoming too obvious. The full case descriptions are available as supplemental materials online (https://osf.io/dn4rv). The initial case descriptions were intended to present equally persuasive arguments for both of the diagnoses, without conclusively supporting either of them.[2] [3] Participants were subsequently (stage 6) asked to decide on a tentative diagnosis (instructed to "*select a diagnosis and indicate your confidence in it*"). The response was made by clicking on a 10-point scale on which the extreme ends represented the highest degree of certainty for the case's two diagnoses. In Experiment 3, participants also had to answer the same question when they were halfway through the initial case description (stage 4), after only symptoms supporting one of the diagnoses had been presented. This was done as a manipulation check for whether the first half of the symptoms led to a compatible decision, and to check whether being forced to make an

[1] Experiment 1 and 3 presented two symptoms for each of the diagnoses. In an attempt to make the manipulation stronger, Experiment 2 presented two symptoms for the first diagnosis and only one for the other diagnosis. Preceding experiments were also used to adjust the wording of the cases in order to make them be selected equally often.

[2] Figure 1 illustrates the extent to which this was successful. When Experiment 1 indicated a preference for one diagnosis independent of symptom order, the case description for Experiment 2 and 3 were edited to adjust for this.

[3] In order to counterbalance for effects of the order that the diagnoses were listed in (rather than the intended effect of the order of presenting the symptoms), two different versions of each experiment were made, with the two diagnoses for each case in the order ABCD and BADC. Approximately half of the class (assigned by seating) was asked to answer one form, while the other half was asked to answer the other form. This counterbalancing variation was collapsed in the analysis, as the operationalization of responses only counted whether the selected diagnosis matched the symptoms presented first or second.

early decision in a given direction would enhance a congruence bias.

After indicating their initial diagnosis (stage 7), the participants were asked to select one of four options for getting more information about the symptoms (such as request A2 "*Reduced language skills may indicate dementia. Ask the patient about her language use and verbal skills.*"). Two of the options were explicitly marked as seeking more information about symptoms for one of the diagnoses, while the remaining two were marked as seeking information about the other diagnosis. After selecting an option (stage 8), participants received additional information (between 33 and 80 words) relevant to the diagnosis they had selected, but worded in a way that did not conclusively point to either of the diagnoses (such as "*[The patient] has thought of herself as polite and articulated but has recently been told by her family that she can be wicked, vulgar and condescending. [The patient] says that this only happens when she talks about topics that upset her.*"). After receiving the follow-up information, the participants were again instructed to indicate their confidence in either diagnosis, by responding on the same 10-point scale as in stage 6. This was followed (stage 9) by a second opportunity to seek follow-up information, choosing among the same four options as before.[4] After receiving the second follow-up information, participants were then (stage 10) asked to set their final diagnosis in the same way as before. Stage 2-10 were then repeated for the second case in the experiment.

For Experiment 3, four follow-up questions were included at the end of the questionnaire (stage 12). The first two explored the participants' thoughts about the aim of our study, while the latter two asked about the strategy they had used in their decision-making. The aim of these questions was to check whether participants may have guessed the research hypotheses, and whether this had affected their responses. Additionally, we wanted to explore whether the participants were aware of their own decision-making strategy. Participant answers were scored separately by 3 coders and compared for inter-rater reliability. The raters initially scored 1,8% of the participants differently, which were resolved by discussion.

After completing the online experiment (stage 13) the participants were quickly debriefed about the purpose of the experiment, any questions the participants had were answered in plenary, and the gift cards were distributed. Due to time constraints only a short debrief was given verbally in class, while a more thorough debrief and results summary was sent on email). Data preparation was done in a Google Sheet synchronized to OSF, to provide transparency and version history for all data transformations, and can show the calculation of the indices described below. All statistical analyses were performed in RStudio. Due to our pre-registered directed hypotheses, we used one-tailed tests, with a standard alpha cut-off of $p < .0$

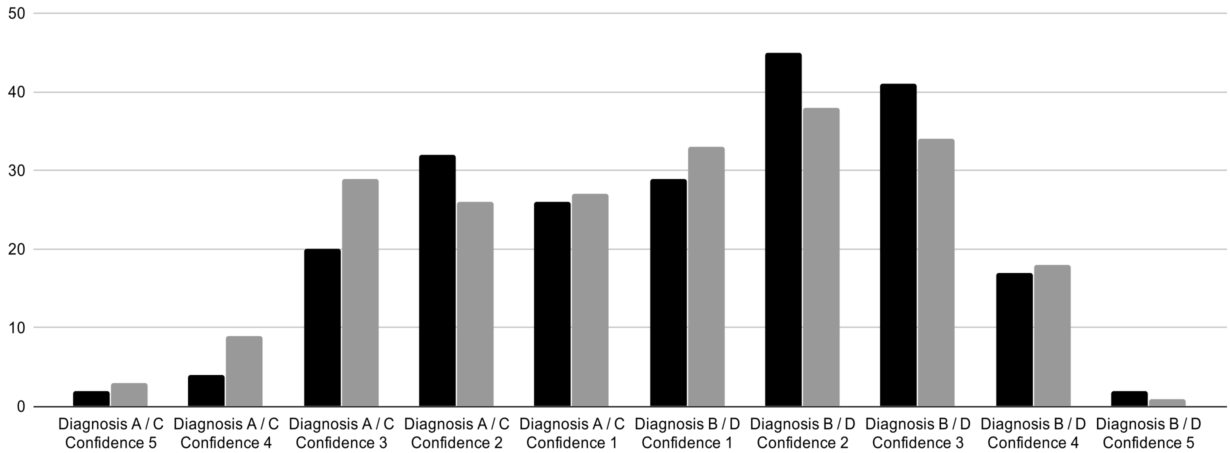

Figure 1: Distribution of answers on the response scale indicating the distribution of confidence in diagnosis for case 1 (black) and case 2 (gray) across all three experiments. As we attempted to balance case descriptions based on previous experiments, the responses were even closer to a normal distribution in later experiments (see figures online).

3: Results

3.1: Tests of primacy

Hypothesis H1 stated that primacy would lead to the information presented first having a larger effect on the decision, despite the fact that equal evidence was presented for both diagnoses. An index was created to reflect whether the initial diagnosis (at T1b) matched the symptoms presented first in the descriptions for each of the two cases (thus having a value of 0, 1 or 2). The null-hypothesis of no effect of symptom order would predict that by chance participants select the diagnosis matching the symptom

[4] It was possible to request the same follow-up information twice, but this only happened for 1.82% of the cases across all participants. This indicates that thefar majority of participants were invested in and attentive to the task.

presented first in one of the two cases. H1 was thus assessed with a one-sample t-test for whether the diagnosis matched the first presented symptoms for more than one case.

3.1.1: Test of H1 in Experiment 1. For Experiment 1 the number of initial diagnoses matching the first symptoms (*M* = 1, *SD* = .74) was identical to the reference constant of 1, thus failing to show a significant difference ($t(70) = 0$, p = .5 one-tailed).

3.1.2: Test of H1 in Experiment 2. Similarly for Experiment 2 the number of initial diagnoses matching the first symptoms (*M* = 1, *SD* = 0.71) was identical to the reference constant of 1, thus failing to show a significant difference ($t(55) = 0$, p = .5 one-tailed).

3.1.3: Test of H1 in Experiment 3. Experiment 3 included a manipulation check by asking participants to make an additional preliminary diagnosis after reading the first half of the symptoms in each case (at T1a). This was done in order to test whether participants in fact indicated the diagnosis supported by the only symptoms they had read so far, and to test whether forcing them to make a decision at this point could lead to primacy or congruence bias (on the T1b question). The number of diagnoses at T1a matching the first symptoms (*M* = 1.53, *SD* = 0.60) was higher than the reference constant of 1, indicating that the participants were consistent with the symptom information presented so far.

As in Experiment 1 and 2, the H1 test of primacy in Experiment 3 used responses from the diagnostic decision made after hearing the full initial case description (T1b). The number of diagnoses matching the first symptoms (*M* = 0.63, *SD* = 0.69) was lower than the constant of 1, and thus not significant in one-tailed testing against a higher value than 1 ($t(90) = -5$, $p = 1$).

3.1.4: Test of H1 across all three experiments. In an exploratory follow-up analysis of data from all three experiments was collapsed to make a more robust assessment of H1. The test remained non-significant despite higher power ($t(217) = -3.06$, p = .99 one-tailed).

3.2: Tests of congruence bias

Hypothesis H2 stated that a congruence bias would lead participants to seek out information that could support the diagnosis they already preferred. An index was created to reflect the number of times the participants requested information that could support the diagnosis they had preferred on the preceding diagnosis question (on T1b and T2 for both cases). The null-hypothesis of no confirmatory information seeking predicted that participants would be equally likely to investigate the preferred diagnosis as the alternate diagnosis, and would thus seek “confirmatory” information at two of the possible four opportunities across the two cases. H2 was thus assessed with a one-sample t-test of whether participants made more than two confirmatory requests.

3.2.1: Test of H2 in Experiment 1. For Experiment 1, the average number of requests for confirming information (*M* = 1.93, *SD* = 0.76) was slightly lower than the reference constant of 2, and was thus not significantly higher than the referent ($t(70) = -0.78$, p = .78 one-tailed).

3.2.2: Test of H2 in Experiment 2. For Experiment 2, the average number of requests for confirming information (*M* = 2.23, *SD* = 1.04) was higher than the reference constant of 2, a difference barely failing to meet our criteria for significance ($t(55) = 1.66$, p = .051 one-tailed).

3.2.3: Test of H2 in Experiment 3. For Experiment 3, the average number of requests for confirming information (*M* = 2.16, *SD* = 0.95) was somewhat higher than the reference constant of 2, again barely failing to meet our criteria for significance ($t(90) = 1.64$, p = .052 one-tailed).

3.2.4: Test of H2 across all three experiments. When collapsing participants across all three experiments, the test for H2 was significant ($t(217) = 1.68$, p = .047 one-tailed). It should be noted that the effect was small ($d = 0.11$), reflecting that there were on average 2.11 (of 4 possible, *SD* = 0.93) cases of seeking confirmatory information.

3.3: Tests of confidence leading to congruence

Hypothesis H3 stated that higher confidence in the diagnostic choice should lead to more confirmatory information seeking. We compared the average confidence rating (a value between 1 and 5 for either diagnosis) on the diagnosis questions preceding request for confirmatory information with those preceding requests for dissenting at both time points (T1b and T2) for both cases. The null-hypothesis predicted no difference in confidence when requesting confirming and when requesting dissenting information. H3 predicted that the confidence would be higher for confirmatory requests than for dissenting requests. H3 was thus tested with a t-test for dependent samples for whether confidence was higher for confirmatory than for dissenting requests.

3.3.1: Test of H3 in Experiment 1. For Experiment 1, the confidence on diagnoses preceding confirmatory request (*M* = 2.43, *SD* = 0.85) was somewhat lower than the confidence on diagnoses preceding dissenting request (*M* = 2.64, *SD* = 0.88). The effect was thus in the opposite direction of that predicted by H3, not reaching significance ($t(66) = -1.76$, p = .96 one-tailed).

3.3.2: Test of H3 in Experiment 2. For Experiment 2, the confidence on diagnoses preceding confirmatory request (*M* = 2.38, *SD* = 0.91) was very close to the confidence on diagnoses preceding dissenting request (*M* = 2.38, *SD* = 1.08), not reaching significance ($t(46) = .01$, p = .49 one-tailed).

3.3.3: Test of H3 in Experiment 3. For Experiment 3, the confidence on diagnoses preceding confirmatory request (*M* = 2.19, *SD* = 0.75) was close to the levels of confidence on diagnoses preceding dissenting request (M = 2.08, SD = 0.85), not reaching significance ($t(80) = 1$, $p = .16$ one-tailed).

3.2.4: Test of H3 across all three experiments. When collapsing participants across all three experiments, the test for H3 remained non-significant ($t(194) = -0.39$, $p = .65$ one-tailed).

3.4: Test of decision process influencing confidence
Hypothesis H4 stated that having a consistent decision and only seeking confirming information should be associated with increased confidence in the decision. To avoid underpowered tests, this was only tested across all three experiments. An index was calculated for participants that preferred the same diagnosis on all three diagnostic questions for a case, and requested congruent information at both opportunities for the case. The index calculated the change in confidence from the first (T1b) to the last (T3) decision for the case (could vary between -3 to +3). For participants where both cases fit the criteria, an average of the confidence change in the two cases was used. Since few participants fit these criteria, this was not tested separately for each experiment but was tested across all three experiments (63 participants, 28%). H4 was thus tested as a one-sample t-test that there would be a positive change in confidence. Participants that had the same diagnosis on all three decisions (T1b, T2 and T3) and sought confirming information at both opportunities (T1b and T2) were included in this test (63 participants across all experiments, 28% of all participants). These participants increased their confidence with (M = .63, SD = 1.39), which was a significant change in the predicted direction ($t(67) = 3.75$, $p < .001$ one-tailed, $d = 0.45$).

3.5: Follow-up analyses
3.5.1: Analysis after removing non-naive participants. Experiment 3 included questions about what the participants thought the research hypotheses were. These were used in pre-registered follow-up analyses that tested whether the hypotheses listed above would be supported when excluding participants that appeared to have correctly guessed the hypotheses. After reviewing the responses, we excluded 19 participants who appeared to have fully or partly guessed the central research questions or any of the hypotheses, leaving 71 participants for a follow-up analysis. The analysis still showed no significant effects for the three hypotheses (H1: $t(71) = -4.39$, $p = 1$ one-tailed, H2: $t(71) = 1.07$, $p = .14$ one-tailed, H3: $t(63) = 0.86$, $p = .2$ one-tailed). Note that the test for H2 no longer approaches significance after removing participants that may have been aware of the hypotheses.

3.5.2: Two-tailed tests. All of the tests listed above were one-tailed due to directed hypotheses based on previous studies. However, some of the tests (H1 and H3) had effects in the opposite direction than predicted. We thus perform exploratory follow-up two-tailed analyses across all three experiments in order to examine these patterns. For H1 there was a significant effect ($t(217) = -3.06$, $p = .003$ two-tailed, $d = .21$) of participants selecting the diagnosis matching the symptoms presented last. There were no significant two-tailed effects across experiments for H2 ($t(217) = 1.68$, $p = .094$, $d = .11$) or for H3 ($t(194) = -0.93$, $p = .7$, $d = .03$). H4 was significant when tested one-tailed across experiments (se 3.4), and will not be tested two-tailed.

4: Discussion

4.1: Indications of decision-making biases

4.1.1: Summary of results. The aim of the current study was to develop a single experimental procedure for studying the interaction of primacy, congruence and confidence across the decision process. Further, we wanted to test the experiment in moderately sized samples of decision-makers within training in the subject matter. To this end, we performed three classroom experiments with online data collection in which advanced medical students were asked to diagnose two hypothetical mental health cases. All three experiments had the same overall structure and tested the same hypotheses, with only minor variations in wording of the cases and an additional question in the final experiment. Across three experiments, we found no support of primacy (H1), marginal support for congruence bias (H2), no support for confidence increasing confirmatory information seeking (H3), and support for confidence increasing when being consistent throughout the decision process (H4). The results for all hypotheses are summarized in Figure 2, and each of them are discussed in more detail below.

4.1.2: No indication of primacy. We expected that (H1) the symptoms presented first would create primacy for reading the rest of the case description, and would thus lead participants to choose the diagnosis matching the symptoms presented first rather than the diagnosis matching the symptoms presented second. However, this was not supported in any of the three experiments. In fact, all three experiments showed an effect in the opposite direction, such that the diagnosis more often matched the symptoms presented second. This is in contrast to previous studies that have found a primacy effect for medical decision-making in similar designs that manipulated the order of symptom presentation (see e.g. Cunnington et al., 1997; Richards & Wierzbicki, 1990)

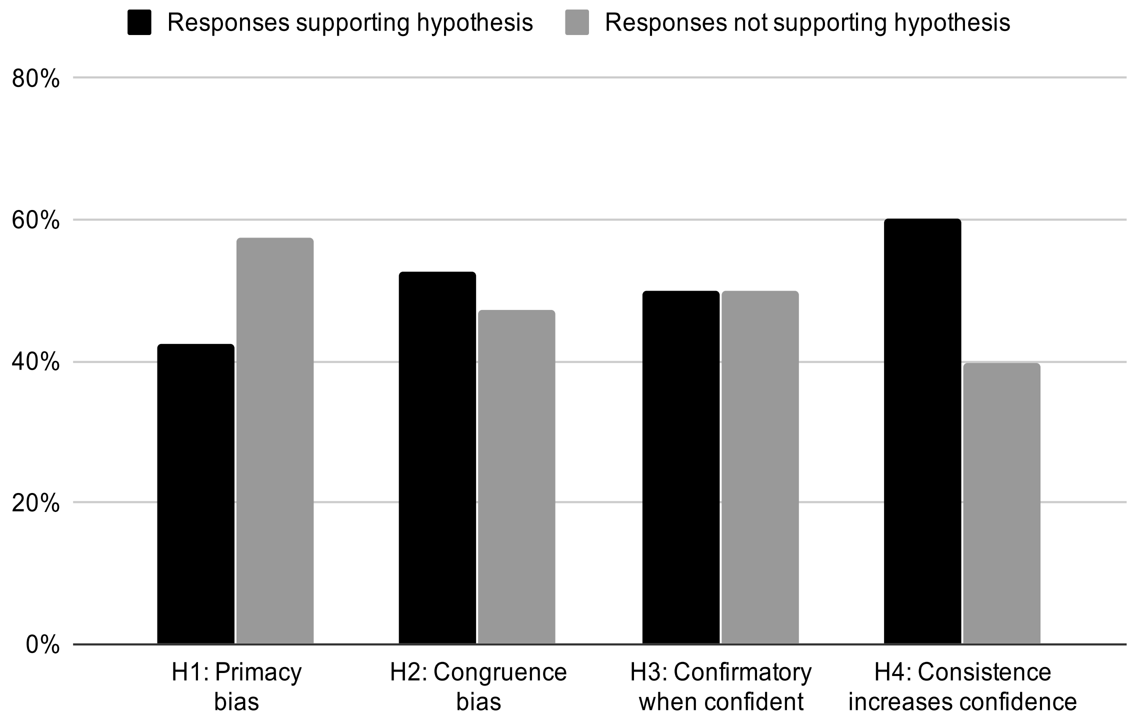

Figure 2: Illustration of the number of responses that support for each hypothesis across the three experiments.

.

Richards and Wierzbicki (1990) argued that it can be challenging to create case materials that are sufficiently balanced so that the mere order of symptoms is sufficient to tip the scales in favor of a given diagnosis, while avoiding imbalances due to the differences in description length or severity of symptoms. When designing the current materials, we also attempted to strike a balance between describing symptoms that pointed towards a specific diagnosis, yet being sufficiently ambiguous to not conclusively eliminate the competing diagnosis. If our case descriptions were not informative or balanced between the diagnoses, this may have prevented the experiments from producing a primacy effect. Our counterbalancing of diagnosis order (i.e. half the participants had symptoms of diagnosis A first, while the other half had diagnosis B first) should decrease any effects of poor balance between the diagnoses. Further, examining the data shows that the initial decisions had a roughly normal distribution centered around having low confidence in one of the diagnoses, which indicates that the case descriptions are fairly balanced.

It should be emphasized that instead of showing primacy in preferring the diagnosis matching the symptoms presented first, our results showed the opposite pattern, of preferring the diagnosis matching the symptoms presented last. This could indicate a *recency effect* (Botto, Basso, Ferrari, & Palladino, 2014; Murdock, 1962; Tzeng, 1973) of the symptoms presented most recently being more available in working memory, and thus having a larger impact on the decision (see similar effects in e.g. Bergus, Chapman, Gjerde, & Elstein, 1995; Tubbs, Gaeth, Levin, & Van Osdol, 1993). The current design and results could also be compared to a "serial anchoring" effect (see Bahník, Houdek, Vrbová, & Hájek, 2019). When using two "anchors" in opposite directions, they found the second anchor to impact the decision. When compared to our task, such a theoretical approach may see the two halves of our initial symptom information to be two sequential anchors in opposite directions.

It should be noted that mixed results have been reported when attempting to replicate previous research designs on medical decision-making (Ellis et al., 1990; Friedlander & Stockman, 1983). It is possible that the complexity inherent in real-life decisions among expert decision-makers makes it difficult to replicate primacy effects that have been shown for naive participants making abstract decisions. This could be due to the decision being affected by various prior assumptions, strategies and preferences (Hutton & Klein, 1999).

4.1.3: Confirmatory information seeking. Congruence bias was operationalized across the three experiments as more often requesting follow-up information that could confirm the diagnosis that the participant preferred on the preceding decision, rather than information to confirm the opposing diagnosis. We expected (H2) that there would be more requests for confirming, rather than dissenting information. This test for congruence bias approached significance in Experiments 2 and 3, and was significant when collapsed across the three experiments.

The study as a whole thus found indications of confirmatory congruence bias when making medical diagnostic decisions, in terms of seeking information that could support the diagnosis one currently holds, rather than seeking information that could falsify the assumption. This finding is compatible with other studies (Martin, 2001; Mendel et al., 2011; Oskamp, 1965; Parmley, 2006). This indicates that the previously identified congruence bias phenomenon (Jonas, Schulz-Hardt, Frey, & Thelen, 2001; Schulz-Hardt, Frey, Lüthgens, & Moscovici, 2000) extends to novel experimental settings. Further, this indicates that the confirmatory congruence bias could be relevant for the mental health domain. Decision-makers in this domain should thus be aware that such a bias can lead to non-optimal information gathering and decision-making (Blumenthal-Barby & Krieger, 2015).

However, while the argument can be made that the current study as a whole found indications for a congruence bias, it should be noted that the effect sizes were small, and the pre-registered analyses of each experiment individually did not show significant effects. This partly contradicts previous research (Martin, 2001; Mendel et al., 2011; Oskamp, 1965; Parmley, 2006) that has indicated that congruence bias is a robust effect that should reliably reproduce in each of the experiments in the current study. It should be noted that most of the studies on biases in diagnostic decision-making also appear to have small effects, which resemble our pooled results.

It should also be noted that our study materials gave ambiguous feedback on the follow-up questions. This may have confused or annoyed the participants, which could motivate them to use a more analytical mode of thinking (see Croskerry, 2009b). A possible limitation of the current experiment design is that we do not know the participants' motivation for asking follow-up questions. We have assumed that participants asked about symptoms in order to support the associated diagnosis if the symptom is present. However, participants could also have asked about a symptom with the intention of discounting the associated diagnosis if the information they received would indicate that symptom is absent. Participants might thus have used an information-seeking strategy intended to confirm their assumptions that would not be registered as congruence in our operationalization. Similar mechanisms may also have been present in previous studies of congruence bias.

4.1.4: Confidence not leading to confirmatory information seeking. We expected (H3) that higher confidence in a diagnostic decision would precede requests for confirmatory information, compared to the confidence measures preceding requests for dissenting information. However, although there was an overall tendency for seeking confirmatory information (see H2), there was no support in any of the experiments for increased confirmatory information seeking when participants were more certain of their diagnostic decision. This null-finding thus fails to support a previous finding (Martin, 2001) of higher confidence leading to more confirmatory information seeking. We tested for the effect of confidence on information seeking in order to explore whether this could be a mechanism for congruence bias. The absence of such an effect may indicate that confirmatory information seeking is not motivated by the degree of confidence in the preferred decision, but instead is dichotomous in seeking information that supports the preferred decision.

It should be noted that we used a novel response mode in the current experiments, where participants indicated on a 10-point scale how confident they were in one of the diagnoses. It is possible that participants' use of the scale did not represent their actual degree of confidence, and cannot be interpreted as such. However, note that most participants indicated low confidence in their decisions (see Figure 1), which could be expected given the ambiguous case descriptions and follow-up information, and may thus be taken to supporting our measure of confidence.

4.1.5: Increased confidence when not exploring alternatives. We expected (H4) that the subset of participants who kept to the same diagnosis and only sought confirmatory information would show an increase of confidence in their decision. This was supported (when analyzed across the three experiments), in the sense that the consistent participants increased their confidence during the decision process. As the follow-up information was designed to be ambiguous and should not provide the participants with any additional conclusive information, the increase in confidence could be said to indicate overconfidence (Oskamp, 1965). Similar effects have been seen in a related study (Martin, 2001), although this used more conclusive information, and which showed that experts were less confident than novices.

4.2: Limitations and further research

4.2.1 Experiment design. The current study had a novel experiment design in order to test the combination of primacy, congruence and confidence on seeking and evaluating information throughout the decision process. In order to ensure sufficient samples, the experiment was designed to be short so that it could be completed in a break between two lectures. Compared to other experiments with more comprehensive materials, the fact that we provided participants with only a limited amount of information may have allowed them to deliberately process all the provided information, and thus reduce heuristic processing. This may account for current results deviating from similar designs where more patient information was provided (see e.g. Mendel et al., 2011). This may detract from the validity of the current study, as more information would be available in most real-life mental health decisions.

The case descriptions were designed to be ambiguous and for follow-up questions to not provide any conclusive information. It could be that there would have been a primacy effect if the case had first provided a clear indication for a given diagnosis, that would then be contradicted. However, such an approach may have stretched the validity of the cases). Another approach could be that participants received feedback on the follow-up questions that to some extent resolve their uncertainty. Such an approach could make the participants more invested in the use of follow-up questions, which may have enhanced a congruence bias. However, this may have made it more difficult to interpret the sequential responses as indicators of the decision process.

A possible reason for the lack of a primacy effect in the current study may be that participants did not commit to a decision after reading the first symptoms. To test for this, Experiment 3 asked participants to make a preliminary diagnosis after hearing only the first half of the symptoms. However, this did not have the intended effect of the decision after hearing the second half of the symptoms to be more in line with the first half. On the contrary, while the decision after hearing all symptoms in Experiment 1 and 2 were evenly distributed between the two diagnoses, in Experiment 3 there was a strong preference for the decision matching the second half of the symptoms. This may be due to demand characteristic effects (Orne, 1962; Strohmetz, 2008), where the participants in Experiment 3 believed that a different response based on the additional information is expected when the same question is asked for the second time.

4.2.2 Participant bias. Based on debrief conversations after Experiment 1 and 2 we suspected that some of the participants made assumptions about the research hypotheses which may have influenced their responses. Attempting to measure this in Experiment 3 indicated that about a fifth of the participants correctly guessed one or more of the research hypotheses. Removing these participants did not change the significant results from Experiment 3. Nevertheless, it is possible that such artefacts may have impacted the results in Experiment 1 and 2 or in previous research. If participants are familiar with the cognitive biases or are suspicious of the research paradigm, they may be more careful in their decision making than they would otherwise be.

4.2.3 Statistical power. The sample size of our current three experiments were restricted by practical concerns (the number of medical students at our local university). The three studies independently (at average n = 72) had sufficient power to detect effects of Cohen's d = .36 or larger (given power of .8 and alpha of .05 one-sided). Alternatively, pooling all participants (n = 224) gives sufficient power to detect effects of d = .21. The studies on decision making in mental health cited above often fail to provide sufficient information to calculate effect sizes for primacy, congruence bias and overconfidence. Some of the studies (e.g. Richards & Wierzbicki, 1990) have described their effects to be between weak and moderate (thus corresponding to between 0.2 and 0.5). If this is accurate, the current study is more or less sufficiently powered to detect effects at the relevant magnitude, at least when collapsing across all three experiments. It should be mentioned that previous experiments that have established these effects typically have had low power. Nevertheless, we should take care to not read too much into the lack of statistically significant effects in the current study, as the experiment may have been underpowered to detect effects. It is possible that the examined biases have weaker effects than previously assumed when tested in somewhat realistic problem-fields in which the decision-maker has relevant expertise, and that higher-powered studies and strong manipulations are therefore required to demonstrate these biases.

4.3: Implications and future research

This is the first study investigating the interaction of primacy, congruence and overconfidence bias during a decision process for trained decision-makers in mental health diagnostics. The current experiment procedure and materials may inspire similar explorations of medical decision-making. A transparent research process with pre-registration, open materials and data may assist the planning of future studies. To improve ecological validity and measurement specificity one may consider expanding the number of clinical cases, symptom information or follow-up questions. Another approach could be to measure meta-knowledge during the decision process, similar to what we did at the end of Experiment 3.

The fact that the current study failed to find primacy and yielded unclear results for a congruence bias, may indicate that the biases are less reliable or have weaker effects for real-life decision-making than what has been indicated in past literature. While several studies have detected these biases in similar settings (Mendel et al., 2011; Parmley, 2006; Richards & Wierzbicki, 1990), there have also been failed replications (Ellis et al., 1990). We argue above that the complexity of the medical diagnostic process may yield more opportunities for biases than in simpler one-off decisions (see similar arguments in e.g. Blumenthal-Barby & Krieger, 2015; Saposnik, Redelmeier, Ruff, & Tobler, 2016). However, such complexity also allows for different processes and interactions being in effect, and that the biases previously shown in simpler situations are only relevant in a subset of these processes, or that they may be countered by other processes. Cognitive biases such as "loss aversion" have been shown to be modulated by factors such as domain knowledge, experience and education (Mrkva, Johnson, Gächter, & Herrmann, 2019).

The findings previously shown in the literature may have relied on using specific manipulations to provoke biases in certain settings. Norman and Eva (2010, p. 97) argued that some previous demonstrations of biases may "induce error for the sake of determining if the biases exist". Alternative approaches such as the "Naturalistic decision making" (Klein, 2015) and "ecological rationality" (Gigerenzer, 2008) have expressed doubts as to the extent that biases have substantial negative impacts on experts' real-life decision-making. Recent development in psychological science (Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014) have emphasized how research practices such as analysis flexibility, selective publication and conceptual rather than direct replications may have generated and propagated false positive findings and makes it difficult to know whether effects generate to other settings. Such practices may have contributed to overestimating the reliability of primacy and congruence biases, and may have prevented us from identifying conditions where the biases would not be in effect. This may partly explain why the current results deviates from the majority view in the literature.

Acknowledgements

The theoretical framework and the experiment designs for the current study was developed by TF, ØKB and BS. TF and ØKB developed the case descriptions and did data collection for Experiments 1 and 2. VTS did data collection for Experiment 3. BS managed the project. All authors contributed in writing the manuscript. Results from Experiments 1 and 2 have previously been part of a master thesis by TF & ØKB, supervised by BS (ref bora). Thanks to our collaborators at the Faculty of Medicine: Anders Bærheim, Øystein Hetlevik, and Thomas Mildestvedt. Thanks to the participants in the three experiments.

References

Bahník, Š., Houdek, P., Vrbová, L., & Hájek, J. (2019). Variations on anchoring: Sequential anchoring revisited. *Judgment and Decision Making, 14*(6), 711-720.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational behavior and human decision processes, 42*(1), 88-110. doi:10.1016/0749-5978(88)90021-0

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *The Quarterly Journal of Experimental Psychology, 40*(2), 269-297. doi:10.1080/02724988843000122

Bergus, G., Chapman, G., Gjerde, C., & Elstein, A. (1995). Clinical reasoning about new symptoms despite preexisting disease: sources of error and order effects. *Family medicine, 27*(5), 314-320.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5), S2-S23.

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making, 35*(4), 539-557.

Botto, M., Basso, D., Ferrari, M., & Palladino, P. (2014). When working memory updating requires updating: Analysis of serial position in a running memory task. *Acta Psychol (Amst), 148*, 123-129. doi:10.1016/j.actpsy.2014.01.012

Croskerry, P. (2002). Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine, 9*(11), 1184-1204.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*(8), 775-780.

Croskerry, P. (2009a). Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in Health Sciences Education, 14*(1), 27-35. doi:10.1007/s10459-009-9182-2

Croskerry, P. (2009b). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022-1028. doi:10.1097/ACM.0b013e3181ace703

Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., . . . Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education, 18*(3), 343-363. doi:10.1007/s10459-012-9374-z

Cunnington, J. P., Turnbull, J. M., Regher, G., Marriott, M., & Norman, G. R. (1997). The effect of presentation order in clinical decision making. *Academic Medicine.*

Curley, S. P., Young, M. J., Kingry, M. J., & Yates, J. F. (1988). Primacy effects in clinical judgments of contingency. *Medical Decision Making, 8*(3), 216-222. doi:10.1177/0272989X8800800310

Ellis, M. V., Robbins, E. S., Schult, D., Ladany, N., & Banker, J. (1990). Anchoring errors in clinical judgments: Type I error, adjustment, or mitigation? *Journal of counseling psychology, 37*(3), 343.

Eva, W. K. (2001). The influence of differentially processing evidence on diagnostic decision-making.

Friedlander, M. L., & Stockman, S. J. (1983). Anchoring and publicity effects in clinical judgment. *Journal of Clinical Psychology, 39*(4), 637-644.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20-29.

Graber, M. (2005). Diagnostic errors in medicine: a case of neglect. *The Joint Commission Journal on Quality and Patient Safety, 31*(2), 106-113.

Graber, M. (2011). Diagnostic error: the hidden epidemic. *Physician executive, 37*(6), 12.

Graber, M. (2013). The incidence of diagnostic error in medicine. *BMJ Qual Saf, 22*(Suppl 2), ii21-ii27.

Graber, M., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of internal medicine, 165*(13), 1493-1499.

Graber, M., Gordon, R., & Franklin, N. (2002). Reducing diagnostic errors in medicine: what's the goal? *Academic Medicine, 77*(10), 981-992.

Hannigan, B., & Coffey, M. (2011). Where the wicked problems are: the case of mental health. *Health policy, 101*(3), 220-227.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*(1), 1-55. doi:10.1016/0010-0285(92)90002-J

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science, 24*(5), 379-385.

Hutton, R. J., & Klein, G. (1999). Expert decision making. Systems Engineering: The Journal of The International Council on Systems Engineering, 2(1), 32-45.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*(5), 235-241.

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology, 80*(4), 557.

Klein, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition, 4*(3), 164-168.

Ludolph, R., & Schulz, P. J. (2018). Debiasing health-related judgments and decision making: a systematic review. *Medical Decision Making, 38*(1), 3-13.

Lund, F. H. (1925). The psychology of belief: IV. The law of primacy in persuasion. *Journal of Abnormal and Social Psychology, 20*(2), 183-191.

Martin, J. M. (2001). Confirmation bias in the therapy session: The effects of expertise, external validity, instruction set, confidence and diagnostic accuracy. ProQuest Information & Learning,

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., . . . Hamann, J. (2011). Confirmation bias: why

psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine, 41*(12), 2651-2659.
Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences, 14*(10), 435-440.
Mrkva, K., Johnson, E. J., Gächter, S., & Herrmann, A. (2019). Moderating Loss Aversion: Loss Aversion Has Moderators, But Reports of its Death are Greatly Exaggerated. *Journal of Consumer Psychology*.
Murdock, B. B. J. (1962). The serial position effect of free recall. *Journal of experimental psychology, 64*(5), 482. doi:10.1037/h0045106
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2*(2), 175-220.
Norman, G. R. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*(1), 37-49. doi:10.1007/s10459-009-9179-x
Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical education, 44*(1), 94-100.
Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776.
Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of consulting psychology, 29*(3), 261.
Parmley, M. C. (2006). The Effects of the Confirmation Bias on Diagnostic Decision Making. Drexel University,
Payne, V. L. (2011). Effect of a metacognitive intervention on cognitive heuristic use during diagnostic reasoning. University of Pittsburgh,
Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical education online, 16*(1), 5890.
Richards, M. S., & Wierzbicki, M. (1990). Anchoring errors in clinical-like judgments. *Journal of Clinical Psychology, 46*(3), 358-365.
Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making, 16*(1), 138.
Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of personality and social psychology, 78*(4), 655.
Strohmetz, D. B. (2008). Research artifacts and the social psychology of psychological experiments. *Social and Personality Psychology Compass, 2*(2), 861-877.
Todd, P. M., & Gigerenzer, G. (2001). Putting naturalistic decision making into the adaptive toolbox. *Journal of Behavioral Decision Making, 14*(5), 381-383.
Tubbs, R. M., Gaeth, G. J., Levin, I. P., & Van Osdol, L. A. (1993). Order effects in belief updating with consistent and inconsistent evidence. *Journal of Behavioral Decision Making, 6*(4), 257-269.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.
Tzeng, O. J. (1973). Positive recency effect in a delayed free recall. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 436-439. doi:10.1016/S0022-5371(73)80023-4
van den Berge, K., & Mamede, S. (2013). Cognitive diagnostic error in internal medicine. *European journal of internal medicine, 24*(6), 525-529.
Xiao, Q., Lan, C. S. E., Piara, M., & Feldman, G. (2020). Revisiting status quo bias: Replication of Samuelson and Zeckhauser (1988). doi:10.31234/osf.io/ta97c