Abstract

Purpose. Unexpected questions have been shown to increase cues to deception, without reducing the information given by truth tellers. Two studies investigated whether asking expected versus unexpected questions first impacted on the amount of detail provided by interviewees.

Methodology. In Study 1, participants (N = 85) were interviewed about their own intentions and in Study 2, participants (N = 84) were interviewed about an intention given to them by the experimenter.

Results. Analyses showed that, in both studies, differences between the expected-first and the unexpected-first order were minimal and lie detection accuracy was not improved by asking the unexpected questions first. Unlike in previous research, differences between truth tellers and liars were not larger for unexpected questions than expected questions.

Conclusions. These results offer important information for forensic interviewers, showing that there is no need to ask unexpected questions at a specific point during the interviews. Link to associated OSF page: https://osf.io/93g7h/?view_only=680db3652bff44cfaa1b6534eb04688e (anonymised version)

Unexpected questions in deception detection interviews: does question order matter?

Interviews in forensic settings can have two goals: to obtain information from the interviewee and to determine whether the interviewee is lying or telling the truth (Fisher, 1995; Vrij, Mann, Kristen and Fisher, 2007; Nahari et al., 2019, commentary 1). How much information is obtained is often not measured in the amount of words used by an interviewee, but in the amount of details provided. Several schemes have been developed to code for the amount of details provided by an interviewee. Reality Monitoring (RM) is one commonly used scheme (e.g. Nahari & Vrij, 2013), which uses several criteria to measure the richness of detail provided by interviewees.

Different interview techniques have been developed to increase the amount of details obtained from an interviewee, which benefits both aims of the interview: to obtain information and detect deception. Asking unexpected questions is one such technique. It is based on the cognitive load theory of deception, which states that lying is more cognitively demanding than telling the truth and that increasing the cognitive load for interviewees should increase the number of deception cues (Vrij, Fisher, Mann, & Leal, S., 2006; Vrij, Fisher & Blank, 2017). Both truthful and deceptive interviewees provide less details in response to unexpected questions than to expected questions, a sign that answering unexpected questions is harder. The reduction in detail for unexpected questions is larger for liars than truth tellers (Lancaster, Vrij, Hope and Waller, 2012). Liars may answer unexpected questions in less detail than truth tellers because liars rely on answers that they have prepared beforehand. For unexpected questions, liars have no answer prepared, therefore they have to create one, which increases difficulty. Truth tellers also find it harder to answer unexpected questions, but they can rely on their memory. Retrieving a memory is likely to be easier than creating an answer (Clemens, Granhag & Strömwall, 2013, Vrij et al., 2009).

There is evidence that the unexpected questions technique can be effective in interviews about intentions as well as those about past events (Warmelink, Vrij, Mann, Jundi and Granhag, 2012), unlike other techniques derived from the cognitive load technique (e.g. Fenn, McGuire, Langleben and Blandón-Gitlin, 2015). In interviews about intentions, liars give more detail than truth tellers in response to expected questions, but less detail in response to unexpected questions (Warmelink et al. 2012). This is a slightly different pattern than 'liars give less detail than truth tellers' found in interviews about other topics (DePaulo et al. 2003). Warmelink et al. (2012) speculated that this difference is caused by lying participants 'over-preparing' their answers. They create an answer that is more detailed than most truth tellers are willing or able to give about their intentions. A similar 'over-preparation' effect was found in written truthful or deceptive statements about intentions (Kleinberg, van der Toolen, Vrij, Arntz, & Verschuere, 2018).

What makes a question unexpected? Some expected questions can be made unexpected by adding a constraint (e.g. use reverse order or a fixed perspective; Lancaster et al. 2012). The downside of this approach is that these unexpected elements are known to be cognitively demanding independently of their unexpectedness. Some questions appear to be unexpected because they ask for information the interviewee did not expect the interviewer to want. Warmelink et al. (2012), Sooniste, Granhag, Knieps & Vrij, 2013, Clemens, Granhag & Strömwall (2013) and Mac Giolla and Granhag (2015) found that questions about the planning of an intention are unexpected for interviewees. Warmelink, et al. (2012) also found that questions requesting details about the most salient aspect of the intention or about travel required for the intention are unexpected.

Clemens, Granhag and Strömwall (2011) manipulated the order of questions about the intention and planning for the intention. They found that asking unexpected (planning) questions first increased the difference between truth tellers' and liars' within-statement consistency, compared to asking expected (intention) questions first. Asking unexpected questions first might disrupt liars' strategies to be as convincing as possible and lead to better deception detection. Vrij et al. (2018) studied similar order effects in interpreter-mediated interviews that included questions about a trip (expected) and questions about the planning of a trip (unexpected). They found that the order of these questions did not affect the amount of detail provided. However, interpreter mediated interviews tend to elicit less detail, which may make detail effects harder to study (Ewens, Vrij, Leal, Mann, Jo & Fisher, 2016).

Asking unexpected questions first may disrupt the interviewee's mental representation of the interview. Fisher (1995) suggests that the order of the questions in an interview should be compatible with the interviewees' mental representation of events. It is possible that it is not only the events discussed, but also the unfolding interview that needs to match the interviewee's representation in order for the interview to be successful. If the unexpected-first order is more difficult for interviewees, this may reduce the amount of information elicited.

In this article, we report two studies of participants being interviewed about the interviewee's plans later that day ("their intention"). The two studies used the two main paradigms that are used to study detecting deception about intentions (Granhag & Mac Giolla, 2014). By using both paradigms, we can get a conceptual replication of the results and discuss differences between the two paradigms. In Study 1, participants were interviewed about their own intentions, which they created independently of the study. In Study 2, the experimenter gave the participants an intention that they were told to complete after the interview. The interview questions, manipulations and hypotheses were the same in both studies.

The interviews involved four types of questions about the participants' intention: 1) general questions about the intention and its familiarity ('general'); 2) details about the most salient event of the intention ('core'); 3) how they would travel in order to reach the location for the intention ('travel') and 4) the planning they already completed or were intending to complete in order to ensure they'd be able to execute the intention ('planning'). Following Warmelink et al. (2012), Clemens, Granhag and Strömwall (2013) and Mac Giolla and Granhag (2015), we expected that general questions would be most expected, then core and travel, with planning questions the least expected.

We hypothesised that the expected-first order (general, core, travel, planning) would be less cognitively demanding than an unexpected-first order (planning, travel, core, general) and that this would be visible in several ways. Firstly, participants in the unexpected-first condition would experience the study as more difficult (Hypothesis 1). Secondly, participants in the unexpected-first condition should give fewer details, particularly perceptual and contextual details (Hypothesis 2). Thirdly, the extra cognitive load induced in the unexpected-first condition should affect liars more than truth tellers, due to the higher cognitive load associated with lying. Therefore, the difference in number of details between truth tellers and liars should be larger in the unexpected-first condition than in the expected-first condition (Hypothesis 3).

Study 1: Self-induced intentions

Method.

Participants. Eighty-five participants took part (61 women, 13 men, 11 non-disclosed). Their mean age was 19.11 (SD = 2.70). All participants were students (62 psychology majors, 16 psychology minors; 7 unknown). The study was conducted at X University, concurrently with the second study. Both were approved by the University Research Ethics Committee. Sample size for both studies were based on Warmelink et al. (2012).

Design. The experiment had a 2 (Veracity, between: truth, lie) x 2 (Order, between: expected-first, unexpected-first) x 4 (Question type, within: general, core, travel, planning) mixed design. Participants were randomly allocated to the conditions. There were three main interviewers: authors Y (N = 15), Z (N = 44) and W (N = 15) and several emergency interviewers with less than 8 interviews each (collapsed into 'other interviewers').

The dependent variables were question expectedness and difficulty and level of detail provided (see Appendix A; all appendices are available on OSF [https://osf.io/93g7h/?view_only=680db3652bff44cfaa1b6534eb04688e).

Materials. The participants were recorded in an interview suite with wall mounted camera's and a ceiling mounted microphone. A hand-held audio recorder was used as back-up.

Procedure.

Pre-Interview Procedure. The participants signed up via an online participant pool (SONA). The study advertisement informed them that they would be interviewed about their intended activities immediately after they completed the study and that they might be asked to lie about these intentions during the study. Participants were emailed 1 to 3 days before the study. They were asked to provide information about their intentions and, for the liars, about the lie they were going to tell. When participants arrived at the lab they signed informed consent forms. Participants in both conditions were reminded that their task was to convince the interviewer they were telling the truth and asked to confirm that their intention had not changed since their initial email.

Interview procedure. The participants were interviewed by an interviewer who was blind to their veracity. The interviewer was informed of their order condition, when the participant arrived. The interviewer asked the participants the interview questions (Appendix B; OSF), in either expected-first order or in unexpected-first order.

Post-Interview procedure. Once the interview was completed, the interviewer rated the participants' veracity and their confidence in this rating, while participants filled in a questionnaire about their experience of the interview. For purposes beyond the scope of this paper (see X, under review), participants then completed a reaction time task and a questionnaire about that task. The participants were then given a debriefing form and their SONA credits.

Data Analysis for both studies

Coding. The coding for both studies was done concurrently and using the same procedure. The interviews were transcribed by a transcription service. The transcripts were coded (by author Y) and 33 transcripts (20%) were coded for reliability (author X). Both coders were blind to the Veracity condition of the transcript and the first coder was blind to the hypotheses of the experiment. After the first and the 15th transcript the coders discussed the coding and the correlation between the coders was checked; the coders clarified any differences between them.

The coding scheme (Table 1) was similar to that used in Warmelink et al. (2012, 2013). These codes can be mapped onto the RM criteria (e.g. Nahari & Vrij, 2013), 'perceptual details' and 'contextual details'. Hedges are a detail type coded for in Warmelink et al. (2012, 2013), that do not map onto RM codes. They were added to the coding as the participants in these studies were speaking about future behaviour. Since future events are less certain than past events, they are often described with more hedges. The only difference with Warmelink et al. (2012, 2013) was that an extra type of detail ("knowledge") was added at the coding stage, to capture some details that did not seem to fit any of the other categories.

The reliability of the coding was measured using correlations and ICC on the numbers of details over the entire transcript (see Table 1). The highly variable nature of this data makes the commonly used Kappa statistic unsuitable (Vierra and Garrett, 2005). All details except for Knowledge had correlations in the moderate or strong range. Due to the weak-to-moderate correlation, knowledge details were excluded from the remaining analyses, leaving 9 detail categories. The detail categories Visual, Auditory, Smell/Taste, Tactile and Action were then combined to form a *perceptual details* category. The Spatial and Temporal details were combined to form a *contextual details* category. *Emotions* were retained as separate category, in line with RM, as were the *hedge* details. A *total details* category was created by summing the perceptual, contextual and emotion details. Hedges were not included as they were not considered a detail as such: they express uncertainty rather than provide information.

Analyses. Repeated measures ANOVAs were conducted with veracity and order as the independent/fixed factors and interviewer was added as a covariate. The interviewer covariate was included to control for any individual effect that the interviewer may have[1]. As results for the interviewer variable do not test any of our hypotheses, they are not reported.

Pallai's trace and Greenhouse-Geisser corrected values are reported for the repeated measures ANOVA's. To correct for the analysis of four different types of detail, the Bonferroni correction was applied, leading to an alpha of 0.01. Word count was not controlled for on any of the variables, as this tends to reduce content cues to deception (Elntib, Wagstaff, & Wheatcroft, 2015) and our hypotheses centre on the amount of detail not on detail density.

Bayes factors were also calculated for all 1-df analyses. Bayes factors provide a continuous measure of evidence for the alternative hypothesis relative to the null hypothesis. We used the Dienes and Mclatchie (2018) R script (full code available in supplementary materials) to calculate Bayes factors. The prior model was specified using the half-normal distribution, setting the SD to the predicted effect size (Dienes, 2014). In the absence of prior research on the order effect, we used the effect size for veracity from previous research (Warmelink, 2012; Warmelink et al., 2012) as an approximate scale of effect that details can be influenced by within an interview context. We treat Bayes factors between 0.33 and 3 as weak evidence, while Bayes

[1] There are significant interviewer by question type interaction effects for total details in both studies (self-induced: multivariate F (3,72) = 3.59, p = 0.02, η^2_p = 0.13; other-induced: (multivariate F (3,74) = 7.88, p < 0.001, η^2_p = 0.24. To test whether the inclusion of the covariate was affecting the results, the total details analysis was rerun without this covariate. There were no differences between the two analyses: all effects that were significant when interviewer was included were significant when it wasn't. The same was true for non-significant effects.

factors in the ranges 0.05-0.33 and 3-20 are moderate for the null and experimental hypothesis respectively, and Bayes factors <0.05 and above 20 are strong evidence for the null and experimental hypothesis respectively (Jarosz & Wiley, 2014).

Results.

Question expectedness. The ANOVA showed a main effect of question type (F (3, 72) = 25.09, $p < 0.001$, $\eta^2_p = 0.51$). No other significant main or interaction effects were found (all Fs < 3.08, all ps > 0.03). The expectedness ratings matched the hypothesis, with general question considered the most expected, then core, then travel, and planning was the least expected question type (see Table 2).

There was a significant main effect of question type on the total number of details reported, F (3, 72) = 17.24, $p < 0.001$, $\eta^2_p = 0.42$. Participants gave the most details in response to travel questions, followed by the core questions, then the most expected general question, with the planning questions eliciting the least details. There was no significant interaction effect between question type and veracity, F (3, 72) = 1.66, $p = 0.18$, $\eta^2_p = 0.07$.

Difficulty. The ANOVA showed a significant main effect of question type (F (2.84, 222) = 4.28, $p = 0.01$, $\eta^2_p = 0.06$). Planning questions were rated as the most difficult, followed by core questions; both were experienced as significantly more difficult than the travel questions and general questions (see Table 2). There were no other significant effects (all Fs < 1.74, all ps > 0.17).

Detail. The ANOVAs showed no significant main effects of order or order by question type interaction effects (all Fs < 3.29, all ps >0.03). However, the Bayes factors indicated that for total details there was moderate evidence, $B_{H(0, 37.83)} = 3.58$, that participants gave more detail in the expected first condition (M = 160.28, SD =

78.08) than in the unexpected-first condition (*M* = 131.58, *SD* = 53.52), F (1,74) = 3.75, p = 0.06, η^2_p = 0.05. The Bayes factor also provided moderate evidence, $B_{H(0, 7.05)}$ = 3.64, that participants gave more contextual detail in the expected first condition (*M* = 44.55, *SD* = 26.42) than in the unexpected-first condition (*M* = 35.14, *SD* = 17.03), F (1,77) = 3.62, p = 0.06, η^2_p = 0.05. The Bayes factors provided weak, inconclusive support for the experimental hypothesis for perceptual, emotion and hedge main effects (see tables 3 and 4 for full results).

Lie detection cues. None of the details showed a three-way effect between veracity, order and question type (all *F*s < 1.86, all *p*s > 0.15) or a veracity by question order effect (all *F*s < 3.71, all *p*s > 0.06). The Bayes factors show weak, inconclusive evidence for the experimental hypothesis for total ($B_{H(0, 37.83)}$ = 1.13), perceptual, contextual, emotion, and hedge details (full results in Table 3 and 4). Further research is required in order to determine whether question order increases the difference between truth tellers and liars in the amount of information provided.

Study 2: Other-induced intentions study

Method.

Participants. Eighty-four participants were recruited (61 women, 18 men, 5 non-disclosed; mean age = 19.18 (1.93)). All were students (62 psychology majors, 13 psychology minors and 9 unspecified students). This study was run concurrently with study 1. No participants took part in both studies.

Materials. The same materials as in the self-induced study were used. Additionally, participants in this study were asked to complete a questionnaire asking them to rate the strength of the intention to execute the experimental task (Appendix C; OSF).

Design. The design was the same as for Study 1. Authors Y (*N* = 27), Z (*N* = 26) and W (*N* = 27) completed the majority of the interviews, 4 were completed by one of the other interviewers.

Procedure.

Pre-Interview Procedure. Participants signed up via the online participant pool and received their instructions after signing the informed consent form in the lab. Truth tellers were instructed to complete an intention (to buy two items from a shop) and to tell the truth about this intentions during the interview. The liars were given their intention (to drop off a USB-stick) and were told that in the interview they should lie about their intention and pretend that they had the truth teller's intention, for which they received the same instructions as the truth tellers. Participants in both veracity conditions were given 8 minutes to "prepare or plan" for completing their intention and for the interview. They were given the opportunity for more time to prepare (only one participant did).

*Interview Procedure. S*ame as the Self-induced study.

Post-interview procedure. After the participants completed the questionnaire and the unreported reaction time task, they were informed they would not have to execute their intentions. The participants were given the opportunity to ask the experimenter questions about the deception that the experimenter had perpetrated. They were reminded that they could withdraw their data if they wished (no one withdrew). They were then asked to complete a questionnaire on whether they had believed that they had to execute the intention and how they prepared for this (see Appendix C; OSF). The participants received a debriefing form and their credit for the SONA participant pool.

Data analysis. See data analysis for the self-induced study.

Results.

Induced Intention manipulation check. Participants indicated that they were convinced that they would have to complete the intention (M = 5.99, SD = 1.14, out of 7) and that they were motivated to do so (M = 5.60, SD = 0.94, out of 7). This did not differ significantly across different conditions (all Fs < 2.74, all ps > 0.12).

Question expectedness. The ANOVA showed a significant main effect of question type, F (2.73, 231) = 5.58, p = 0.001, η^2_p = 0.07. As hypothesised, the general questions were rated as most expected, core, travel and planning questions were all expected less. Core, travel and planning questions did not differ significantly from each other (see Table 5). No other significant main or interaction effects were found (all Fs < 1.06, all ps > 0.36).

There was a significant main effect of question type on total details reported, F (3, 72) = 40.41, p < 0.001, η^2_p = 0.63. Participants gave the most details in response to travel questions, followed by the core questions, then the general question, with the planning questions eliciting the least details. This pattern matches Study one (see Table 5). There was no significant interaction effect between question type and veracity, F (3, 72) = 0.31, p = 0.82, η^2_p = 0.01.

Difficulty. The ANOVA showed a significant main effect of veracity, F (1, 76) = 19.84, p < 0.001, η^2_p = 0.21. This interacted with question type, F (2.89, 228) = 4.73, p = 0.003, η^2_p = 0.06. For all question types, except travel, the liars rated the questions as more difficult than truth tellers. However, there was no effect of question order, F(1, 76) = 2.47, p = 0.12, η^2_p = 0.03, nor did veracity and question order interact, F (1, 76) = 1.65, p = .20, η^2_p = 0.02.

Detail. The ANOVAs showed no significant question order main effects or order by question type interaction effects (All *Fs* < 6.53, all *p*s > 0.01) at the corrected alpha level of 0.01. The Bayes factors provided weak evidence for the null hypothesis for total, perceptual, contextual and emotion details, and moderate evidence for the null hypothesis for hedge details, $B_{H(0,\ 1.58)}$ = 0.21, (see Tables 6 and 7).

Lie detection cues. None of the details showed a three-way effect between veracity, order and question type (all *F*s < 1.85, all *p*s > 0.15). No veracity by question order effects were found for any of the details (all *F*s < 077, all *p*s > 0.38). Bayes factors showed weak evidence for the null hypothesis for all detail types except for emotion details, which showed weak evidence for the experimental hypothesis (see Tables 6 and 7).

Both studies combined.

Overall, the mean length of the interview was 5 minutes and 59 seconds (*SD* = 1.51). There was a high correlation between the total number of details and the length of the interviews ($r = 0.69$, $p < 0.001$).

When combining the two datasets, the null results remain. For total details, there was no three-way effect between veracity, order and question type (*F* (2.56, 397.40 = 0.15, $p = 0.90$); nor were there any significant two-way effects (all *F*s <2.49, all *p*s > 0.07), except for a question type by interviewer effect (*F* (2.56, 397.40)= 13.34, $p < 0.001$, $\eta^2_p = 0.08$). No significant main effects were found either (all *F*s <2.53, all *p*s > 0.11), except for a significant main effect of Questions type (*F* (3, 397.40)= 63.34, $p < 0.001$, $\eta^2_p = 0.29$). Participants gave the most details in response to travel questions (*M* = 54.08, *SE* = 2.28, *CI* [49.58-58.58]), followed by the core questions (*M* = 38.86, *SE* = 1.90, *CI* [25.114-42.60]), then the most expected general question (*M* = 29.25, *SE* = 1.33, *CI* [26.63-31.87]), with the planning questions eliciting the lowest number of details (*M* = 21.26, *SE* = 1.04, *CI* [19.21-23.32]). The Bayes factor for total details showed weak evidence for the experimental hypothesis for the order main effect, $B_{H(0, 37.83)} = 1.56$, and weak evidence for the null hypothesis for the order by veracity interaction, $B_{H(0, 37.83)} = 0.60$)

Discussion

In both studies the expectedness of the questions was similar to that reported in Warmelink et al. (2012) and Sooniste, et al. (2013). General questions were the most expected, with core, travel and planning questions were less expected. Although the difference between core, travel and planning questions was not consistently significant, the replication of high expectedness of general questions and low expectedness of planning questions suggests that question expectedness is relatively consistent across samples. However, despite the geographical variation in the samples in the literature (Warmelink et al., 2012 and Sooniste, et al., 2013 were run in different countries), they have similar demographics (i.e. WEIRD; Henrich, Heine and Norenzayan, 2010). Whether question expectedness results would be robust across cultures is uncertain.

The results also show that question expectedness is not the only driver of difficulty and the level of detail provided. The expectedness, difficulty and the level of detail provided were not directly related to each other, as may have been expected. Core and travel questions elicit large numbers of details, although they are moderately expected and core questions are considered difficult. Presumably these questions, which explicitly ask for detail about the intention itself, are perceived as difficult, but also lend themselves to a large number of details being given, independently of their unexpectedness.

The studies do not replicate previously reported findings that there is an interaction between question type and veracity, with truth tellers giving more detail than liars to unexpected questions, but not for expected questions (Warmelink et al., 2012; Sooniste et al., 2013). In both current studies, no such effect was found. It is notable that this interaction is also not found in Vrij et al. (2018). Why this effect from the 2012 and 2013 studies does not replicate to these newer studies is unclear. Further replications should be conducted in order to draw conclusions about the existence and size of the effect. Although a significant question type by veracity effect is not necessary to answer our hypotheses about the order of the questions, it may indicate that this dataset differs in some unknown way from the previous studies (Warmelink et al., 2012; Sooniste et al., 2013).

Hypothesis 1 was not supported. In neither study did the order of expected and unexpected questions significantly affect the perceived difficulty of the questions. However, in both studies the questions were perceived to differ in difficulty, although this was not consistent across the studies. In the self-induced study, planning questions were considered the most difficult, while in the other-induced study core questions were. This is perhaps related to the source of the intention: self-induced intentions may be more salient, especially for details about the intention itself, while the core details of an other-induced may be harder to imagine and describe. The planning for the other-induced intentions was completed shortly before the interview, while in the self-induced condition the planning may have been done some time ago, making it harder for the participants to remember.

That question difficulty was not influenced by question order could indicate that asking unexpected questions first does not increase cognitive load. Alternatively, asking participants to rate each question - rather than the whole interview – may not be not the best way to detect cognitive load effects of question order. This may lead the participants to focus on the relative differences between the questions.

Hypothesis 2 was only partly supported. For self-induced intentions, participants in the expected-first order provided more details overall than those in the unexpected first order, although this was not significant and Bayes factors indicated that the data provided only moderate evidence for this hypothesis. This finding was not replicated in the other-induced study, nor was this effect present in the analysis that combined the data from both studies. It may be that this order effect is only present when people are discussing certain topics. Self-induced intentions may be more salient to participants than other-induced intentions, which may have influenced the order effect. This result does indicate that the two paradigms used to study deception about intentions in the literature may not be fully equivalent. Larger effects may be present in paradigms that use self-induced intentions rather than other-induced ones.

Hypothesis 3 was not supported: asking questions in the unexpected-first order did not increase the differences between truth tellers and liars. The results are in line with previous findings that the order of expected and unexpected questions has little or no influence on the amount of information given in an interview (Vrij et al. 2018) and contrast with studies that suggested that asking unexpected question increases lie detection accuracy (Clemens, Granhag and Strömwall, 2011). It is possible that order only affects consistency-based lie detection cues and not detail-based ones.

The current studies have several limitations. One issue is whether the studies were sufficiently powered. The studies provide moderate evidence for total details in Study 1, but in Study 2 Bayes factors provided only weak support for either hypothesis. Again, this points to the possibility that the paradigm used to induce the intention is influencing effects sizes. This underpowering was also influenced by an a priori overestimation of the size of the interaction between veracity and the expectedness of a question, which was based on previous studies (e.g. Warmelink et al., 2012; Sooniste et al., 2013), but was found to be not significant in these studies. Future research can use these results to conduct more accurate a priori power analyses.

A second limitation is that the Bayes factors require the plausibility of different effect sizes (a “prior model”) to be specified. We specified the prior models using the size of the veracity effect in a study that used a very similar interview and coding protocol. While this provided an idea of how much psychological factors can be affected within an interview context, we ideally would have specified the priors based on previous studies of the order effect. Unfortunately these were not available to us, and we hope future research into order effects builds on the results that we have reported here.

In conclusion, the results show that participants in the unexpected-first condition do not consistently differ from participants in the expected-first condition. Most importantly, the unexpected-first order does not increase the differences between liars and truth tellers in a way that benefits lie detection. This suggests that researchers and practitioners who are using unexpected questions for eliciting detail may use them at any point in the interview without detriment.

References

Clemens, F., Granhag, P.A., & Strömwall, L.A. (2013). Counter-Interrogation Strategies when Anticipating Question on Intentions. *Journal of Investigative Psychology and Offender Profiling, 10, 1,* 125-138. DOI: 10.1002/jip.1387.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin*, *129*(1), 74. DOI: 10.1037/0033-2909.129.1.74

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology, 5, 781.* DOI: 10.3389/fpsyg.2014.00781

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin and Review, 25, 1,* 207-218. DOI: 10.3758/s13423-017-1266-z

Elntib, S., Wagstaff, G. F., & Wheatcroft, J. M. (2015). The Role of Account Length in Detecting Deception in Written and Orally Produced Autobiographical Accounts using Reality Monitoring. *Journal of Investigative Psychology and Offender Profiling*, *12*(2), 185–198. https://doi.org/10.1002/jip.1420

Ewens, S., Vrij, A., Leal, S., Mann, S., Jo, E., & Fisher, R.P. (2016). The effect of interpreters on eliciting information, cues to deceit and report. *Legal and Criminological Psychology, 21,* 286-304. DOI: 10.1111/lcrp.12067

Fenn, E., McGuire, M., Langleben, S., and Blandón-Gitlin, I. (in press). A reverse order interview does not aid deception detection regarding intentions. *Frontiers in psychology, ?.* DOI: 10.3389/fpsyg.2015.01298

Fisher, R.P. (1995). Interviewing Victims and Witnesses of Crime. *Psychology, Public Policy, and Law, 1, 4,* 732-764. DOI: 10.1037/1076-8971.1.4.732

Granhag, P. A., & Mac Giolla, E. (2014). Preventing Future Crimes: Identifying Markers of True and False Intent. *European Psychologist*, *19*(3), 195–206. https://doi.org/10.1027/1016-9040/a000202

Henrich, J., Heine, S.J., & Norenzayan, A. (2010) Most people are not WEIRD. *Nature*, *446,* 29, DOI: 10.1038/466029a.

Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, *7*(1). https://doi.org/10.7771/1932-6246.1167

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology*, *32*(3), 354–366. https://doi.org/10.1002/acp.3407

Lancaster, G.L.J., Vrij, A. Hope, L., & Waller, B. (2013). Sorting the Liars from the Truth Tellers: The Benefits of Asking Unanticipated Question on Lie Detection. *Applied Cognitive Psychology, 27, 1,* 107-114. DOI: 10.1002/acp.2879.

Mac Giolla, E., & Granhag, P.A. (2015). Detecting False Intent Amongst Small Cells of Suspects: Single Versus Repeated Interviews. *Journal of Investigative Psychology and Offender Profiling, 12, 2,* 142-157. DOI: 10.1002/jip.1419

Nahari, G. & Vrij, A. (2013). Are you as good as me at telling a story? Individual differences in interpersonal reality monitoring. *Psychology, Crime & Law, 20, 6,* 573-583. OI: 10.1080/1068316x.2013.793771

Nahari, G., Ashkenazi, T., Fisher, R.P., Granhag, P.A., Hershkowitz, I., Masip, J., Meijer, E., Nisin, Z., Sarid, N., Taylor, P.J., Verschuere, B., & Vrij, A. (2019).

Language of lies: Urgent issues and prospects in verbal lie detection research. *Legal and Criminological Psychology, 24,* 1-23. DOI: 10.1111/lcrp.12148

Sooniste, T., Granhag, P.A., Knieps, M., & Vrij, A. (2013). True and false intentions: Asking about the past to detect lies about the future. *Psychology, Crime & Law, 19, 8,* 673-685. DOI: 10.1080/1068316X.2013.793333

Vierra, A.J, & Garret, JM. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine, 37, 5,* 360-363.

Vrij, A., Fisher, R.P., & Blank, H. (2017). A cognitive approach to lie detection: a meta-analysis. *Legal and Criminological Psychology, 22, 1.* 1-21. DOI: 10.1111/2Flcrp.12088

Vrij, A., Fisher, R.P., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *TRENDS in Cognitive Sciences, 10, 4,* 141-142. DOI: 10.1016/j.tics.2006.02.003.

Vrij, A., Mann, S., Kristen, S., & Fisher, R.P. (2007). Cues to Deception and Ability to Detect Lies as a Function of Police Interview Styles. *Law and Human Behavior, 31,* 499-518. DOI: 10.1007/s10979-006-9066-4.

Vrij, A., Leal, S., Mann, S., Fisher, R. P., Dalton, G., Jo, E., … Houston, K. (2018). Using unexpected questions to elicit information and cues to deceit in interpreter-based interviews. *Applied Cognitive Psychology, 32*(1), 94–104. https://doi.org/10.1002/acp.3382

Vrij, A., Leal, S., Granhag, P.A., Mann, S., Fisher, R.P., Hillman, J., & Sperry, K. (2009). Outsmarting the Liars: The benefit of Asking Unanticipated Questions. *Law and Human Behavior, 33, 2,* 159-166. DOI: 10.1007/s10989-008-9143-y

Warmelink, L. (2012). *Lying about Intentions.* (Unpublished doctoral thesis). University of Portsmouth, Portsmouth, UK.

Warmelink, L., Vrij, A., Mann, S., Jundi, S., & Granhag, P.A. (2012). The effect of question expectedness and experience on lying about intentions. *Acta Psychologica*, *114*, 178-183. DOI: 10.1016/j.actpsy.2012.07.011.

Warmelink, L., Vrij, A., Mann, S., & Granhag, P.A. (2013). Spatial and Temporal Details in Intentions: A Cue to Deception. *Applied Cognitive Psychology*, *27*, 101-106. DOI: 10.1002/acp.2878

X (2017). *The Implicit Association Test as a cue to deception about intentions.* Manuscript submitted for publication.

Table 1

Coding Information and Reliability per Detail

Detail	Definition	Example	Correlation	ICC (95% Confidence interval)
Visual[1]	Details that are seen	The *grey staircase* (2 Visual details)	.89*	.93 (.86-.97)
Auditory[1]	Details that are heard	The *creaky* staircase	.81*	.87 (.74-94)
Smell/taste[1]	Details that are smelt or tasted	It smells of *chips*	.75*	.86 (.71-93)
Tactile[1]	Details that are felt	It's *cold*	.80*	.83 (.66-.92)
Action[1]	Details about the participants' actions	I'll *walk* there	.94*	.96 (.92-.98)
Spatial[2]	Details about spatial arrangements	*Next to* the store	.92*	.87 (.73-.93)
Temporal[2]	Details about temporal order	*First*, I'll walk there	.92*	.96 (.92-.98)
Emotion	Emotional details	I'm *worried*	.73*	.84 (.67-.92)
Hedge	Participants expressing uncertainty	I'll *probably* walk there	.80*	.76 (.51-.88)
Knowledge	Details the participants knows, but doesn't experience	I have to spend *as close as possible* to £20	.49*	.65 (.29-.84)

Note. * indicates $p < .01$. [1] indicates perceptual details. [2] indicates contextual details.

Table 2: expectedness, difficulty, and detail per question type for study 1: self-induced.

Question type	Expectedness Mean (SE), CI	Difficulty Mean (SE), CI	Number of details Mean (SE), CI
General	8.47 (0.28), *CI* [8.17-9.30]	5.82 (0.31), *CI* [5.21-6.43]	30.26 (2.14), *CI* [25.99-34.54]
Core	7.07 (0.30), *CI* [6.47-7.67]	6.92 (0.32), *CI* [6.28-7.56])	41.02 (3.06), *CI* [34.92-47.12]
Travel	6.55 (0.29), *CI* [5.98-7.11]	6.05 (0.29), *CI* [5.46-6.62]	52.99 (3.52), *CI* [45.97-60.01]
Planning	5.80 (0.33), *CI* [5.14-6.46]	7.48 (0.37), *CI* [6.75-8.22]	22.18 (1.62), *CI* [18.95-25.42]

Table 3

Estimated marginal means, standard errors and 95 % confidence intervals for total details in study 1: self-induced

Veracity	Order	Question type	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Truth	Expected-first	General	28.93	4.28	20.40	37.46
		Core	34.28	6.11	22.10	46.45
		Travel	51.58	7.03	37.56	65.59
		Planning	24.32	3.24	17.86	30.78
	Unexpected-first	General	28.51	4.16	20.23	36.79
		Core	32.12	5.93	20.30	43.94
		Travel	43.10	6.83	29.50	56.70
		Planning	18.40	3.15	12.13	24.67
Lie	Expected-first	General	30.06	4.40	21.28	38.83
		Core	56.09	6.29	43.57	68.62
		Travel	69.95	7.23	55.54	84.37
		Planning	26.27	3.34	19.63	32.92
	Unexpected-first	General	33.56	4.37	24.85	42.27
		Core	41.58	6.24	29.15	54.02
		Travel	47.33	7.18	33.02	61.64
		Planning	19.74	3.31	13.14	26.33

Table 4

P-values and Bayes factors (Mean of theory, SD of theory) for all details in Study 1: self-induced.

Detail	P-value Order Main effect	Bayes factor Order main effect	P value Order by Veracity effect	Bayes factor Order by Veracity effect
Total	0.06	$B_{H(0, 37.83)} = 3.58$	0.44	$B_{H(0, 37.83)} = 1.13$
Perceptual	0.11	$B_{H(0, 13.59)} = 2.46$	0.52	$B_{H(0, 13.59)} = 1.14$
Contextual	0.06	$B_{H(0, 7.05)} = 3.64$	0.60	$B_{H(0, 7.05)} = 1.09$
Emotion	0.78	$B_{H(0, 0.19)} = 1.07$	0.35	$B_{H(0, 0.19)} = 1.05$
Hedge	0.22	$B_{H(0, 1.58)} = 1.43$	0.06	$B_{H(0, 1.58)} = 2.15$

Table 5

Expectedness, Difficulty and detail per question type for study 2: other-induced.

Question type	Expectedness Mean (SE), [CI]	Difficulty Mean (SE), [CI]	Number of details Mean (SE), [CI]
General	7.48 (0.30), *CI* [6.88-8.08]	6.41 (0.24), *CI* [5.93-6.89]	28.35 (1.63), *CI* [25.10-31.59]
Core	6.22 (0.26), *CI* [5.70-6.75]	7.51 (0.28), *CI* [6.96-8.06]	37.17 (2.24), *CI* [32.70-41.64]
Travel	6.35 (0.30), *CI* [5.75-6.95]	6.23 (0.28). *CI* [5.68-6.77]	55.47 (2.85), *CI* [49.80-61.15]
Planning	6.31 (0.30.), *CI* [5.71-6.91]	6.38 (0.25), *CI* [5.88-6.87]	20.43 (1.32), *CI* [17.80-23.05]

Table 6

Estimated marginal means, standard errors and 95 % confidence intervals for total details in study 2: other-induced

Veracity	Order	Question type	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Truth	Expected-first	General	26.15	3.46	19.25	33.05
		Core	39.73	4.77	30.22	49.23
		Travel	58.95	6.07	46.87	71.02
		Planning	21.63	2.81	16.04	27.21
	Unexpected-first	General	32.51	3.28	25.98	39.03
		Core	34.26	4.52	25.26	43.25
		Travel	51.65	5.74	40.22	63.08
		Planning	19.63	2.66	14.35	24.92
Lie	Expected-first	General	28.00	3.12	21.79	34.21
		Core	38.19	4.30	29.63	46.76
		Travel	53.82	5.46	42.94	64.70
		Planning	19.62	2.53	14.59	24.65
	Unexpected-first	General	26.72	3.19	20.37	33.08
		Core	36.51	4.40	27.75	45.26
		Travel	57.48	5.59	46.35	68.60
		Planning	20.83	2.58	15.68	25.98

Table 7

P-values and Bayes factors (Mean of theory, SD of theory) for all details in Study 2: other induced.

Detail	P-value Order Main effect	Bayes factor Order main effect	P value Order by Veracity effect	Bayes factor Order by Veracity effect

Total	0.06	$B_{H(0, 37.83)} = 0.36$	0.44	$B_{H(0, 37.83)} = 0.49$
Perceptual	0.77	$B_{H(0, 13.59)} = 0.66$	0.67	$B_{H(0, 13.59)} = 0.74$
Contextual	0.96	$B_{H(0, 7.05)} = 0.44$	0.69	$B_{H(0, 7.05)} = 0.59$
Emotion	0.28	$B_{H(0, 0.19)} = 0.60$	0.38	$B_{H(0, 0.19)} = 1.29$
Hedges	0.01	$B_{H(0, 1.58)} = 0.21$	0.42	$B_{H(0, 1.58)} = 0.52$