An investigation of the detectability of false intent about flying

Kleinberg, B.; Nahari, G.; Arntz, A.; Verschuere, B.

DOI
10.17605/OSF.IO/XX397

Publication date
2017

Document Version
Submitted manuscript

TITLE PAGE

TITLE: An investigation of the detectability of false intent about flying

AUTHORS: Bennett Kleinberg[1], Galit Nahari[2], Arnoud Arntz[1], Bruno Verschuere[1]

AFFILIATIONS:
[1]Department of Psychology, University of Amsterdam, The Netherlands.
[2]Department of Criminology, Bar-Ilan University, Ramat Gan, Israel.

KEYWORDS: deception detection; false intentions; Reality Monitoring; airport screening; information protocol

WORD COUNT ABSTRACT: 218

WORD COUNT MANUSCRIPT (excl. References and Appendix): 8,284

Abstract

Background: Academic research on deception detection has largely focused on the detection of past events. For many applied purposes, however, the detection of false reports about someone's intention merits attention. Based upon the cognitive load perspective, we explored whether true statements on intentions were more detailed and more specific than false statements on intentions, particularly when instructed to be as specific as possible.

Method: Participants (n = 354) either lied or told the truth about their upcoming flight plans, either providing 'as much information as possible' (standard instructions) or being 'as specific as possible' (specific instructions), resulting in four conditions (truthful vs. false intention and standard instructions vs. specific instructions). We collected data via a custom-made web app and performed automated, algorithmic verbal content analysis of participants' written answers.

Findings: We did not find a significant difference in participants' statements specificity. The instruction to be as specific as possible promoted more specific information but did not help to discern honest from deceptive flying intentions.

Conclusion: The experiment reported here attempted to demonstrate automated detection of verbal deception of intentions. The difficulty in capturing genuine, autobiographical implementation-intentions, and the non-intrusive, non-interactive questioning approach might explain the null findings and raise questions for further research towards large-scale applicability. We conclude with suggestions for a novel framework on semi-interactive information elicitation.

1. Introduction

With an increased demand for security systems like airport border control, researchers and practitioners alike have identified the need for applications to detect deception at large-scale (e.g. Vrij, Granhag, & Porter, 2010; Honts & Hartwig, 2014). For example, the context of airport border control excludes many tools used in deception research due to their limited applicability. A promising paradigm of deception research is the cognition-based approach to deception based on the early observation that lying is cognitively more demanding than telling the truth (Zuckerman, DePaulo, & Rosenthal, 1981; Vrij, Fisher, & Blank, 2015). Several techniques have been developed to increase the difference in cognitive demand for liars to induce leakage of cues to deception (Vrij et al., 2015). Since the majority of studies within the cognitive deception paradigm was conducted on the verbal content of face-to-face interviews, a key challenge is a transition towards large-scale applicable methods. This paper reports a first attempt to apply verbal cognitive deception detection tools on large-scale in an airport security context.

1.1 Cognition-based deception detection

From the cognitive perspective on deception detection, liars have to cope with a more challenging situation when being interviewed as a suspect than truth-tellers. While a truth-teller can simply remember an event, a liar will have to convince the interviewer of a false story (Vrij et al., 2010). In doing so, the liar has to maintain a plausible (fabricated) story while carefully avoiding any inconsistencies or non-believable elements. For example, if a liar aims to convince the interviewer that they were at a concert and not at the crime scene, they must take care to provide a credible account of their concert alibi story. The credible alibi implies that the liar will have to avoid any inconsistencies or too vague a description of the concert. Consequently, in a suspect interview, the task is more difficult for someone pretending to have been at the concert compared to someone else who truly was at that concert. The rationale of the cognitive approach is to increase the differences in cognitive load between liars and truth-tellers further through the use of interviewing techniques.

Vrij et al. (2015) provided a meta-analytical overview of cognition-based deception detection identifying three key techniques that can be used to increase the differences between truth-tellers and liars. First, imposing additional cognitive load means to create an interviewing situation that is cognitively more difficult for the liar than for the truth-teller (e.g. asking suspects to recall events in reverse order; Evans, Meissner, Michael, & Brandon, 2013). Similarly, administering an additional, simultaneous task during a suspect interview (e.g. holding an object or maintaining eye-contact, Vrij et al., 2015) is thought to require cognitive resources, which the liar has less of than the truth-teller. Second, asking unexpected questions has been used to exploit the differences between liars and truth tellers differently (e.g. Warmelink, Vrij, Mann, Leal, & Poletiek, 2011). Assuming that liars prepare for an interview more extensively than truth tellers, the interview can be designed to make use of this preparation difference. For example, a liar may prepare for questions like 'Where have you been

yesterday?' But they might not be equally prepared to answer questions like 'What did the spatial arrangement look like in the cafe?'. Third, demanding interviewees to provide detailed information creates an interview situation that is more difficult for the liar than for the truth teller. The more information is provided, the more likely it is that the false account becomes incoherent, implausible or unbelievable due to the lack of details (Nahari, Vrij, & Fisher, 2014a).

Meta-analytical findings (Vrij et al., 2015) supported cognitive load-based deception detection. Interviews that applied at least one of the three techniques had an average overall detection accuracy (i.e. truth-tellers and liars combined) of 71% compared to 56% without cognitive load interviewing techniques.[1] The highest accuracy was found for interviews where participants were encouraged to provide much information. Using these three techniques, it is especially the verbal cues to deception that become more apparent.

1.2 Verbal cues to deception

The idea to use verbal content as an indicator of deception is famously rooted in the Undeutsch Hypothesis (1967, 1982) stating that truthful statements differ from false declarations in quality and content because the process through which the particular statement is produced is different (see Fornaciari & Poesio, 2013). Research by Johnson and Raye (1981; Masip, Sporer, Garrido, & Herrero, 2005) has specified further that the source of one's memory determines how a remembered event is recalled. Genuine memories are obtained through sensory experiences whereas non-genuine memories are constructed through cognitive operations. Therefore, the narratives of these memories should differ so that descriptions of genuine memories should be richer in sensory experiences (e.g. perceptual, spatial, temporal information), whereas non-genuine memories should contain more references to cognitive operations (Johnson & Raye, 1981; Masip et al., 2005). Reality Monitoring (RM) is a theoretical and analytical framework that incorporates this idea. Parallel to genuine and non-genuine memories, truthful statements are expected to contain more references to sensory information compared to false statements since truthfully experienced events are supposed to be retrieved with more detail (also labeled as Interpersonal Reality Monitoring, see Nahari & Vrij, 2014; Johnson, Bush, & Mitchell, 1998). Masip et al. (2005) have examined how RM performs for the detection of deception across multiple studies. Their findings indicate that RM is a tool suitable for deception detection in verbal statements. Especially the amount of temporal, spatial and perceptual detail has been found to be higher in truth tellers' statements compared to those of liars. Overall accuracy rates of RM in laboratory studies seem to be around 72% (Vrij, 2008).

[1] The base rates in the studies included in that meta-analysis are typically 0.5 for truth-tellers and liars. Note that such a base rate is fundamentally different from an applied airport passenger screening context (see Honts & Hartwig, 2014).

Recently, Nahari et al. (2014a, 2014b) introduced the Verifiability Approach (VA) and suggested that there might be an additional dimension to the amount of details, namely the verifiability of details. The VA is based directly on strategies used by liars to provide a reasonable, false account. During an interview, the liar faces the dilemma between being inclined to describe an event in sufficient detail to sound convincing and at the same time avoiding that information that could potentially be verified (Nahari, Vrij, & Fisher, 2012). For example, an answer like 'I spoke to my friend James in the Vondelpark' might be a detail that theoretically could be investigated further by the interviewer (e.g. by consulting James and asking for confirmation), whereas 'I spoke to someone in the park' would not count as a verifiable detail. Derived from this rationale, a series of studies led to the establishment of VA (e.g., Harvey, Vrij, Nahari, & Ludwig, 2016; Nahari & Vrij, 2014), showing that the number of verifiable details discriminates liars from truth teller. For example, Harvey et al. (2016) showed that the number of verifiable details is higher in truthful insurance claims than in deceptive ones when both liars and truth tellers are instructed to mention as much verifiable information as possible. The proportion of verifiable details yielded an overall accuracy rate of 80%. The working definition (see Nahari et al., 2014) of verifiable information includes any activity that i) has been done with an identifiable person, ii) has been witnessed by an identifiable person, or iii) has been documented or recorded through technology (e.g. CCTV, email, social networks). Whereas the VA, like most verbal deception detection research, has predominantly been limited to the detection of past events, recent developments have called for research on deceptive intentions.

1.3 Detecting false intentions

Whereas academic deception research focused on the detection of deception about recent events, for many practical purposes in law enforcement and intelligence services, it is becoming increasingly important to detect people with potentially malicious intent to prevent crimes from happening (Vrij & Granhag, 2012). For example, in border control settings, it might be more important to determine what someone is planning to do upon entering a country rather than learning what they have been doing before coming to border control. Recently, researchers have begun to shift the temporal dimension of research paradigms towards that of intentions (e.g. Sooniste et al., 2015; MacGiolla et al., 2016; see Granhag, 2010). Most of the studies conducted in this domain adopted the cognitive deception paradigm. For example, Sooniste, Granhag, Knieps and Vrij (2013) developed an experimental model designed to have participants lie about their intentions.

In their experiment, Sooniste et al. (2013) instructed half of their participants to plan and enact an innocent activity in a shopping mall (i.e. buy two gifts). The other half was told to prepare and enact a mock-crime (i.e. placing a USB stick illegally in a shop in the same shopping mall). Those instructed to plan the mock-crime were also told to develop a cover story similar to that of the innocent participants. Before any of the participants enacted their assigned task in the shopping mall, they were intercepted and interviewed about their intended behavior in the mall. Before the interview, liars were

told to hide their true intentions, so that each interviewee tried to convince the interviewer of having planned the innocent activity. Most importantly, during the interview, Sooniste et al. (2013) asked questions about the planning phase and the intentions, reasoning that the former were less expected and hence more diagnostic than the latter. They found that truth tellers' answers to planning-related questions were rated as more detailed than those of liars, whereas there was no such difference for intentions-related questions. In a related study, Vrij, Granhag, Mann, and Leal (2011) asked airport passengers in an in-vivo semi-experimental setup to either lie or tell the truth about their upcoming trip. They found that truth tellers' answers were rated as more plausible than liars' answers but did not differ on the perceived amount of detail.

1.4 The current investigation

The aim of the present study was to examine whether we could identify people lying about their intentions of traveling by airplane. To work towards potentially large-scale applicable methods of deception detection within the promising cognitive deception paradigm, we built an online platform where we asked questions about people's upcoming flight plans. To elicit information similar to face-to-face interview studies, we stayed close to the questions asked in previously successful studies on cognition-based deception detection about intentions (e.g. Warmelink et al., 2011, 2013; Sooniste et al., 2013). In addition to being able to collect data on a large scale, another requirement for implementable tools is an automated analytical framework. Here, we aimed to address this by analyzing verbal content both computer-automated as well as employing human coding. Specifically, participants in our study were instructed to either tell the truth or lie about their future or most recent flight and were subsequently asked ten questions about this trip. Besides, we manipulated the way in which we asked questions. Studies by Nahari et al. (Nahari et al., 2014; Harvey et al., 2016) have shown that informing both liars and truth tellers about the expected verifiable information in truthful answers may benefit deception detection (but see Nahari & Pazuelo, 2015). We instructed half of the liars and truth tellers in each condition to provide as much information as possible and the other half to provide highly specific information (e.g. names of persons or locations, dates). Nahari et al. (2014) reasoned that the latter would pose a difficulty to liars, who might avoid providing damaging information, whereas truth tellers could quickly provide verifiable information. This manipulation is thus a modification of the Verifiability Approach specifically targeted at computer-automated analysis via named entity recognition. Named entity recognition is a method originating from computational linguistics used to extract information from text. Specifically, information is classified into pre-defined categories (e.g., persons, locations, dates; see Kleinberg et al., 2016), whereby a mixture of methods to perform that classification is used (e.g. machine learning, dictionaries, and rule-based classification).

1.5 Hypotheses

Our primary hypothesis in this study was motivated by findings from deception research on both past events (e.g. Nahari et al., 2014) and intentions (e.g. Mann et al., 2011, Warmelink et al., 2013): Truthtellers' accounts contain more detailed information than those of liars (Detailedness Hypothesis). Similar to Nahari et al.'s (2014) information protocol procedure, we further hypothesized that truth tellers can provide more specific information than liars if they are explicitly told to do so (Information Protocol Hypothesis)For exploratory purposes, we were interested in investigating i) human coded variables for differentiating truthful from deceptive statements that may be harder to automatize (e.g. plausibility); ii) how the question type affected the detailedness (i.e., questions about the planning phase may be more suitable to evoke lie-truth differences than questions about the intentions themselves, see Sooniste et al., 2013), iii) how detailedness differed between past events and future intentions, and iv) how the temporal immediacy of flight plans moderated the effect of detailedness.

2. Method

2.1 Participants

We aimed to collect data from 518 participants based on a priori power analysis for the 2 (Veracity: truthful vs. deceptive) by 2 (Information Protocol: standard vs specific) interaction for Cohen's f = 0.25, power = .95, alpha = .05.[2] We opened spots for participation on the online platform crowdflower.com until this number of participants was reached. Of the initial sample of 518 participants, there were no data for nine participants, and we further excluded data of participants whose IP address has been registered more than once, resulting in an additional 94 exclusions (for similar exclusion criteria, see Kleinberg & Verschuere, 2015; Verschuere et al., 2015). The relatively high exclusion number for double IP addresses here might be due to the block-wise data collection that made double participation possible. From the remaining sample (n = 415), we excluded participants who were outliers (larger than 2.5 SD above the mean) on the number of weeks until their flight and the number of times having visited the destination before (n = 33) and those who indicated to not have provided genuine information at the beginning (n = 28), resulting in 354 participants for analysis.[3] Participants were randomly allocated to the truthful or deceptive condition and further to the standard or specific information protocol condition. Of those who were going to fly (n = 222)[4], 109 participants were in the truthful condition (standard: n = 49, M_{age} = 32.51, SD = 9.10, 32.65% female; specific: n =

[2] Note that power calculations indicate a sample size of 210 for the given parameters (i.e. 53 per cell in a 2x2 betw.-subjects design). We deliberately added 20% to this number for potential drop-out (n per cell: 64) in this online study (see Kleinberg & Verschuere, 2015; 2016) and multiplied this number by eight given that the overall design was 2x2x2 betw.-subjects.

[3] Post-hoc power calculations showed that the final sample size of 354 was indeed sufficient to pick up an effect of f = 0.25; achieved power: 0.99.

[4] Achieved power for f = 0.25 and n = 222: 0.96

60, M_{age} = 35.78, *SD* = 9.35, 41.94% female), and 103 were in the deceptive condition (standard: *n* = 52, M_{age} = 33.90, *SD* = 10.21, 34.54% female; specific: *n* = 61, M_{age} = 32.25, *SD* = 7.61, 34.43% female). For demographics on the non-flyer group see Appendix.

2.2 Materials

2.2.1 Experimental task

The experimental task was presented in a custom-made web app programmed in HTML and JavaScript. Participants needed an Internet connection and a standard web browser to do the task on their computer. To ensure additional control over the experimental task, we disabled the translation function, the copy-and-paste function, tested whether the input was provided where necessary, whether text input was real English, and whether the required minimum length of answers was provided. If not, the participants were informed about this via a pop-up and alerted that this could result in invalidation of their participation. The original task is accessible via this link: tinyurl.com/jny6p9w and the source code of the web app are obtainable via this link (later). Also, the data collected in this study is available in the Open Science Framework repository (later) in anonymized form.

2.2.2 Computer-automated analysis

2.2.2.1 Detailedness: Linguistic Inquiry and Word Count software

Many studies that adopted a computer-automated approach to verbal deception detection have used the Linguistic Inquiry and Word Count software (LIWC, Pennebaker et al., 2015; Mihalcea & Strappavara, 2009). Text statements processed with LIWC return proportions of word categories occurring in the text. Each word category is intended to model different psycholinguistic variables such as the category 'affect' which is used to model emotional processes. Underlying each category are extensive dictionaries of words against which the words in the statements are analyzed. LIWC has successfully been employed in multiple contexts (Ott et al., 2011, 2013) and was shown to be acceptable for modeling human-coded RM annotation (Bond & Lee, 2005). In the current investigation, we used the LIWC word categories “percept,” “space” and “time,” to model perceptual, spatial and temporal details, respectively. For each participant, we summed the three categories across all ten answers to derive the dependent variable detailedness.

2.2.2.2 Sentence specificity: Speciteller

Motivated by the observation that two sentences can contain the same propositional meaning but convey that information with different degrees of specificity, Li and Nenkova (2015) developed speciteller. Speciteller is a python-implemented machine learning-based classifier giving the specificity of a sentence ranging from 0 (lowest) to 1 (highest). Li and Nenkova (2015) had five independent annotators code a sample of 885 sentences from the Wall Street Journal, New York Times, and Associated Press. The coders determined that 54.58% of the sample were specific

sentences which were then used to build a classifier with shallow surface features (e.g. number of words, estimated number of named entities) and dictionary features (e.g. subjective words, concreteness). Using machine learning (supervised logistic regression, semi-supervised and co-training classification), they derived a final classifier model they released as open-source software under the name speciteller. We modified the speciteller tool to automatically preprocess bunches of text statements to calculate the average sentence specificity per statement as a dependent variable.

2.2.2.3 Information specificity: Named Entity Recognition

We operationalize information specificity as the number of named entities recognized by the SpaCy python natural language processing tool. To be able to extract information from text, a technique from computational linguistics detects the occurrence of so-called named entities. Named entities refer to specific information that falls into one of the multiple categories of persons, places or dates. In general, the approach to developing a named entity recognition (NER) system is to define grammar-based rules, regular expressions and machine learning classification to identify entities in text automatically. In this investigation, we use the NER of the Python library SpaCy (Honnibol, 2016) to extract named entities in the categories persons, nationalities or religious groups, facilities, organizations, geopolitical entities, locations, products, events, works of art, law documents, languages, dates, times, percentages, quantities, ordinals, and cardinals (e.g. Kleinberg et al., 2016). Pre-processing was limited to standard tokenization. We obtained the dependent variable information specificity by summing all named entities per statement divided by the number of words.

2.2.3 Human coding

In addition to automated report coding, two students coded the declarations of those participants that were flying in no more than four weeks (n = 110). The coders were trained in rating the statements on detailedness ("the inclusion of specific descriptions of place, time, persons, objects and events in the statement"), plausibility ("the coherency of the statement in terms of not containing logical inconsistencies or contradictions and the degree to which the message seems plausible, likely, or believable."), complications ("the reporting of either an unforeseen interruption or difficulty, or spontaneous termination of the event"), occurrence of how-utterances ("concrete descriptions of activities"), occurrence of why-utterances ("first, wider motivations/reasons why someone planned an activity; second, motivations/reasons for doing something in a certain way"), and truthfulness. Coders were presented with all ten answers per participant and were instructed to make an overall judgment rather than rate the separate answers. All variables were scored on a 7-point Likert scale (1 = very low; 7 = very high). Definitions were adopted from Vrij (2015) and MacGiolla et al. (2013). Both coders received a 2.5-hour training session on statements of non-flying participants and a subset (n = 16) of the selected statements which were excluded from the analysis. Of the remaining 94 statements (48 truthful, 43 deceptive), 31 statements were coded by both coders (ICCs: plausibility = 0.67,

detailedness = 0.85, how-utterances = 0.36[5], why-utterances = 0.86, complications = 0.71, truthfulness = 0.82). The remaining 63 statements were randomly distributed between the two coders.

2.3 Procedure

The experimental task was advertised on crowdflower.com as a survey about people's flying behavior. Upon accessing the custom-made web app via a link provided in the task description (tinyurl.com/jny6p9w) participants were introduced to the task and told that serious participation was necessary and would be rewarded with the chance of winning a $100,- Amazon.com voucher.[6] After giving informed consent, on the next page, all participants were asked whether they would be flying in the next three months (answer options: "yes", "no", "not sure yet"). If they indicated that they would fly in the next three months, the next page asked the following flight-related questions; i) how many weeks this flight was away, ii) what the purpose of this flight was (pre-defined selection menu, e.g. "work"; see Appendix), iii) what the final destination of their trip was (e.g. "London", and iv) how often they had visited that place before. For all pages where any input was required, participants could only proceed after providing the required information. The next page specific task instructions: Participants were either instructed to lie or to tell the truth about their flight. Those who were in the truthful condition were told to provide honest answers about their trip to (say) "London for work." Those who were allocated to the deceptive condition were assigned a new destination (e.g. "Madrid") and a new purpose (e.g. "holiday") and were told to pretend they are planning to fly to this new destination with the new purpose.

Besides, in both conditions, we told participants that they should either provide as much information as possible on the next ten questions about their flight or that they should provide as specific information as possible (e.g. names, locations, dates). The instructions were repeated in bullet points on the next page, and all participants had 30 seconds to prepare for the upcoming questions. In total, all participants answered eleven questions including one test question ("Please describe your task for this experiment") to get help participants become acquainted with the task (see Table 1). The remaining ten questions were identical to all participants whereby the destination and purpose were filled in according to the participants' experimental condition and their assigned destination/purpose pair. During all questions, below the actual wording of the questions, the instructions regarding the veracity and information protocol manipulation were repeated (e.g. "Remember: please lie about your original trip by giving very specific information (persons, locations, times, etc.) about a trip to Madrid for a holiday". Questions were presented one at a time in predefined order.

[5] Due to the low ICC of how-utterances, we did not include this variable further into the analysis.

[6] Of all participants that provided their email address for this, one was chosen at random and was awarded the voucher.

After typing in the answers to these questions, we asked for demographic variables (age, gender, education, country of origin, native language) and asked for each question, how expected they found it on a Likert scale from 1 (not expected at all) to 10 (absolutely expected). Also, we asked how motivated they were and had them rate their language proficiency as well as doing two language assessment tasks which were part of another study and are not reported here. Those participants who indicated at the beginning that they were not flying in the next three months proceeded through the same task but answered all flight-related questions about their most recent past flight. The truthful/deceptive manipulation was adjusted accordingly (i.e. answer truthfully or lie about the last past trip). The wording of the questions changed automatically.

At the end of the task, as a control check, participants were asked whether they provided accurate information at the beginning of the task regarding their upcoming or past flight (answer options: "yes", "no"). Participants were then debriefed and could provide their email address for the draw on the $100, - voucher. The task took approx. 15min.

2.4 Experimental manipulations

There were two experimental manipulations as well as one semi-experimental manipulation in this study. First, we manipulated the veracity of people's answers by allocating them to either the truthful or the deceptive condition. If participants indicated that the purpose of their upcoming flight would be returning home, they responded to questions about their past trip. Those in the truthful condition (for both past and upcoming trip) were asked questions about the self-reported destination and purpose whereas those in the deceptive condition were allocated a different destination/purpose pair. This allocation ensured that neither the purpose nor the destination for liars matched the original one. We further attempted to apply a semi-yoked matched design by randomly allocating a destination/purpose pair that genuine flyers reported in pilot studies. Second, we manipulated the information protocol by changing the additional instructions to answer the questions. Those in the standard information protocol condition were told to provide as much information as possible, whereas those in the specific information protocol condition were told to provide as much specific information as possible. The latter also received examples of what specific information was (names, times, locations, etc.). Third, the semi-experimental manipulation was the temporal focus of flying (past flight or upcoming flight) that was self-reported by participants.

Table 1. *Questions asked in the experimental task (verbatim).*

#	Question	Rationale/ reference	Minimum length (characters)	Example of given answer[7]
1	“This is a test question and a check whether you understood the instructions. Please briefly state your task in this experiment.”	Control question	15	"To accurately provide information on my trip to London to visit family. I plan to fly to London to visit family and friends. I will also be traveling to Brighton and maybe Southampton."
2	“What is the main purpose of your flight to [DESTINATION]?”	General question (Warmelink et al., 2013)	50	"The main purpose is to visit family in London. I will also be going to Brighton."
3	“Who will you meet in [DESTINATION] and for which reason?”	General question	50	" I will be visiting family that live in London. I will visit some friends as well."
4	“Please describe in which order you did the planning for your trip to [DESTINATION]. What was first, what second, and what last?”	Planning question (Warmelink et al., 2013)	50	"The first thing I had to do was check the flights to London. The second was to book the flights according to my schedule."
5	“What was the hardest to plan?”	Planning question (Warmelink et al., 2013)	50	"The hardest thing to plan was booking a hotel. There are so many hotels with so many reviews. It was difficult to choose one and pick the location."
6	“What is the most pleasant event you expect to happen during your trip?”	Emotion-related question (Hauch et al., 2015)	50	"The most pleasant event that I expect to happen during my trip is to see family that I haven't seen in a couple of years."
7	“What is the most unpleasant event you expect to happen during your trip?”	Emotion-related question (Hauch et al., 2015)	50	"The most unpleasant event will likely be the travelling part. I will be departing at 6am, so it is likely to be an early morning."
8	“If you have to wait during your journey, for example in the airport or changing train stations, what will you do while you're waiting?”	Transportation question (Warmelink et al., 2012)	10	"While waiting on my journey, I will likely be on my phone or laptop."
9	“How will you get from the airport to your accommodation?”	Transportation question (Warmelink et al., 2012)	10	"I will travel from the airport to my accommodation via rental car."
10	“What is the first thing you will do when you arrive at your final destination?”	Other specific question	50	"The first thing I will do when I arrive will be to look for a Starbucks."
11	“What is the first thing you will do when you return home from your trip to [DESTINATION]?”	Other specific questions	50	"The first thing I will do when I return home is unpack and shower."

Note. The tense of the question was adjusted when the questions pertained to the most recent past flight.

[7] The locations and times are anonymized here. The examples are taken from a participant in the truthful condition.

2.5 Analytical plan

Although the full design of this study was 2 (Temporal focus: upcoming vs. past flight, between-subjects) by 2 (Veracity: truthful vs. deceptive, between-subjects) by 2 (Information Protocol: standard vs. specific, between-subjects), the primary aim of the analysis were those participants who went on an upcoming flight. Therefore, for the main hypotheses tested, the particular design was 2 (Veracity: truthful vs. deceptive, between-subjects) by 2 (Information Protocol: standard vs. specific, between-subjects). As the dependent variable, we tested detailedness, average sentence specificity, and information specificity in the written answers.

Also, in exploratory analyses, we provide human coding of verbal content variables of a subset of statements. For exploratory analyses, we included an additional factor into the analysis, namely Question type (general vs. planning vs. emotion-related vs. transportation vs. other specific). All analyses were conducted with an alpha level of .05.

3. Results

The data reported in this investigation is available via:
https://osf.io/knhz4/?view_only=2ab543b62da84e2db643aa5ac8b88285

3.1 Descriptive statistics

Table 2 shows descriptive statistics for the final sample.

3.2 Manipulation check

There was no difference in the distribution of participants who failed the control question between the flyers in the truthful and deceptive condition, $X^2(1) = 0.07$, $p = .795$, Cramer's $V = 0.05$. A one-way ANOVA on the question expectedness revealed that expectedness differed across Question type, $F(4, 880) = 13.87$, $p < .001$, $f = 0.13$. Follow-up tests indicated that the general questions were perceived as more expected ($M = 7.26$, $SD = 2.04$) than questions of all other types ($M_{collapsed} = 6.24$, $SD_{collapsed} = 1.80$, ps $> .05$, see Table 2)

3.3 Main analysis

For detailedness, the 2 (Veracity: truthful vs deceptive) by 2 (Information protocol: standard vs specific) between-subjects ANOVA revealed that there was no significant main effect of Veracity, $F(1, 218) = 0.07$, $p = .787$, $f = 0.02$, nor for Information protocol, $F(1, 218) = 3.57$, $p = .060$, $f = 0.13$. This main effect of Information protocol suggests a trend that those who received the instruction to provide specific information ($M = 14.97$, $SD = 3.54$) provided more detailed information than those with standard instructions ($M = 14.07$, $SD = 3.53$). The interaction between Veracity and Information protocol was not significant, $F(1, 218) = 1.00$, $p = .754$, $f = 0.07$ (see Table 2).

For information specificity, there was no significant main effect of Veracity, $F(1, 218) = 0.70$, $p = 0.40$, $f = 0.06$, and no significant Veracity by Information protocol interaction, $F(1, 218) = 0.28$, $p = .596$, $f = 0.04$. However, the main effect of Information protocol was significant, $F(1, 218) = 6.78$, $p = .010$, $f = 0.18$, suggesting that those instructed to provide specific information ($M = 1.47$, $SD = 1.19$) did in fact provide more information than those with standard instructions ($M = 1.09$, $SD = 1.00$). For sentence specificity, there was no significant main effect of Veracity, $F(1, 218) = 0.04$, $p = .836$, $f = 0.04$, or Information protocol (specific: $M = 0.55$, $SD = 0.31$; standard: $M = 0.55$, $SD = 0.32$), $F(1, 218) = 0.01$, $p = .941$, $f = 0.01$. The interaction between Veracity and Information protocol not significant either, $F(1, 218) = 0.51$, $p = .475$, $f = 0.05$.

Table 2. *Descriptive statistics for participants who reported upon their upcoming flight (means and standard deviation in parentheses).*

	Standard		**Specific**	
	Truthful	Deceptive	Truthful	Deceptive
Final *n*	49	52	60	61
Weeks until flight	5.12 (3.64)	7.08 (7.06)	6.07 (5.48)	5.85 (7.73)
Times visited before	4.24 (5.02)	4.82 (6.89)	4.73 (5.84)	4.70 (7.23)
Motivation	8.00 (1.99)	8.12 (1.83)	7.81 (1.87)	8.23 (1.53)
Failed control question (%)	3.92 (19.60)	3.70 (19.06)	4.76 (21.47)	1.61 (12.70)
Number of words*	195.18 (66.26)	224.33 (112.04)	210.22 (75.52)	214.75 (80.54)
Expectedness general questions	6.84 (2.10)	7.53 (1.72)	7.08 (2.33)	7.57 (1.90)
Expectedness planning questions	5.94 (2.30)	6.10 (2.29)	6.37 (2.30)	6.70 (2.41)
Expectedness emotion-related questions	5.46 (2.47)	6.65 (1.93)	6.24 (1.90)	7.07 (2.25)
Expectedness transportation questions	6.22 (2.22)	6.62 (1.96)	6.43 (2.14)	6.95 (1.85)
Expectedness other specific questions	6.21 (2.25)	6.41 (2.09)	6.26 (2.14)	6.78 (2.26)
Detailedness (LIWC)	13.93 (3.30)	14.21 (3.75)	14.98 (3.94)	14.96 (3.14)
Average sentence specificity*100	57.41 (32.48)	53.43 (32.60)	54.00 (30.60)	56.20 (32.43)
Named entity-based information specificity*100	0.98 (0.92)	1.18 (1.06)	1.45 (1.16)	1.49 (1.23)

Note. *There were no main effects or an interaction between Veracity and Information Protocol for the number of words, all *p*'s > .143.

3.4 Exploratory analyses

3.4.1 Human coded variables

Trained, human coders blind to the experimental conditions and hypotheses scored a subset (n = 94)[8] of statements (i.e. those who fly within no more than four weeks) on plausibility, complications, detailedness, why-utterances, and truthfulness on 7-point Likert scales. For each statement that was coded by the two coders, we used an odd-even split to determine which scoring to use for the analysis. We conducted 2 (Veracity: truthful vs. deceptive) by 2 (Information Protocol: standard vs. specific) between-subjects ANOVAs on each of the five variables (see Table 3). There was a significant main effect of Information Protocol for detailedness, $F(1, 87) = 12.32$, $p < .001$, $f = 0.37$;[9] and for why-utterances, $F(1, 87) = 3.97$, $p = .050$, $f = 0.21$.[10] Statements were rated as more detailed and containing more why-utterances when the instructions for participants were to provide specific information. There were, however, no effects of Veracity, Fs < 1.

3.4.2 Question type

To test how the Question type affected the detailedness of the answers, we added Question type as within-subjects factor and conducted a 2 (Veracity: truthful vs deceptive) by 5 (Question type: general, planning, emotion-related, transport, other) ANOVA on the LIWC-scored detailedness. There was no significant main effect of Veracity, $F(1, 220) = 0.07$, $p = .789$, $f = 0.02$, and no significant Veracity*Question type interaction, $F(4, 880) = 1.49$, $p = .203$, $f = 0.04$. The main effect of Question type was significant, $F(4, 880) = 18.95$, $p < .001$, $f = 0.14$. Table 4 shows the means (SDs) per Question type and follow-up contrasts between the different question types. For the average sentence specificity, the same pattern emerged with only a significant main effect of Question type, $F(4, 1408) = 40.11$, $p < .001$, $f = 0.20$; as well as for information specificity with a significant main effect of Question type, $F(4, 880) = 48.49$, $p < .001$, $f = 0.22$.

3.4.3 Past events versus future intentions

We examined exploratory whether the temporal dimension of the flight moderated the detailedness of participants' answers and potentially the effect of Veracity. A 2 (Veracity: truthful vs deceptive) by 2 (Temporality: past flight vs. upcoming flight) between-subjects ANOVA on the LIWC-based detailedness revealed only a significant main effect of Temporality, $F(1, 350) = 9.80$, $p = .002$, $f = 0.17$. Answers about past flights regardless of Veracity contained more detailed information (M =

[8] Of the total 110 statements, 16 were randomly chosen for the training of the coders and therefore excluded from the analysis.

[9] The main effect of Information Protocol, $F(1, 87) = 1.48$, $p = .227$, $f = 0.13$; as well as the Veracity*Information Protocol interaction, $F(1, 87) = 0.25$, $p = .616$, $f = 0.05$, were non-significant.

[10] The main effect of Information Protocol, $F(1, 87) = 0.87$, $p = .354$, $f = 0.10$; as well as the Veracity*Information Protocol interaction, $F(1, 87) = 0.07$, $p = .783$, $f = 0.03$, were non-significant.

15.80, *SD* = 3.57) than answers about upcoming flights (*M* = 14.56, *SD* = 3.55). There was no such effect for average sentence specificity or information specificity.

3.4.4 Temporal immediacy of intentions

In order to test whether the immediacy of flying (i.e. how long away in the future/past the flight was) was related to detailedness, we included the number of weeks until/after the flight in the ANOVA model. There was no significant main effect of or interaction with the number of weeks, all *p*s > .05.

Table 3. *Means (SDs) of the human-coded variables per Veracity and Information Protocol.*

	Standard		**Specific**	
	Truthful	Deceptive	Truthful	Deceptive
n	24	24	14	29
Detailedness	1.38 (0.65)	2.67 (1.88)	1.93 (1.00)	2.90 (1.76)
Plausibility	5.33 (1.55)	5.50 (1.50)	5.57 (1.83)	5.24 (1.38)
Complications	2.20 (1.38)	2.08 (1.59)	2.29 (1.33)	2.48 (1.09)
Why-utterances	2.88 (1.39)	3.42 (1.59)	2.50 (1.29)	3.21 (1.42)
How-utterances	3.33 (1.09)	3.79 (1.41)	3.36 (1.08)	3.62 (1.32)
Truthfulness	3.71 (1.57)	4.83 (1.86)	4.71 (2.16)	4.62 (1.76)

Table 4. *Detailedness, average sentence specificity and information specificity per Question type.*

	General question		Planning question		Emotion-related question		Transportation question		Other specific question	
	T	D	T	D	T	D	T	D	T	D
Detailedness (LIWC)	12.92 (9.24)	12.94 (8.21)	16.52 (8.54)	16.36 (7.91)	14.48 (9.85)	14.87 (9.60)	12.47 (10.17)	11.28 (8.94)	14.05 (8.26)	15.62 (8.21)
Average sentence specificity*100	14.91 (21.47)	16.46 (22.73)	15.75 (24.42)	17.07 (24.35)	9.26 (17.98)	9.14 (18.97)	7.54 (15.87)	8.87 (19.56)	6.91 (15.41)	8.01 (17.44)
Named entity-based information specificity*100	22.53 (32.60)	24.54 (30.68)	13.56 (27.76)	13.96 (20.31)	6.79 (14.81)	9.07 (20.01)	5.35 (15.88)	5.90 (18.76)	6.75 (17.88)	7.38 (13.96)

Note. T = truthful; D = deceptive.

4. Discussion

In this study, we examined whether computer-automated verbal content analysis could differentiate between participants who provided truthful or deceptive statements about their upcoming flight. To address challenges of large-scale applicability, we tried to adopt an online data collection methodology and a computer-automated analytical approach. We used validated linguistic software to model the

detailedness of statements. Our core hypotheses were that truthful statements contained more detailed information than false statements, particularly when being asked to be as specific as possible.
First, the data did not support the hypothesis that truthful statements contain more detailed information than false statements which is in contrast to studies by Sooniste et al. (2015) and Warmelink et al. (2013). Those studies found that truthful statements tended to be richer in detail than false statements. In our data, none of the dependent variables indicated a significant main effect of the veracity of the answers given. Therefore we do not interpret any of the observed patterns. Second, our data partially supported the hypothesis that promoting specific answers resulted indeed in slightly more detailed and specific answers than giving standard instructions. These findings corroborate one part of Nahari et al.'s (2014) information protocol hypothesis: promoting specific answers did seem to result in more specific answers, although not to the effect of eliciting differences between truthful and deceptive answers. In exploratory analyses, human judgments of the statements corroborate the finding that the information protocol manipulation facilitated the elicitation of information. However, the gain in information was found not to be conducive to differences in deceptive and truthful statements.
In several ways, the results from this study are not in line with previous studies on the detection of false intentions (Vrij et al., 2011; Sooniste et al., 2013, MacGiolla et al., 2013), the detection of deception on past events (Nahari et al., 2014). We will first discuss limitations related to the experimental design and data collection and then elaborate on those related to the data analysis and operationalization of constructs.

Experimental design and data collection

There are some differences between our and previous studies in this domain on deceptive intentions. First, the most striking difference is that out setting was entirely non-interactive whereas all studies within the cognitive load paradigm on intentions so far use face-to-face interview settings (e.g. Sooniste et al., 2013, Ormerod & Dando, 2015). Our data collection procedure may have affected the statements in two important ways. On the one hand, the participants were merely filling in forms in our study implying that the interviewing process was passive (i.e. without an interviewer as conversation counterpart) rather than actively engaging as championed by the cognitive load paradigm (see Vrij, Granhag & Porter, 2010). On the other hand, the flow of the questions was pre-scripted and non-dynamic. Such a static interviewing may have precluded the possibility of asking follow-up questions or providing clarifications and implied that we had to assume participants' same understanding of the questions. In this sense, our procedure was fundamentally different from the Cognitive Interview promoting active engagement of the interviewer (Vrij et al., 2015; Memon, Meissner, & Fraser, 2010). Moreover, the latter might also have caused order effects resulting in a steadily declining amount of information provided in the course of the task. Our data, however, do not support this order effects hypothesis since we did not find a linear trend towards shorter answers.

Second, the non-interactive data collection process used in the current investigation poses interesting questions for the cognitive deception detection approach. Contrary to the vast majority of studies on verbal deception detection (see Vrij et al., 2015), there was no interviewer present and hence no time pressure for the interviewee to reply. From a theoretical perspective, it is possible that the assumption that additional cognitive load makes lying harder than telling the truth might be moderated by the temporal immediacy with which must reply to a question. For example, in a face-to-face interview, an interviewee may be inclined to respond within very short time to avoid any irregularities in the conversation. To meet this objective, the interviewee has to reply quickly. On the contrary, if there is no interviewer, there is no time pressure, and therefore, it seems to be irrelevant to the interviewee how long they take to reply. Although there was no difference in the response time (see Appendix), future research could shed light on the interplay of cognitive load and time pressure.

Third, in contrast to the experimental setup by Sooniste et al. (2013) and the semi-experimental study by Mann et al. (2011; Warmelink et al., 2013), we had to rely on participants' self-reported flight plans as ground truth. Although we cannot find an obvious reason why participants could have lied about their flight plans in the first place, this might have blurred our data. We did try to address this limitation by asking a control question after the experimental task on which 7.78% of our participants admitted providing false information.

Fourth, a critical assumption made by us, inspired by previous studies (Sooniste et al., 2013; Warmelink et al., 2012), was that planning and transportation questions, in particular, would be perceived as unexpected. The unexpectedness should have put truth tellers as compared to liars in an advantage of being able to report on their genuine trip freely. Our data suggest that this assumption was only partially met: participants did indicate that the general questions were more expected than all others, but there was no differentiation between the questions of the remaining four topics. In Sooniste et al.'s (2013) study, the general (intention-related) question was perceived as less difficult than the planning-related questions, but it remains to be examined how planning-questions differ from transportation-questions, for example. Further research might have to pre-test question expectedness to yield the expected differences that might eventually benefit deception detection (for a pre-testing method, see Warmelink et al., 2012).

Fifth, the information protocol manipulation we used, could have worked in two opposite directions. By instructing participants in the standard information protocol condition to provide *as much detail as possible*, it is imaginable that this gave the participants, especially the liars, a hint that detailedness is a cue of interest. The information protocol has been shown to work with instructing participants to provide as much *verifiable* information as possible (e.g. Nahari et al., 2014a; Harvey et al., 2016). But research by Nahari and Pazuelo (2015) suggests that the information protocol pointing towards detailed (rather than verifiable) information might impede truth-lie differences. In the current study, the beneficial effect of the 'specific' instructions on the detectability of the statements' veracity could have been canceled out by the detrimental effect of the 'as detailed as possible' instructions. Although

the instruction to provide *as detailed answers as possible* has been used as interviewing tool in other studies (e.g. Sooniste et al., 2013), further research should try to adopt novel ingredients for verbal deception detection like the model statement technique (e.g. Leal et al., 2015) and the VA information protocol (e.g. Harvey et al., 2016) for the detection of intention.

Fifth, although our aim was similar to that of previous intentions experiments, we think we have measured multiple types of intentions here. Being able to report upon immediate and specific intentions was a major factor in studies of Sooniste et al. (2013, 2015), but we might not have been able to grasp the same to-be-implemented intentions here (Malle, Moses & Baldwin, 2001). On average the flight upon which participants reported was five to seven weeks away, whereas Sooniste et al.'s (2013) participants had direct plans to implement their intention on the spot. When we selected only those who were flying within two weeks, the veracity effect remained non-significant. However, our data tend to support the notion that reporting upon a future trip yields fewer details than reporting about a past flight regardless of statement veracity. Combined with the vagueness of the intentions that might have played a role here, the current procedure might simply have put both liars and truth tellers in the difficult situation of providing information regarding transportation, for example, of a flight that they had not yet planned in sufficient detail. One way to address this limitation is in-vivo studies (e.g., Vrij et al., 2011) directly at the airport which allows for direct intentions and ground truth checking.

Data analysis and operationalization

In our operationalization procedure, it merits attention that we adhered predominantly to fully computer-automated scoring of verbal content. While the quantification of qualitative measures remains a key challenge for social sciences and computational disciplines alike, related studies (Bond & Lee, 2005; Bachenko, Fitzpatrick, & Schonwetter, 2008; Fitzpatrick, Bachenko, & Fornaciari, 2015) have shown that this is feasible. Skepticism towards automated text analysis has been voiced elsewhere (e.g. Vrij, 2008) for context-sensitive scoring tools like Reality Monitoring. The argument is that human coders are more attentive to context-dependency than dictionary-based approaches like the LIWC (but see Newman Pennebaker, Berry & Richards, 2003; Strappavara & Mihalcea, 2009). In the current experiment, even with human-scored verbal content variables, no differences between truthful and deceptive statements became apparent. Moreover, the moderate to high intra-class coefficients (ICCs 0.67 - 0.86, except for how-utterances) between the two trained coders suggest that the lack of differences in cues like plausibility and detailedness was not due to insufficient reliability in the coding itself. It is noteworthy that the coders in the current study rated the presence of criteria rather than counting the occurrence of, for example, details. Nahari (2016) suggests that rating the presence of criteria is more sensitive to individual differences between coders. This effect can also explain the moderate ICCs achieved. However, previous intentions studies that did find truth-lie differences (e.g. Sooniste et al., 2013) used the same rating procedure that we did (but see, Warmelink et al., 2012, 2013). Since we had human-coders also rate the statements' truthfulness, we were able to

test whether there were truth-lie differences that were merely not picked up by our operationalization of variables. The human-coded truthfulness was not better than chance (*AUC* = 0.56; 95% CI: 0.44 - 0.68), from which we conclude that there were no obvious truth-lie differences that automated or poor human coding might have blurred.

Future research

Although all of the limitations mentioned above merit the attention of future research on deceptive intentions, we believe an essential requirement for applied purposes is that of large-scale applicability. Therefore, research efforts might be directed towards remote approaches to information elicitation within the cognitive load paradigm. For example, rather than providing participants a form to be filled in, one could develop instant-messaging frameworks for the detection of deception on both intentions as well as past events. None of the existing instant messaging studies in a deception context (Zhou, 2005; Derrick et al., 2013) targeted scalability and deceptive intentions. A novel framework of the cognitive load interviewing paradigm could be developed for instant-messaging ideally i) allowing for active information elicitation through interviewee-interviewer interaction, ii) providing higher information gain (i.e. shorter replies with more information), iii) facilitating quick interview procedures, and iv) laying the foundation for automated chatbot-like systems that would be a step towards large-scale applicability of verbal deception detection.

5. Conclusion

The reported experiment was an attempt to investigate false intentions using a remote data collection procedure. Participants' truthful or deceptive statements about their upcoming or past flight did not reveal any differences on verbal content variables with both a computer-automated analysis approach as well as human-coding of the declarations. The findings suggest that it is hard to detect false intent if the experimental manipulation does not directly target implementable actions and if the question procedure is too far from face-to-face interviewing. The future research agenda on large-scale cognitive approaches might want to explore strategies more responsive to the information elicitation processes allowing for cognitive manipulations of the interviewing procedure.

References

Bachenko, J., Fitzpatrick, E., & Schonwetter, M. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. *Proceedings of the 22nd International Conference on Computational Linguistics*, 41-48.

Bond, G.B., & Lee, A.Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19,* 313-329. doi:10.1002/acp.1087

Derrick, D. C., Meservy, T. O., Jenkins, J. L., Burgoon, J. K., & Nunamaker, J. F. (2013). Detecting Deceptive Chat-Based Communication Using Typing Behavior and Message Cues. *ACM Transactions on Management Information Systems*, *4*(2), 1–21. https://doi.org/10.1145/2499962.2499967

Evans, J. R., Michael, S. W., Meissner, C. A., & Brandon, S. E. (2013). Validating a new assessment method for deception detection: Introducing a Psychologically Based Credibility Assessment Tool. *Journal of Applied Research in Memory and Cognition*, *2*(1), 33–41. https://doi.org/10.1016/j.jarmac.2013.02.002

Fitzpatrick, E., Bachenko, J., Fornaciari (2015). *Automatic detection of verbal deception.* Morgan & Claypool Publishers.

Fornaciari, T., & Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, *21*(3), 303–340. https://doi.org/10.1007/s10506-013-9140-4

Granhag, P. A. (2010). On the psycho-legal study of true and false intentions: Dangerous waters and some stepping stones. *The Open Criminology Journal*, *3*, 37–43.

Harvey, A. C., Vrij, A., Nahari, G., & Ludwig, K. (2016). Applying the Verifiability Approach to insurance claims settings: Exploring the effect of the information protocol. *Legal and Criminological Psychology*, n/a-n/a. https://doi.org/10.1111/lcrp.12092

Honnibol, M. (2016). spaCy (Version 0.100.6) [Computer software]. Available from https://spacy.io/

Honts, C.R., & Hartwig, M. (2014). Credibility assessments at portals. In D.C. Raskin, C.R. Honts, & J.C. Kircher (Eds.) *Credibility assessment: Scientific research and applications* (pp. 37-62). Academic Press.

Johnson, M. K. and Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67-85.

Johnson, M. K., Bush, J. G., & Mitchell, K. J. (1998). Interpersonal reality monitoring: Judging the sources of other people's memories. *Social Cognition, 16*, 199-224. doi:10.1521/soco.1998.16.2.199

Kleinberg, B., & Verschuere, B. (2015). The role of motivation to avoid detection in reaction time-based concealed information detection [Electronic resource]. *Journal of Applied Research in Memory and Cognition*.

Kleinberg, B., & Verschuere, B. (2015). Memory Detection 2.0: The First Web-Based Memory Detection Test. *PLOS ONE, 10*(4), e0118715. https://doi.org/10.1371/journal.pone.0118715

Kleinberg, B., Nahari, G., & Verschuere, B. (2016). Using the verifiability of details as a test of deception: A conceptual framework for the automation of the verifiability approach. In *Proceedings of NAACL-HLT* (pp. 18–25). Retrieved from http://www.anthology.aclweb.org/W/W16/W16-0803.pdf

Li, J. J., & Nenkova, A. (2015). Fast and Accurate Prediction of Sentence Specificity. In *AAAI* (pp. 2281–2287). Retrieved from https://pdfs.semanticscholar.org/69f5/a7032605a88e7bed7bf0c9c2218c5e3f2512.pdf

Mac Giolla, E., Granhag, P. A., & Ask, K. (2016). Task-related Spontaneous Thought: A Novel Direction in the Study of True and False Intentions. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2016.04.010

Mac Giolla, E., Granhag, P. A., & Liu-Jönsson, M. (2013). Markers of good planning behavior as a cue for separating true and false intent: Good planning behavior and true and false intent. *PsyCh Journal*, *2*(3), 183–189. https://doi.org/10.1002/pchj.36

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: a review of the empirical evidence. *Psychology, Crime & Law*, *11*(1), 99–122. https://doi.org/10.1080/10683160410001726356

Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. *In Proceedings of the Association for Computational Linguistics*, 309-312.

Nahari, G. & Pazuelo, M. (2015). Telling a convincing story: Richness in detail as a function of gender and information. *Journal of Applied Research in Memory and Cognition, 4*, 363-367. http://dx.doi.org/10.1016/j.jarmac.2015.08.005

Nahari, G. & Vrij, A. (2014). Are you as good as me at telling a story? Individual differences in interpersonal reality monitoring. *Psychology, Crime & Law, 20* (6), 573-583. 10.1080/1068316X.2013.793771

Nahari, G., & Vrij, A. (2014). Can I Borrow Your Alibi? The applicability of the verifiability approach to the case of an alibi witness. *Journal of Applied Research in Memory and Cognition, 3,* 89 – 94.

Nahari, G., Leal, S., Vrij, A., Warmelink, L., & Vernham, Z. (2014). Did Somebody See It? Applying the Verifiability Approach to Insurance Claim Interviews: The verifiability approach in insurance interviews. *Journal of Investigative Psychology and Offender Profiling*, *11*(3), 237–243. https://doi.org/10.1002/jip.1417

Nahari, G., Vrij, A., & Fisher, R. (2012). Does the truth come out in the writing? SCAN as a lie detection tool. *Law & Human Behavior, 36*, 68–76. doi:10.1007/s10979-011-9264-6

Nahari, G., Vrij, A., & Fisher, R. P. (2014a). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, *19*(2), 227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x

Nahari, G., Vrij, A., & Fisher, R. P. (2014b). The Verifiability Approach: Countermeasures Facilitate its Ability to Discriminate Between Truths and Lies: The verifiability approach and countermeasures. *Applied Cognitive Psychology*, *28*(1), 122–128. https://doi.org/10.1002/acp.2974

Newman, M.L., Pennebaker, J.W., Berry, D.S., & Richards, J.M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychological Bulletin, 29* (5), 665-675 DOI: 10.1177/0146167203251529

Ormerod, T. C., & Dando, C. J. (2014). Finding a Needle in a Haystack: Toward a Psychologically Informed Method for Aviation Security Screening. *Journal of Experimental Psychology: General, 144*(1), 76–84. doi:10.1037xge0000030

Ott, M., Cardie, C., & Hancock, J.T. (2013). Negative deceptive opinion spam. *Proceedings of NAACL-HLT 2013*, 497–501.

Ott, M., Choi, Y., Cardie, C., & Hancock, J.T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 309- 319.

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates (www.liwc.net).

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. https://doi.org/10.1080/1068316X.2013.793333

Sooniste, T., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2015). Statements about true and false intentions: Using the Cognitive Interview to magnify the differences. *Scandinavian Journal of Psychology*, *56*(4), 371–378. https://doi.org/10.1111/sjop.12216

Undeutsch U (1967) Beurteilung der Glaubhaftigkeit von Aussagen [Veracity assessment of statements]. In: Undeutsch U (Ed.) *Handbuch der Psychologie*: vol 11. Forensische Psychologie. Hogrefe, Gottingen, pp 26–181

Undeutsch U (1982) Statement reality analysis. In: Trankell A (Ed.) *Reconstructing the past: the role of psychologists in criminal trials*. Kluwer, Deventer, pp 27–56

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition, 4*(1), 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001

Vrij, A. (2008). Reality Monitoring. In A. Vrij (Ed.) *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley and Sons, Sussex, UK.

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110–117. https://doi.org/10.1016/j.jarmac.2012.02.004

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, n/a-n/a. https://doi.org/10.1111/lcrp.12088

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection. *Psychological Science in the Public Interest*, *11*(3), 89–121. https://doi.org/10.1177/1529100610390861

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Lying about flying: the first experiment to detect false intent. *Psychology, Crime & Law*, *17*(7), 611–620. https://doi.org/10.1080/10683160903418213

Warmelink, L., Vrij, A., Mann, S., & Granhag, P. A. (2013). Spatial and Temporal Details in Intentions: A Cue to Detecting Deception: Spatial and temporal details in lie detection. *Applied Cognitive Psychology*, *27*(1), 101–106. https://doi.org/10.1002/acp.2878

Warmelink, L., Vrij, A., Mann, S., Leal, S., & Poletiek, F. H. (2013). The Effects of Unexpected Questions on Detecting Familiar and Unfamiliar Lies. *Psychiatry, Psychology and Law*, *20*(1), 29–35. https://doi.org/10.1080/13218719.2011.619058

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*, *141*(2), 178-183.

Zhou, L. (2005). An Empirical Investigation of Deception Behavior in Instant Messaging. *IEEE Transactions on Professional Communication*, *48*(2), 147–160. https://doi.org/10.1109/TPC.2005.849652

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1–59). New York: Academic Press.

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, *19*(4), 307-342.

Appendix

For further exploratory analyses, we added response time, the number of deletions, and gaps between key-press events as dependent variables

Para-linguistic keyboard-related variables

Following Zhou's (2005) taxonomy of deceptive behavior, we measured the para-linguistic keyboard-related variables for participants' answers to the ten main questions. The response time in milliseconds from the first key-press event to finishing the question, the number of occasions the backspace/delete key was pressed, the number of gaps between key-press events that were longer than 100ms, 200ms, and 300ms. These variables were intended to represent participants' editing behavior and have previously been shown to be informative (e.g. Zhou, 2005; Derrick et al., 2013).

Specifically, we recorded the following para-linguistic variables: response time in milliseconds, number of deletions, number of gaps between key-presses larger than 100ms, 200ms and 300ms. All of these three non-verbal measures might be a way to model cognitive processes such as cognitive load involved in formulating answers through non-verbal behavior in a non-interactive setting (e.g. Zhou, 2005; Derrick et al., 2013). We conducted separate one-way ANOVAs with Veracity as factor. There were no significant effects of Veracity on the average response time ($M_{truthful}$ = 2953, $SD_{truthful}$ = 2595; $M_{deceptive}$ = 2984, $SD_{deceptive}$ = 2606); on the average number of deletions ($M_{truthful}$ = 12.57, $SD_{truthful}$ = 11.74; $M_{deceptive}$ = 13.16, $SD_{deceptive}$ = 13.72); nor on the average number of key-press gaps larger than 100ms ($M_{truthful}$ = 120.91, $SD_{truthful}$ = 47.45; $M_{deceptive}$ = 128.81, $SD_{deceptive}$ = 58.91). However, for the average number of key-press gaps larger than 200ms ($M_{truthful}$ = 62.79, $SD_{truthful}$ = 30.65; $M_{deceptive}$ = 69.84, $SD_{deceptive}$ = 33.03) and larger than 300ms ($M_{truthful}$ = 35.10, $SD_{truthful}$ = 19.90; $M_{deceptive}$ = 40.04, $SD_{deceptive}$ = 22.74), there were significant differences, $F(1, 352) = 4.30$, $p = .038$, $f = 0.11$, and $F(1, 352) = 4.68$, $p = .031$, $f = 0.11$, respectively. Although these differences might be an indicator of participants' hesitation when formulating answers and might then be in line with the cognitive load rationale, these conclusions are merely tentative and merit further replication in future studies. Furthermore, the effect sizes are small and thereby as single cues to deception, these para-linguistic variables are of limited relevance for the detection of individual cases.