“Language of Lies”: Urgent Issues and Prospects in Verbal Lie Detection Research

Since its introduction into the field of deception detection, the verbal channel has become a rapidly growing area of research. The basic assumption is that liars differ from truth-tellers in their verbal behavior, making it possible to classify them by inspecting their verbal accounts. However, as noted in conferences and in private communication between researchers, the field of verbal lie detection faces several challenges that merit focused attention. The first author therefore proposed a workshop with the mission of promoting solutions for urgent issues in the field. Nine researchers and three practitioners with experience in credibility assessments gathered for three days of discussion at Bar-Ilan University (Israel), in the first international verbal lie detection workshop.

Practitioners were invited to take part in the workshop, as a comprehensive understanding of any research area that has direct real-world applications can be achieved only by taking into account both theoretical and practical perspectives. One can see science as a parasail tied to a boat that represents the field. The practitioners are in the boat, dealing with the real-life challenges. They navigate the boat in accordance with their aims, in real-life conditions. In contrast, academic researchers often keep their distance from “real life.” Height gives them a broader perspective. They can see further than the practitioners on the boat and help with navigation by suggesting better directions to reach desired destinations. In our view, science must be connected to real-life goals, which is why the parasail is tied to the boat. If the rope is cut, the parasail will lose contact with reality. Still, researchers do not need to sit on the boat and practitioners do not need to join the researchers in the air. What we do need is cooperation, with each side playing its own important role. Sharing

knowledge and experience will provide us with the most comprehensive picture. With this aim in mind, researchers and practitioners worked together throughout the workshop.

The primary session of the workshop took place the morning of the first day. In this session, each of the participants had up to10 minutes to deliver a brief message, using just one slide. Researchers were asked to answer the question: "In your view, what is the most urgent, unsolved question/issue in verbal lie detection?" Similarly, practitioners were asked: "As a practitioner, what question/issue do you wish verbal lie detection research would address?" The issues raised served as the basis for the discussions that were held throughout the workshop. The current paper first presents the urgent, unsolved issues raised by the workshop group members in the main session, followed by a message to researchers in the field, designed to deliver the insights, decisions, and conclusions resulting from the discussions.

Commentary #1 by Granhag: Deception detection research must be more sensitive context

Researchers tend to study one topic at a time and this makes sense. To put too many factors into play will make it difficult to assess what is for responsible for any observed effects. However, reality does not care much about researchers' preferences. Reality is complex and fluid. I believe that the field of verbal lie detection faces several future challenges; challenges that will stimulate and reshape the field. One such challenge is the need for a more context sensitive research agenda. I will offer two concrete examples.

First, police interviewers rarely approach a suspect with one single objective, for example to detect deceit. They always have *multiple objectives*; for example to elicit

information about critical aspects of a crime, to assess whether the information collected is reliable and to make the suspect willing to talk again. In most real life interrogations there is an intricate interplay between (a) the collection of information and (b) the elicitation of cues to deceit. Sometimes an increase of information can result in enhanced cues to deceit. Sometimes the elicited cues to deceit can be used to draw out new, critical information. Most of today's research ignores this interplay.

My second example concerns the *output*. Typically, 'truth/lie-judgements' made by lie-catchers are aggregated and translated into a score of accuracy. I am clear on that we need to have some sort of measure to address the comparative effectiveness of different lie detection techniques. But I also believe that we need to acknowledge that the interrogation of a suspect is not an isolated stage. The outcome may be fed into the ongoing investigation, and parts of the investigation may be fed into the prosecutor's case-construction. A police interviewer's assessment of (a specific part of) a suspect's statement is of no evidentiary value. The assessment as such has no relevance for a prosecutor building a case.

Today there is a paradigmatic format for many of our studies and we tend to stay very loyal to that format. Researchers are sometimes rather creative with respect to what they come up with within this format – but there are few signs that they are willing to change the format as such. In today's research the truth/lie assessment is treated as an end-point. In the real world it is, if anything, a starting point.

I think deception researchers face a more challenging task than do for example researchers who developed memory enhancing techniques. The output of an interview technique used for witnesses produces information. This is evidence that can be used in

court. The output of a technique used for detecting deceit is often a 'truth-lie' judgment. This assessment has no relevance in court.

To increase the impact I think we as deception researchers need to start acknowledge and address a larger portion of the problem. We need to account for context and we need to set up our studies a bit different. I do not think we should trust someone else to contextualize our findings. This responsibility falls on us, the researchers in the field. Reality will not change because it is ignored. Police interrogators will continue to have *multiple objectives* and prosecutors will *not care very much about police officers' subjective views* about the veracity of a suspect's statement. If we continue to drill deeper in our own favorite niche of reality - if we continue to ignore context - we run the risk offering solutions that have to seek their problems.

Commentary #2 by Nahari: A Call for Theory

Research in verbal lie-detection has grown intensely and impressively in the last few decades, resulting in the identification of diagnostic indicators of truthfulness and deception, the development of techniques to determine veracity, and the examination of these techniques in different contexts and situations. Today, we are able to outline valid directions for further promotion of this field, and to suggest practical approaches for detecting lies. However, though valuable, current scientific knowledge cannot yet provide a comprehensive understanding of deceptive verbal behaviour. While we are able to provide theoretical explanations for specific pieces of empirical evidence, we still cannot see the whole picture, and do not yet have established theories of deception.

We have, however, "adapted" theories from other, related fields. While this generally seems like a reasonable and practical approach, it may involve pitfalls, and should be done with care. A significant example, demonstrated by Nahari (2018a), is the extension of Reality Monitoring (RM) theory (Johnson & Ray; 1981) from its original use in the field of memory source monitoring to the field of deception detection. This extension did not take into consideration a factor that plays a key role in lies but not in false memories: the motivation of liars to deceive and their resultant tendency to apply strategies in order to be convincing. As a result, the diagnostic indicators yielded from RM do not fulfil their potential utility in determining veracity (Nahari, 2018b; Nahari & Nisin, 2018).

I believe that we have reached the point at which our aims can be more ambitious, and more significant resources can be directed to the establishment of deception theories. We are keen to provide the field with the solutions it requires, such as effective diagnostic techniques. Yet, however appropriate and desirable this aim, it will be beneficial to direct some of our resources to the establishment of deception theories, and thereby fulfill another part of our mission as scientists: the provision of explanations. After all, it is the theoretical underpinnings that provide scientific value to the techniques. It is indeed necessary to know *what* works. Still, we must also be able to explain *why* it works, and thereby to predict and define *in advance*, not by a trial and error process, when it works better and when it may fail, when it is appropriate to apply it and when it may lead to wrong decisions. As such, underlying theoretical knowledge serves not only science but also practice, by enabling us to provide practitioners in the field with more reliable tools.

Clearly, the development of theories requires a different type of research. The establishment of a theoretical understanding of a phenomenon sometimes entails distance

from “real world” settings. This distance should not worry us, as it is only temporary. By taking a few steps back, we enable the wider perspective required for the examination of the mechanisms behind observed behaviors. We can then come back to apply the insights gained in more ecological research settings.

To sum up, I wish to use this stage to call for the establishment of deception theories, for the execution of more basic research, and the development of theoretically grounded techniques. Presumably, this will require patience and effort, but it will surely make a worthwhile contribution to both science and practice.

Commentary #3 by Taylor: The Ecological Challenge: Ensuring our Aggregate Results are Individually Relevant[1]

One of the gratifying advantages of working in legal psychology is the impact that theoretically-grounded research can have on practice. This is certainly true for deception research, where evidence of verbal cues and the methods that elicit them are components of practice worldwide (See Comment #11 above). However, there is one question asked by investigators that research struggles to answer: is **this** person lying?

This question is a challenge because most research compares variance in participants’ scores across truth-teller and liar groups. For example, 55 of 69 studies in DePaulo et al.’s (2003) meta-analysis of verbal cues to deceit compared groups, as did 13 of 14 studies in Vrij, Fisher, and Blank (2017) comparison of cognitive and standard approaches to interview. The problem is that analysing the variance of individuals’ scores

[1] This work was part funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1).

from their group's best predicted value (means in ANOVA's case) says little about the degree to which unknown group membership can be predicted from an individual's known score. As Guttman (1988) demonstrated, standardised effect size (e.g., η^2) can be far less than 1.0 in cases when the distribution of each group's scores do not overlap; a scenario where the prediction of veracity would be perfect.

What is needed to answer this question is a single point criterion and an assessment of the extent to which this achieves accurate predictions. In deception research this criterion is likely to be a score on a (reliably implemented) scale, above or below which a statement would be determined deceptive. Investigators satisfied with the performance of a method could then implement the criterion. Researchers seeking to develop a method could refine the criterion, or develop multiple criteria to accommodate moderators (e.g., culture; Taylor, Larner, Conchie, & Menacere, 2017).

Fortunately, the statistics for this kind of analysis exist and they can be calculated from published research. Table 1 presents such calculations alongside customary metrics (i.e., *M, SD*, *d*) for two methods: model statement and unanticipated questions. We chose these methods because they both score the number of details provided, assume truth-tellers will provide more than liars, and have no agreed criterion. Critically, however, the model statement has a uniform delivery while unanticipated questions can vary in the number and type of questions asked. We anticipate that a uniformed delivery will lead to greater consistency across studies, since participants' behaviour is less a function of delivery. The statistic U3 (Cohen, 1977) indicates the proportion of liars whose scores fall below the mean truth-teller score (i.e., what happens with a .50 false alarm rate). DISCO (Guttman, 1988) indicates the proportion of liars whose score falls below the lowest score observed

for truth-tellers (i.e., what happens with a .00 false alarm rate).[2] This lower- and upper-bound gives a sense of how a criterion performs under two risk appetites.

Table 1 begins to answer the opening question. The DISCO coefficients suggest the ability of either method to identify liars without risking false accusations varies widely across studies (Range: 00–88%), with the average (M=.405, SD=.255) implying that approximately 60% of lies would be missed under this criterion. Even when allowing a 50% false alarm rate, neither method has a consistent correct hit rate above .75 (U3: $M_{\text{ModelStatement}}$=.743, $SD_{\text{ModelStatement}}$=.103; $M_{\text{UnanticipatedQuestions}}$=.624, $SD_{\text{UnaticipatedQuestions}}$=.114). The model statement is more discriminatory than unanticipated questions, $t(18) = 2.45$, p=.025, d=1.10, 95%CI [0.09, 2.10], and under the conservative criterion of DISCO it is also more consistent (U3: $CV_{\text{ModelStatement}}$=.191, $CV_{\text{UnantipcatedQuestions}}$=.171; DISCO: $CV_{\text{ModelStatement}}$=.445, $CV_{\text{UnantipcatedQuestions}}$=.750). Finally, the performance of each method depends on the testing context, as evidenced by the range of means and discrimination statistics.

This last observation is important. If researchers are to answer the opening question, then any proposed criterion must be agnostic of context. Supplementary Figures 1 and 2 represent the data in Table 1 by graphing randomised normal distributions derived from the group means and standard deviations of each study. In deriving these distributions, we retained each study's n, which is why some curves appear non-normal. Our assumption of

[2] For this illustration, we compute DISCO not from the original data but from random normal distributions produced using the M, SD, and n of the sample. These statistics approximate what would be observed given the effect size and so are equally useful for evaluating future discrimination performance, which would be on data also not identical to the original sample (Snook, Zito, Bennell, & Taylor, 2005).

normal distributions is also why the tails of some samples extend below zero details (which is impossible); the original samples must be negatively skewed.

The question to consider when viewing each Figure is: what single, vertical criterion line should be drawn to discriminate truth from lie? Critically, assuming any one criterion will result in all truth-tellers from one sample being assumed guilty, or all liars from one sample being exonerated. Thus, what plausible criterion exists? The statistics calculated from a single merged data set reassert the problem: Model statement, *d*=.180, U3=.572, DISCO=.000; Unanticipated questions, *d*=.070, U3=.537; DISCO=.100.

We began by highlighting a critical question for deception researchers. We close by suggesting, based on our examples, that no answer is forthcoming when details are the criterion; scores are too dependent on coding precision (Comment #6), cue quality (Comment #5), and context (Comment #1). For researchers, this means avoiding statements about predictive accuracy without considering how a criterion will play out elsewhere. For investigators—and grant awarders and policy makers—this reaffirms that methods designed to elicit information will not reveal a statement's veracity. What these methods do provide is more information against which investigators can try to deduce veracity. The field of verbal cues to deception is rather the field of information elicitation.

Commentary #4 by Masip: The Need to Complement CBCA (and RM) with Lie Criteria

Criteria-based Content Analysis (CBCA; Steller & Koehnken, 1989) was created to assess the credibility of children's allegations of child sexual abuse (CSA). However, its usefulness to also assess adults' credibility in a variety of contexts has also been examined (e.g., Vrij, 2008).

CBCA contains 19 verbal criteria. The presence of the criteria suggests the statement is truthful, but their absence does not necessarily mean that the person is lying, as there might be alternative reasons for the lack of criteria. Thus, using CBCA one can either conclude that the statement is truthful or that one doesn't know whether it is truthful or deceptive, but one cannot conclude the statement is deceptive. The Reality Monitoring approach (RM) is a similar procedure that contains one lie criterion (cognitive operations), but empirical support for cognitive operations is limited (Masip, Sporer, Garrido, & Herrero, 2005; Vrij, 2008). It would be useful for practitioners if CBCA (and RM) were expanded to also include several lie criteria. Also, if CBCA contained lie criteria, the relative global scores on lie vs. truth criteria could ideally be compared to assess the relative likelihood that the person is lying vs. telling the truth (within-person comparison). Note, however, that this involves combining scores across several criteria, which is very problematic because, as pointed out by Hauch, Sporer, Masip, and Blandón-Gitlin (2017), (a) different CBCA criteria reflect different constructs, (b) individual CBCA criteria differ strongly in terms of validity, and (c) some criteria that are rarely present are nevertheless very diagnostic when they appear (e.g., accurately reported details misunderstood).

Adding lie criteria involves some issues. First, what criteria should be added? An examination of the scientific literature on verbal credibility assessment reveals that it has little to offer in terms of potentially valid lie criteria. Indeed, most of the verbal credibility criteria developed in Germany and Sweden during the 20th century (see Steller & Koehnken, 1989) were *truth* criteria. The so-called SCAN (Scientific Content Analysis) Technique contains deception criteria, but SCAN is not supported by the empirical evidence (e.g., Bogaard, Meijer, Vrij, & Merckelbach, 2016). Also, although eight

linguistic cues were significantly associated with deception in Hauch, Blandón-Gitlin, Masip, & Sporer's (2015) meta-analysis, their effect sizes were extremely small.

Second, CBCA is used (within the more encompassing framework of Statement Validity Assessment; see, e.g., Vrij, 2008) in applied settings to assess the credibility of children's allegations of CSA, being admitted as evidence in court in several countries (e.g., Hauch et al., 2017). Because CBCA contains only *truth* criteria, it can establish the child is telling the truth, but not lying. This favors the child—at the risk of incarcerating an innocent defendant if false positives occur. If *lie* criteria were added, then this child-protective bias would be cancelled out, and errors in the opposite direction could also occur—the children's account could be misclassified as deceptive. Thus, child advocates can object to adding lie criteria to CBCA. This reflects a controversy that can sometimes arise between the quest for scientific objectivity on the one hand, and ethical, social, or ideological considerations on the other hand when scientifically-based procedures are used in applied settings. Adding lie criteria to CBCA to have a more balanced and objective "instrument" would probably be more welcome in contexts other than courts dealing with child abuse—e.g., in police investigations.

Commentary #5 by Vrij: In Search of Verbal Cues to Deceit

Research into differences between truth tellers and liars in speech content accelerated 30 years ago after publication in English of the verbal veracity tool Criteria-Based Content Analysis (CBCA). Since then an alternative tool frequently examined has emerged, Reality Monitoring (RM). Both tools have in common that they focus on how truth tellers typically recall events. As a result, the tools primarily include cues of truthfulness. That is, the presence of cues examined in CBCA and RM give an indication

that someone is telling the truth, but the absence of these cues does not necessarily indicate that someone is lying. This makes it difficult to determine whether someone is lying based on CBCA and RM scores in individual cases. It is therefore important to examine how liars typically express themselves because that could lead to cues to deception. If truth tellers report more cues to truthfulness and liars more cues to deception, the result would be that investigators could determine cut-off scores in individual cases.

We started to examine cues that liars report and came up with an index that includes both cues to truth and deception: the proportion of complications. It constitutes a cue to truthfulness, complications, and two cues to deception, common knowledge details and self-handicapping strategies. A complication is an occurrence that makes a situation more difficult to report than necessary (e.g., "On my way back I got lost and could not find the entry to the tube station"). Common knowledge details refer to strongly invoked stereotypical information about events (e.g., "We went to the top of the Eiffel Tower from where we had a wonderful view of Paris"). Self-handicapping strategies refer to justifications as to why someone is not able to provide information ("I can't tell you about the beginning of the BBQ, because I arrived late"). The proportion score is defined as complications / (complications + common knowledge details + self-handicapping strategies). In the five studies in which it has been examined so far, the proportion score successfully discriminated truth tellers from liars and did so to a better extent than the verbal cue total details (the amount of information reported) (Vrij et al., 2017, 2018a, b, c, d). This is promising because total details has emerged as one of the strongest cues to truthfulness to date (Amado, Arce, & Fariña, 2015).

The proportion of complications score has limitations. Truth tellers typically report more complications than liars, but they do not always report fewer common knowledge details; and although self-handicapping strategies distinguish truth tellers from liars, they do not occur frequently. The proportion of complications index would thus become stronger if we would design interview protocols that make truth tellers to report fewer common knowledge details and liars more self-handicapping strategies. An alternative way to strengthen the index would be to add cues liars report other than common knowledge details and self-handicapping strategies. Whatever solution researchers come up with, the issue is clear: Verbal lie detection would become more successful if we find stronger indicators of verbal cues to deceit.

Commentary #6 by Verschuere & Meijer: Lie Detection: Everyone Their Own Truth?

Verbal lie detection aims to discriminate lie from truth based upon the content of a statement. There seems to be an emerging consensus among scholars about the validity of a limited number of cues. For example, ample research has shown that truthful statements are more detailed than deceptive statements. At first glance, the level of detail seems straightforward to measure. The description 'in the green house' could qualitatively be judged as "detailed". But to allow for a quantitative analysis, researchers have developed specific schemes to code the level of detail. Such coding schemes may, however, contain arbitrary choices. In the example above, one may score 'in the green house' as consisting of three details (the *house*, which is *green*, and the location is specified as *in* the house). Others may discount *in* (e.g., because the word *in* does not add information value and/or follows from the previous sentence) and consequently count two details. If the house has been mentioned before or is deemed to be unrelated to the core event, the researcher may

discard the information altogether (zero details). In sum, the 4-word sentence could be coded as consisting of any value between 0 and 3 details. Manuscripts typically report acceptable within-lab inter-rater reliability, but it remains unclear, to what extent these coding schemes differ between labs. This inter-laboratory reliability is the topic of our proposal. Specifically, we propose several means to improve the coding of statements for verbal credibility assessment:

1. Laboratories should specify their coding scheme prior to data collection and make them available to others. These schemes should be accompanied with coded example statements (explaining the coding), and exercise statements that allows others to adopt the coding scheme and assess their coding skill.
2. Laboratories should collaborate to examine the reliability (and validity) of different coding schemes, preferable on openly available datasets (Kleinberg, Nahari, Arntz, & Verschuere, 2017; Kleinberg, van der Toolen, Vrij, Arntz, & Verschuere, 2018; Mihalcea, Narvaez, & Burzo, 2014; Ott et al., 2013; Ott, Choi, Cardie, & Hancock, 2011; Pérez-Rosas, Abouelenien, Mihalcea, & Burzo, 2015).
3. Perfectly reliable, automated scoring might currently lack an understanding of contextual information. We should, however, explore whether verbal coding can be operationalized as a joint effort between computer and human (e.g., human-in-the-loop where computer codes for detail, and a human coder makes adjustments based on well-specified contextual considerations).

Commentary #7 by Fisher: Detecting Deception among Skilled Interviewees

Most research on detecting deception has been conducted with college students as the interviewees, because they are a convenient sample to work with. However, their lack of relevant experience may lead to their generating data that is not representative of real-world, skillful interviewees. I report here a recent study in which the interviewees were skilled participants, who have had extensive experience in investigative interviews, namely, law enforcement officers. The law enforcement officers participated in a simulated spying mission, in which they encountered several key people, instructions, objects, and actions. The participants were then debriefed about their mission in a setting where it was either appropriate to be truthful (debriefed by their supervisor to learn as much as possible about their experiences during the spying mission) or to be deceptive and to withhold or alter their knowledge in order to protect secure information—because they were captured by the enemy. I describe some novel behaviors that discriminated between truth-telling and lying, and other behaviors that did not discriminate—although they are often recommended as indicators of deception.

I also describe a second, research-oriented approach to detecting deception. The idea revolves around the argument that truth-tellers and liars may differ because of their motivations during an interview. Truth-tellers attempt to be helpful, and thus their behaviors are guided by their cognitive processes (e.g., memory, communication). By comparison, liars attempt to convince interviewers that they are being truthful, and as such, their behaviors are guided by their beliefs about how truth-tellers behave. Researchers should be able to use this distinction between truth-tellers (guided by cognition) and liars (guided by metacognition) to devise better techniques to detect deception.

Finally, I describe a novel source of insight about detecting deception among the criminal (or security-risk) population. An untapped source of information about how criminals (or security-risk) think is the criminal (security-risk) him/herself. I describe the real-world benefits of interviewing criminals (or security risk personnel) about their own thought processes to learn about how the criminal mind works.

Commentary #8 by Hershkowitz: Interviewing to Detect Deception in Children, Alleged Abuse Victims

Beyond the challenge of detecting deceptions in adult suspects, forensic investigators interview children who are suspected victims of abuse on a daily basis, and need to assess the veracity of their allegations. The developmental literature provides important insights into the progression of cognitive and meta-cognitive skills needed to intentionally formulate false statements during childhood years, suggesting that at any given moment, their skills are partial. Despite that, when children do make false statements, professionals' ability to detect them proved to be poor and only slightly over chance.

Lie detection with verbal means in these children faces a double challenge. First, children's verbal skills are partial and their narratives limited, thus poorer in terms of verbal indicators. For example, the occurrence of CBCA (Criterion Based Content Analysis; Steller & Köhnken, 1989) criteria seems to be age related, with very few manifestations of criteria among preschoolers. Second, abuse victims tend to under-report (rather than over-report) their abuse for socio-emotional reasons, often denying suspicions in their investigation, and limiting the information they provide when they disclose. These documented limitations of children suggest that they need both cognitive and social-

emotional support to overcome barriers to providing satisfactory forensic statements that can be subject to valid credibility assessment.

The notion of encouraging the production of forensically relevant details in the service of lie detection has been somewhat explored. One promising direction suggests conveying to the witness the amount of information expected, by exposing him/her to a model statement, or by using a practice interview focused on neutral events before switching to the exploration of the criminal events. Other directions involve the optimization of memory retrieval strategies, prioritizing free-recall prompts. Both manipulating the expectations for a rich production, and using free recall prompts to exhaust memory have been associated with a selective improvement of sincere over fabricated statements, thus improving lie detection.

However, the role of social and emotional support has been rarely addressed. There is first evidence that rapport-building encourages cooperation and provision of information among both victims and suspects. Emotional support that is responsive to reluctance, expression of negative emotions or mention of conflicts has been effective in improving cooperation and enhancing production in substantiated cases, and is therefore expected to improve the judgment of veracity.

In real life investigations of children, using some interviewing strategies to improve lie detection can be a twofold sword. While strategies such as increasing cognitive load challenge lie tellers, revealing indicators of lie, they risk hindering truth tellers' statements as well. However, strategies that allow witnesses to feel more supported, both emotionally and cognitively, did not show harmful effects as long as they were not suggestive. Instead, supportive strategies can enhance children's truthful statements, increasing indices of

credibility, while affecting to a lesser extent lie tellers, who may have difficulty to invent rich and coherent information on the spot. When it comes to children or other vulnerable witnesses, supportive yet non-suggestive interviewing strategies seem to be safer, yet promising in terms of facilitating lie detection. Future research may shed light on this direction.

Commentary #9 by Sarid: The Challenge of Credibility Assessment of Victims who Experienced a Traumatic Offense

During my longstanding work with a large number of criminal investigators, I found that most of their attention is commonly directed towards identifying liars, who aim to conceal their criminal involvement. Importantly however, also victims may lie or mislead when filing a true complaint based on real events. This could be due to different reasons such as revenge, fear of getting hurt, or a desire to protect someone. Here, I would like to refer to another aspect related to victims, which I occasionally experience.

When filing a complaint regarding violent or rape events, both women and men provide a description of the criminal event which is often in contrast to the "tell it all" strategy that characterizes truth tellers. Specifically, the description is not always coherent in several aspects:

A. It contains gaps in schedules, lacks a chronological order, and the occurrence is told in irregular sections.
B. The events are not detailed and many facts are either missing or not reported.
C. There is a mental barrier to an explicit verbal description of sex offenses and of the actions of the aggressor.

D. While the case description includes many peripheral details, illustrating the context of the event, the core of the event lacks many details. In other words, there seem to be "black holes" in the victim's memory.

E. When the victim overcomes the inner barrier, many details are provided, some of them are dramatic in nature and include emotional elements.

F. Statements like "I'm just a soldier and he's an officer", "He's a very strong man" or "He's manipulative and will persuade you that it did not happen the way I say" are repeatedly heard.

Since verifying the details of the event is a critical stage in the investigative process, and since certain complainants are not eligible for a polygraph exam, often because of post-traumatic stress syndrome (PTSD), investigators seek alternative means of credibility assessment. This situation raises several questions. How can the factual basis of the complaint be separated from the addition of details designed to strengthen and intensify the weight of the complaint? Is it possible, and how, to separate the factual core from the influence of conceptual distortions, defense mechanisms or tactics to intensify the complaint?

The difficulty in addressing these situations stems, among others, from the fact that violence and rape victims express, as a result of the traumatic experience, characteristics attributed to deception. This includes lack of realism, especially in the estimation of times, as well as defense mechanisms, such as dissociation and repression, which may affect the ability to assess the reliability of a report in an attempt to separate experienced events from imaginative events. For these reasons, I believe that it is crucial

to develop measures to separate real statements from falsehoods among victims in large, and individuals with PTSD in particular.

Commentary #10 by Ashkenazi: The "Have You Ever?" Deception Detection Scenario

Typologies of lies have used several dimensions. Lies have been classified according to beneficiary, motivation, content, referent, and severity (Vrij, 2008). Such typologies are based on specific details, events, or persons involved. However, there are also some process-related typologies of deception (e.g., complexity or difficulty of the lie). One such major typology was coined by DePaulo et al. (1996) who differentiated between outright lies, exaggerations, and subtle lies. Outright lies are ones in which the liar presents the receiver with facts that sharply negate or contradict the truth, e.g. telling A did a certain action, while he actually did not (or conversably). A liar using exaggerations presents the facts in an over or understated way, e.g., exaggerates the intensity of certain feelings or underplays the amount of time spent on a certain activity. Subtle lies use statement-facts that can be interpreted as a literal truth, but are presented in a misleading way e.g. using words that have more than one meaning. Outright lies can further be categorized by lies of commission, omission (or a combination of both, e.g. false Alibi). Lies of commission are presenting events or facts that never happened. Lies of omission are concealing events or facts that did happened. Although the above presents different types of lies, what they all have in common is that the sender and the receiver are focused on a specific event that is anchored in a specific space- and time-frame.

The “Have You Ever?” Deception Detection Scenario I would like to present here is characterized by the fact that the receiver is not focused on a specific space- and time-anchored event. In this omission-type scenario the receiver is interested in the occurrence

of a certain type of event, asking the sender questions such as: Have you ever used drugs? Have you ever revealed classified information to an unauthorized person? And the sender's typical reply is: No. Contrary to the preceding types of lies, in this scenario, the interviewee (whether he is a liar or a true teller) is not presenting any "story" or report. Hence it is not clear how one can apply the existing cognitive-verbal deception detection methods or indicators, as all these methods are based on analysing content. Contributing to the difficulty here is the fact that if, for example, a person did reveal classified information to an unauthorized person, the receiver cannot know when and where it happened, what information was revealed, to whom and so on, which also makes it further unclear as to how to proceed in the interview after getting the answer "No".

The "Have You Ever?" scenario (using highly general questions) has a crucial importance for practitioners working in the intelligence field because the answers and conclusions derived from it can shape the nature of the relationship with the information provider, the risk assessment and risk taking procedures, and the value of other pieces of information he or she provides. This scenario becomes even more challenging because: (a) that guilty knowledge techniques usually cannot be applied here because liars are not all exposed to the same body of knowledge (using drugs is very versatile, and as mentioned above – unauthorized revealing of classified information has no common or even known characteristics), and (b) intention or opinion based techniques (e.g. devil's advocate; Leal, Vrij, Mann, & Fisher, 2010) do not apply because the concealed act does not necessarily reflect the actor's opinion, preference and even free will. To the best of my knowledge, the only deception detection method that declares its applicability to the "Have You Ever?" scenario is the polygraph test with its much controversial CQT protocol. Since for

practitioners this scenario is of great importance and need, this author sees it as the most urgent unsolved question in verbal lie detection.

Commentary #11 by Nisin: Theory - Protocol - Procedure (TPP) paradigm for the implementation of credibility assessment tools.

To me, as a practitioner, an application of credibility assessment (henceforth CA) program would be no less than a standardized and organization-wide implementation of authorized operational procedures. Seldom, local scale, personal-initiative-based use of any lie detection tool cannot be regarded as institutionalized implementation in real life settings. Following the relevant body of research, it was the verbal path of credibility assessment in content (henceforth VCA) that emerged to be the most promising in the field, passing easily the questionable behavioral path of assessment. Yet, until today, only the path of Psychophysiological Detection of Deception (PDD) have reached the status of fully established CA programs all through security and law enforcement agencies around the world. Despite the potential of the verbal path, there is no full-scale VCA program known to me to be applied on an organizational level. The reasons may be because the VCA still has to resolve significant issues, such as embedded lies, or individual case decision rules. I believe that the biggest gap yet to overcome is the absence of adapted to field challenges VCA protocols of engagement with the interviewee.

Field challenges are a variety of situations and contexts, in which credibility assessment is applied. A specific challenge (a unique context and situation), is translated to *institutional-level* operational procedures, determining the way to carry out credibility assessment process in that specific challenge, detailing the activities in accordance with the organizational missions, policies, legal restrictions, and regulatory requirements. Protocols

are *interpersonal-level* step-by-step instructions, with a definite start and end points, that describe clearly and precisely the course of action and the sequence of activities that must be followed to apply accurately and efficiently the CA technique in the specific challenge. For imaginary example, let us look at the challenge of "preventing hazard on board an airplane", translated into a procedure of approaching every passenger by a VCA expert with preliminary questioning prior to their check-in, using the Verifiability Approach protocol for an airport security (Nahari, 2018c), which will detail how to engage the passengers, how to present them with a fixed sequence of relevant questions and stimulations, and how to analyze their verbal responses. The procedure may further establish a scoring cut-off for "clearance" and require following questioning, sometimes even with the use of a different protocol of engagement, for passengers, who did not get "clearance" at the preliminary stage.

To reach their *institutional-level* aims the practitioners are in a constant need for valid and reliable means, hence should look for research-validated instruments, with solid theoretical and empirical underpinning. In that regard, VCA tool-box is the richest and is updated constantly by contemporary research, refining the relevant criteria definition and identification marks, enhancers and inhibitors of accurate assessment, sensitivity and specificity tradeoffs, etc. Often this tool-box will include guidelines and even specifics to support protocols of engagement with the "to be assessed for credibility" targets. Yet, adaptations of these *interpersonal-level* protocols must be made to meet the specific challenge ahead, since theory-based protocols often are lacking contextual and situational adaptations. For example, most of the VCA protocols are based on "free narrative", while in a forensic context, during an interrogation of suspects, after following the legal

procedures and presenting them their rights, the received narrative can hardly be called ”free”. Another issue is the timeframe of existing VCA protocols which does not always meet the forensic nor the security timeframe restrictions.

The *theory-level* tool-box, the *interpersonal-level* protocol, and the *institutional-level* procedures are represented by Theory - Protocol - Procedure (TPP) paradigm (see Figure 3) for implementation of CA tools. The TPP paradigm conceptualizes my perspective on how any CA program should be implemented in the field.

VCA possesses significant advantages, using a ubiquitous platform of human utterance as the raw material for its analysis. No High-Technology and no special skills are required, but the mere ability for conversation, making the possibility of the analysis available almost everywhere, anytime and at a very low cost. The robust body of research is yet another quality to promote the verbal path. To reach the status of a fully deployed VCA program the verbal path should incorporate the ecological specifics of targeted challenges, which can be accessed through collaboration with practitioners.

As a practitioner, I wish VCA research to address the issue of ecologically valid protocols, preferably through a systematic review of field challenge’s factors, such as timeframe restrictions, preventative vs. reactive context, or interrogation vs. screening settings.

The workgroup messages to researchers in verbal lie detection field

During the three information-packed days of the workshop, we experienced a variety of knowledge-exchange modalities (lectures, group problem-solving discussions, and one-on-one interactions), from different professional perspectives (researchers and

practitioners), and across several methodologies (experimental, survey, and artificial intelligence). Naturally, the specific topics we discussed also varied considerably, but they all converged around three over-riding themes: Interactions between researchers and practitioners, the need to moderate global findings by the specific investigative context, and the difficulties of classifying individual cases. Based on these themes, we formulated three messages to researchers in the field.

1. **The need for theoretically-based and ecologically-adaptable solutions to field challenges**

A prime role of psychological research is to provide a theoretical understanding of phenomena and behaviors. In verbal lie detection, a field intensively developed in the last few years, we are still establishing a comprehensive understanding of deceptive and truthful behaviors (see Commentary #2 above). While the importance of this understanding stands alone, we recognize that it also serves research applications, such as verbal lie detection techniques. We further realize that for achieving effective applications, suitable to be implemented in the field, one should first define the specific field challenges and then provide not only *theoretically-based* but also *ecologically-valid*, solutions (see Commentaries #1 and #11 above). This task foresees a cooperation with practitioners. However, to achieve maximal effectiveness it is crucial to specify the timing and the framework of such cooperation.

The Theory - Protocol - Procedure (TPP) paradigm, proposed by Nisin (see Commentary #11 above), portraits the field implementation of a credibility assessment technique as organization-specific since every organization has to deal with a designated

challenge. To fulfill this mission the organization will approach a designated specialist within his employees for solutions. The specialist should refer first to relevant techniques and their protocols (the technique may be associated with several protocols), as were developed and validated in the academia, sometimes just to find out that they should be adjusted to the operational environment and to the definite requirements of the particular organization. For example, consider that the application of a technique requires a couple of hours, while this is not possible in the specific organizational settings because of limited human resources, or a tight timetable. In such a case, the existing protocols should be shortened to meet the organizational timeframe. Such an adjustment can be then validated empirically. Such an adjustment can be validated, theoretically and empirically, exclusively by researchers. Obviously, a solid theoretical underpinning will enable easier adaptations of existing protocols to new challenges as well as successful development of new ones (see Commentary #2 above).

Once the protocol was adjusted to the specific organizational demands, the specialist can proceed to its assimilation in their specific organizational settings, authorizing the operational procedures. Those operational procedures will determine what work should be performed, for what reason, in what way and by whom. To explain further the difference between protocol and procedure, consider that an identical *protocol* is applied in two credibility assessment contexts: forensic and border security. In both contexts, the instructions for the interviewee, the questions asked and their order, and the coding system applied are all the same. Yet, the forensic context demands fewer "false alarms" (i.e., false-positive errors) while the security context demands the opposite – fewer "misses" (i.e., false-negative errors). Consequently, to meet their challenges, the forensic

procedure will authorize more lenient cut-off points for "incrimination" and the security procedure will establish harsh bar of "clearance". As such, theory-driven data collection rules are part of the universal protocol, while challenge-driven decision rules are part of organization-specific procedures. Apparently, the theoretical establishment and empirical validation of techniques are exclusively an academic playground as much as the organizational procedures are the territory of the practitioners only, leaving the establishment of operational and valid protocols as the only possible rendezvous point.

The rendezvous stress out the need for a common "language" between the academic researchers and the field specialists. This can be achieved by collaborative effort to define operationally the essence of the field challenges. Field challenges are imposed by the reality, which in turn determines the *context of the evaluation* (e.g., police investigative interview, intelligence information gathering and verifying, pre-employment interviewing) and the *environment of the evaluation* (e.g. interviewee status, interviewee demographic profile, and characteristics of physical environment). Thus, the researcher–specialist dialog can be conducted across the prime axes of the reality determinants. During the workshop, we identified several such axes: the *target event*, being physical (e.g., criminal act, conversations) or mental (e.g., intentions, attitudes, guilty knowledge); the risk management considerations of *accuracy level* (e.g., false positive vs. false negative errors tradeoff); the *resolution of the event*, being a token event (i.e., the interview is regarding the carrying out of a space- and time-anchored misdeed) or type event (i.e., the interview is regarding the carrying out of a misdeed in an unknown time and location); and the *diagnostic stage* being a secondary-order (i.e., preliminary group-level screening) or a primary-order (i.e., final individual-level decision). We are aware of the defined axes being

preliminary and in need of further discussion, yet we acknowledge that defining the prime axes will promote the development of challenge-focused rather than scenario-focused methods. Taking as an example, the "Have you ever…?" operational challenge, described by Ashkenazi (see Commentary #10 above), we expect a dialog between a specialist and a researcher to result in defining the challenge as a *'physical event'*, which requires *'high accuracy'*, to minimize both false positive and negative errors, for a *'type event'* scenario, requiring a *'primary-order decision'*. Once a specific challenge is operationally defined, one can start the process of constructing a solution.

During the workshop, we recognized that the course of cooperation between researchers and field specialists described here is an effective practice for the development and application of theoretical-based techniques to real-life challenges. Yet, a necessary condition for such a cooperation course to occur is the accessibility of the scientific knowledge to field specialists. We are aware of the fact that in search for solutions to field challenges, it is more likely for the organizations to turn to industry rather than to academia. We therefore believe there is an urgent need to find ways to make the scientific knowledge more accessible to the field and to construct permanent platforms for the desired cooperation.

2. The "one size fits all" approach

A lie detection tool that is theoretically sound and usable in all circumstances is fantasy. Nor is it likely to be developed, since deception and lie detection is too complex. As researchers, we must find ways to convey to practitioners theoretically sound techniques

along with a clear understanding of when they can and cannot be used. And we need to convince practitioners not to discredit techniques that cannot be used all the time.

Verbal lie detection tools—like all other lie detection tools—have several restrictions and we will outline four of them. First, at least some verbal cues are culturally specific (Taylor, Larner, Conchie, & Van der Zee, 2014). Cross-cultural variations in the language of liars has been found that are consistent with known cultural differences in self-construal and episodic memory (Taylor, Larner, Conchie, & Menacere, 2017). For example, individualistic White British participants reduced their first-person pronoun use when lying compared to telling the truth. By contrast, collectivist North African participants increased their use of first-person pronouns when lying, in part to compensate for their reduction in use of third-person pronouns and references to family.

Second, some verbal cues are age specific. For example, it is understood that the verbal lie detection tool Reality Monitoring cannot be used with younger children because they have more difficulty than adults to distinguish fact from fantasy (Lindsay, 2002; Lindsay & Johnson, 1987).

Third, verbal cues to deceit are context specific. For example, although truth tellers often report more details than liars when describing past activities, this finding does not necessarily occur when they discuss future activities (Sooniste, Granhag, Knieps & Vrij, 2013). Similarly, collective interviewing research (when pairs of interviewees are interviewed together) has shown that truth tellers communicate more with each other when discussing jointly experienced past activities than fabricated activities (Vernham & Vrij, 2015). However, this finding cannot be automatically generalized to instances where they discuss future actions (for example at border crossings).

Finally, verbal lie detection tools are more or less effective depending on the type of lie that is told. Two types of lies that we believe to be particularly difficult to detect are embedded lies and omissions. An example of an embedded lie is reporting a truthful experience (for example visiting a restaurant) while lying about when the visit took place. Verbal lie detection tools can be used in such situations but only if an investigator asks questions about the deceptive element of the story: When the experience took place. Omissions refer to deliberately not reporting a specific activity. For example, when asked to describe all activities during a specific day, a liar could be entirely truthful but leave out a crucial interaction with a specific person. Research into how to detect omissions is needed urgently.

We encourage researchers and practitioners to work closely together and to think of circumstances in which verbal lie detection is important, but in which the verbal lie detection techniques which have been developed to date cannot be (easily) used (see Commentaries #9 and #10 above). We encourage researchers to think of innovative ways to detect deceit in those circumstances. We expect this to be challenging at times and there is no success guaranteed. However, this should not refrain researchers from attempting to find solutions.

3. The group to individual inference challenge

One of the topics addressed during the workshop is how to convert group averages to individual classification (see Commentary #3 above), a challenge also referred to as group to individual inference (G2i; Faigman, Monahan, & Slobogin, 2014). The considerable heterogeneity between studies makes it difficult to specify an adequate cutoff point. A cutoff resulting in acceptable sensitivity and specificity in one study may yield an

unacceptably high error rate in another. We suggest three research lines that could help with the G2i challenge. These lines seek to reduce or explain the heterogeneity across studies, increasing the generalizability of cutoff points.

The first line of research entails making use of within examinee comparisons (see also Vrij, 2016). Within examinee comparisons to a large extent determine the internal validity of deception tests (Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakhar, 2016), and examples of verbal tools that already include such comparison are the Verifiability Approach (Nahari, Vrij, & Fisher, 2014) and the Strategic Use of Evidence (Granhag, & Hartwig, 2015). A case in point here is that frequently used verbal tools such as CBCA and RM primarily include truth criteria (see Commentaries #4 and 5 above). That is, the presence of cues indicates truthfulness, yet the absence of these cues does not necessarily indicate deception (Steller & Wellershaus, 1992; Rassin, 2000). Adding lie criteria to these tools would allow for a within examinee comparison; if more truth than lie criteria are present this indicates deception, and vice versa. Examples of potential lie cues that could be empirically validated include being overly consistent, using more abstract language, question repetitions (to have time to make up a deceptive response), a lack of proportionality in responses, and unnecessary explanations and justifications. For example, while truth tellers might give explanations relative to *how* they performed some actions, liars might explain *why* they did so. Although such a within examinee comparison does not eliminate the need for a cutoff point (i.e., one still needs to decide how much of a lie-truth difference needs to be present for a lie judgement), such an approach should reduce variance between studies, making decision rules more generalisable.

A second line of research that could help with the G2i challenge involves standardizing the coding schemes (see Commentary #6 above). Coding of statements can require a number of choices, and a simple sentence such as 'in the green house' could, for example, be coded as three details (the *house*, which is *green*, and the location is specified as *in* the house), or two details (e.g., because the word *in* does not add information value and/or follows from the previous sentence it does not count as a separate detail). When researchers compare different conditions within one study (e.g., with versus without model statement; expected versus unexpected questions), these decisions do not necessarily threaten the internal validity of the study. As long as within one study the coding is performed systematically, any effect reported can be considered a true effect. Such coding inconsistency does, however, add to the lack of generalizability of a decision rule, especially when criteria are expressed in absolute numbers (Vrij, Leal, Jupe, & Harvey, 2018). As verbal lie detection encompasses a wide variety of cues, an initial step towards standardization could entail focusing on amount of (different types of) details, as this criterion is both included in many tools, and has been shown most robust in discriminating between deceptive and truthful statements.

A third line of research could examine moderators of verbal criteria either within or between methods. By identifying critical moderators, researchers can either identify criteria that are resistant to variance across a moderator, or propose adjustments to the cut-off point to take into account the role of that moderator. Research has already shown that verbal cues will vary across individual differences (see e.g., Nahari & Pazuelo, 2015; Schelleman-Offermans & Merckelbach, 2010), culture (Taylor, Larner, Conchie, & Menacere, 2017), and lie or scenario type (see e.g., Vrij, Granhag, Mann, & Leal, 2011).

To narrow down the potential number of moderators, it will be important for researchers to focus on those that theory suggest as likely to drive observed differences. For example, a comparison between two cultural groups without a rationale for why there might be a difference is less insightful that a comparison grounded in a known cultural difference in social norm or cognition. The former at best tells us that there is a difference (or not) across the two groups; the latter allows for inferences about other groups and contexts that vary on the same theoretical dimension.

We proposed three research lines that could help with the G2i challenge. It is important to note that their implementation comes with a delivery challenge. Within examinee comparison takes time, more refined coding may be harder to deliver, and introducing moderators adds to the complexity of a technique. This is antithetical to the desire for simple, easy to deliver methods in practice. We suggest the balance to be struck is something that must emerge over time.

Summary

Although we addressed many areas during the workshop, we certainly left a vast array of issues to be examined in the future. Much research still remains to be conducted, and we invite you to extend our efforts by refining and critiquing the work we presented and to explore other important issues we did not examine.

References

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, *7*, 3-12. Doi:10.1016/j.ejpal.2014.11.002

Bogaard, G., Meijer, E., Vrij, A., & Merckelbach, H. (2016). Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event. *Frontiers in Psychology, 7*, 243. Doi: 10.3389/fpsyg.2016.00243

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Routledge. doi:10.1016/b978-0-12-179060-8.50006-2

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*, 979–995.

DePaulo, B. M., Lindsay, J. L., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129,* 74-118. doi:10.1037/0033-2909.129.1.74

Faigman, D.L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review, 81*, 417-480.

Guttman, L. (1988). Eta, disco, odisco, and F. *Psychometrika, 53,* 393-405. doi:10.1007/bf02294220

Granhag, P. A., & Hartwig, M. (2015). The Strategic Use of Evidence (SUE) technique: A conceptual overview. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Deception detection: Current challenges and new approaches* (edn, pp. 231 – 251). Chichester, England: Wiley.

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*, 307-342. Doi: 10.1177/1088868314556539

Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of Criteria-based Content Analysis. *Psychological Assessment, 29*, 819-834. Doi:10.1037/pas0000426

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

Kleinberg, B., Nahari, G., Arntz, A., & Verschuere, B. (2017). An Investigation on the Detectability of Deceptive Intent about Flying through Verbal Deception Detection. *Collabra: Psychology*. http://doi.org/10.1525/collabra.80

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*. http://doi.org/10.1002/acp.3407

Kleiner, M. (2002). *Handbook of polygraph testing*. San Diego, CA: Academic Press.

Leal, S., Vrij, A., Mann, S., & Fisher, R. P. (2010). Detecting true and false opinions: The devil's Advocate approach as a lie detection aid. *Acta Psychologica*, *134*, 323-329.

Lindsay, D. S. (2002). Children's source monitoring. In H. L. Westcott, G. M. Davies, & R. H. C. Bull (Eds.), *Children's testimony: A handbook of psychological research and forensic practice* (pp. 83-98). Chichester: Wiley and sons.

Lindsay, D. S., & Johnson, M. K. (1987). Reality monitoring and suggestibility: Children's ability to discriminate among memories from different sources. In S. J. Ceci, J. Toglia, & D. F. Ross (Eds), *Children's eyewitness memory* (pp. 91-121). New York: Springer-Verlag

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,* 99–122. Doi:10.1080/10683160410001726356

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*, *53*, 593-604.

Mihalcea, R., Narvaez, A., & Burzo, M. (2014). A Multimodal Dataset for Deception Detection. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

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.

Nahari, G., Vrij, A., & Fisher, R. P. (2014). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology, 19*, 227–239

Nahari, G. (2018a). Reality Monitoring in the Forensic Context: Digging Deeper into the Speech of Liars. *Journal of Applied Research in Memory and Cognition, 7,* 432 — 440.

Nahari, G. (2018b). The applicability of the verifiability approach to the Real world. In J. P. Rosenfeld (Ed.), *Detecting Concealed Information and Deception: Recent Developments* (pp. 329 - 349). London: Elsevier

Nahari, G. (2018). Verifiability approach: Applications in different judgmental settings. In T. Docan-Morgan (Ed.). *The Handbook of Deceptive Communication*. Hampshire: Palgrave Macmillan

Nahari, G., & Nisin, Z. (2018). Digging further into the Speech of Liars: Future Research Prospects in Verbal Lie Detection. *Frontiers in psychiatry. Manuscript submitted for publication.*

Ott, M., Cardie, C., Hancock, J. T., Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. http://doi.org/10.1016/j.scitotenv.2014.07.054

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 309–319). Association for Computational Linguistics.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. *IEEE International Conference on Multimodal Interaction*. http://doi.org/10.1063/1.4816640

Rassin, E. (2000). Criteria based content analysis: The less scientific road to truth. *Expert Evidence, 7*, 265-278.

Schelleman-Offermans, K., & Merckelbach, H. (2010). Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, *7*, 247-260.

Snook, B., Zito, M., Bennell, C., & Taylor, P. J. (2005). On the complexity and accuracy of geographic profiling strategies. *Journal of Quantitative Criminology, 21,* 1-26. doi: 10.1007/s10940-004-1785-4.

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, 673-685. Doi: 10.1080/1068316X. 2013.793333

Steller, M., & Köhnken, G. (1989). Criteria-Based Content Analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217-245). New York, NY: Springer-Verlag.

Steller, M. & Wellershaus, P. (1992). Information enhancement and credibility assessment of child statements: The impact of the cognitive interview technique on Criteria-based Content Analysis. In G. Davies, S. Lloyd-Bostock, M. McMurran, and C. Wilson (Eds.), *Psychology, law, and criminal justice. International developments in research and practice* (pp. 118-126). Berlin: Walter de Gruyter.

Taylor, P. J., Larner, S., Conchie, S. M., & Menacere, T. (2017). Culture moderates changes in linguistic self-presentation and detail provision when deceiving others. *Royal Society Open Science*, *4*, 170128. Doi: 10.1098/rsos.170128.

Taylor, P. J., Larner, S., Conchie, S. M., & Van der Zee, S. (2014). Cross-cultural deception detection. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Deception detection: Current challenges and cognitive approaches* (pp. 175-202). Chichester: Wiley-Blackwell.

Vernham, Z., & Vrij, A. (2015). A review of the collective interviewing approach to detecting deception in pairs. *Crime Psychology Review*, *1*, 43-58.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester, UK: Wiley.

Vrij, A. (2016). Baselining as a lie detection method. *Applied Cognitive Psychology, 30,* 1112-1119.

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

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Lying about flying: The first experiment to detect false intent. *Psychology, Crime & Law, 17*, 611–620.

Vrij, A., Leal, S., Fisher, R. P., Mann, S., Dalton, G. Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2018a). Sketching as a technique to elicit information and cues to deceit in interpreter-based interviews. *Journal of Applied Research in Memory and Cognition, 7,* 303-313. Doi: 10.1016/j.jrarmac.2017.11.001

Vrij, A., Leal, S., Jupe, L., & Harvey, A. (2018). Within-subjects verbal lie detection measures: A comparison between total detail and proportion of complications. *Legal and Criminological Psychology, 23*, 265-279.

Vrij, A., Leal, S., Mann, S., Fisher, R. P., Dalton, G. Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2018b). Using unexpected questions to elicit information and cues to deceit in interpreter-based interviews. *Applied Cognitive Psychology*, *32*, 94-104. Doi: 10.1002/acp.3382

Vrij, A., Leal, S., Mann, S., Fisher, R. P., Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2018c). *Eliciting information and cues to deceit through sketching in interpreter-based interviews*. Manuscript submitted for publication.

Vrij, A., Leal, S., Mann, S., Dalton, G. Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2017). Using the Model Statement to elicit information and cues to deceit in interpreter-based interviews. *Acta Psychologica*, *177*, 44-53. Doi: 10.1016/j.actpsy.2017.04.011

Table 1.

Mean, SD, Cohen d, and Discrimination Statistics for Available Studies in Model Statement and Unanticipated Questions.

Paper	Mean (SD)		*d*	U3	DISCO
	Truth	Lie			
Studies of model statement					
Ewens et al. (2016) Interpreter	67.07 (27.97)	27.97 (22.65)	0.70	.758	.470
Ewens et al. (2016) English	80.73 (26.43)	62.80 (23.04)	0.72	.764	.610
Ewens et al. (2016) Non-native English	51.74 (22.11)	37.13 (19.71)	0.70	.758	.560
Harvey et al. (2017)	50.38 (42.80)	11.05 (16.01)	1.22	.888	.880
Kleinberg (2017) Exp. 2.[1]	243.49 (170.3)	287.0 (179.1)	-0.25	.599	.230
Porter et al. (2018). Spatial statement.[1]	148.00 (75.66)	65.19 (42.19)	1.35	.912	.830
Porter et al. (2018). Temporal statement.[1]	94.09 (45.25)	61.05 (39.99)	0.77	.779	.680
Vrij et al. (2018)	85.93 (65.62)	65.27 (39.93)	0.39	.652	.310
Vernham et al. (2018) Total interactions	80.63 (36.53)	68.46 (33.89)	0.35	.637	.180
Leal et al. (in press) Total new details	70.00 (59.71)	48.19 (29.16)	0.47	.681	.620
Studies of unexpected questions					
Vrij et al. 2009	4.78 (1.50)	3.30 (1.00)	1.19	.883	.680
Warmelink et al. (2012). Experienced	122.77 (70.01)	83.46 (41.79)	0.68	.752	.160
Warmelink et al. (2012). Not experienced	92.75 (35.16)	100.59 (60.68)	0.16	.564	.250

Lancaster et al. (2013). Temporal question.[1]	28.20 (13.67)	23.00 (11.65)	0.41	.659	.580
Lancaster et al. (2013). Fixed perspective 1.[1]	2.70 (1.97)	2.30 (2.06)	0.20	.579	.020
Lancaster et al. (2013). Fixed perspective 2.[1]	2.50 (1.97)	2.20 (2.05)	0.15	.560	.230
Lancaster et al. (2013). Fixed perspective 3.[1]	7.00 (4.81)	5.80 (3.52)	0.28	.610	.220
Shaw et al. (2013). Reverse order questions	17.53 (7.20)	16.10 (8.10)	0.19[3]	.575	.220
Vrij et al. (2018)	31.68 (20.87)	32.22 (16.72)	0.03	.512	.100
Jupe et al. (2018). Using RM coding	496.53 (309.78)	460.36 (345.40)	0.11	.544	.260

Note: [1]Total of categories reported. [2]The authors do not report the Means or SD needed. [3]Calculated assuming lie and truth groups contained half the participants. Some papers were excluded due to qualitatively different methodology; see Supplementary Materials for details.

Figure 1. Synthetic data approximating the results from 10 model statement studies shown in Table 1

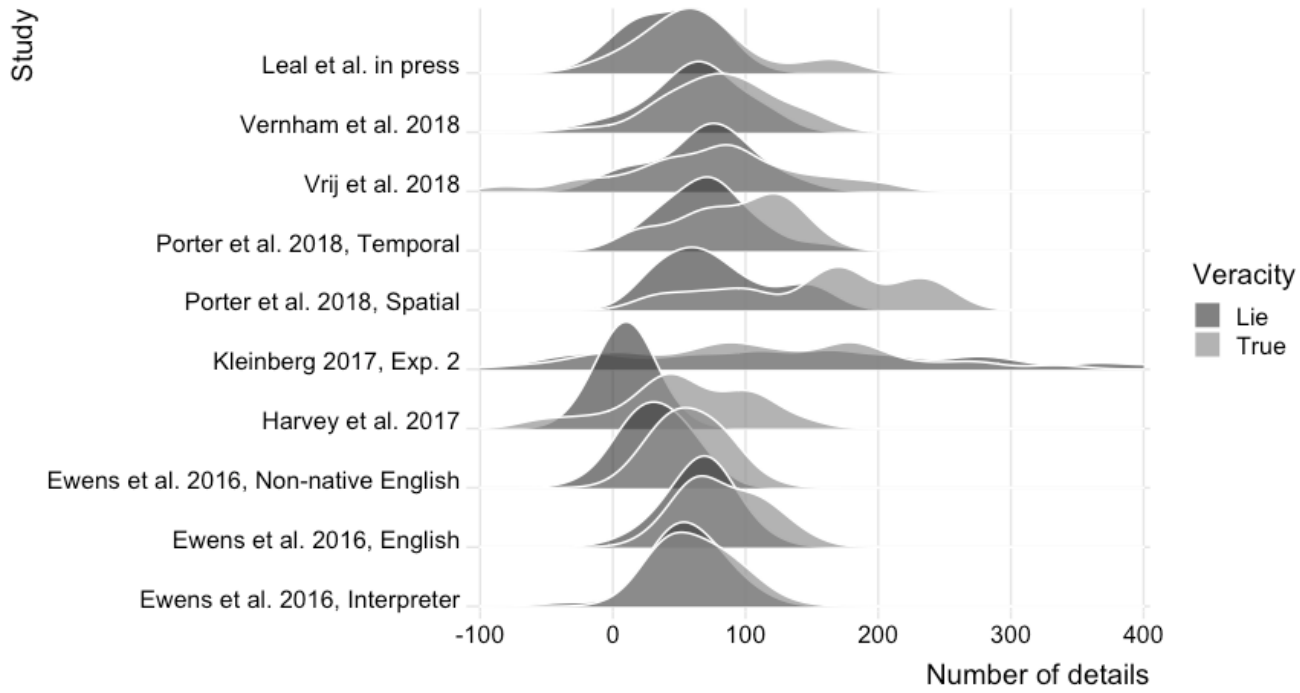

Figure 2. Synthetic data approximating the results from 10 unanticipated question studies shown in Table 1

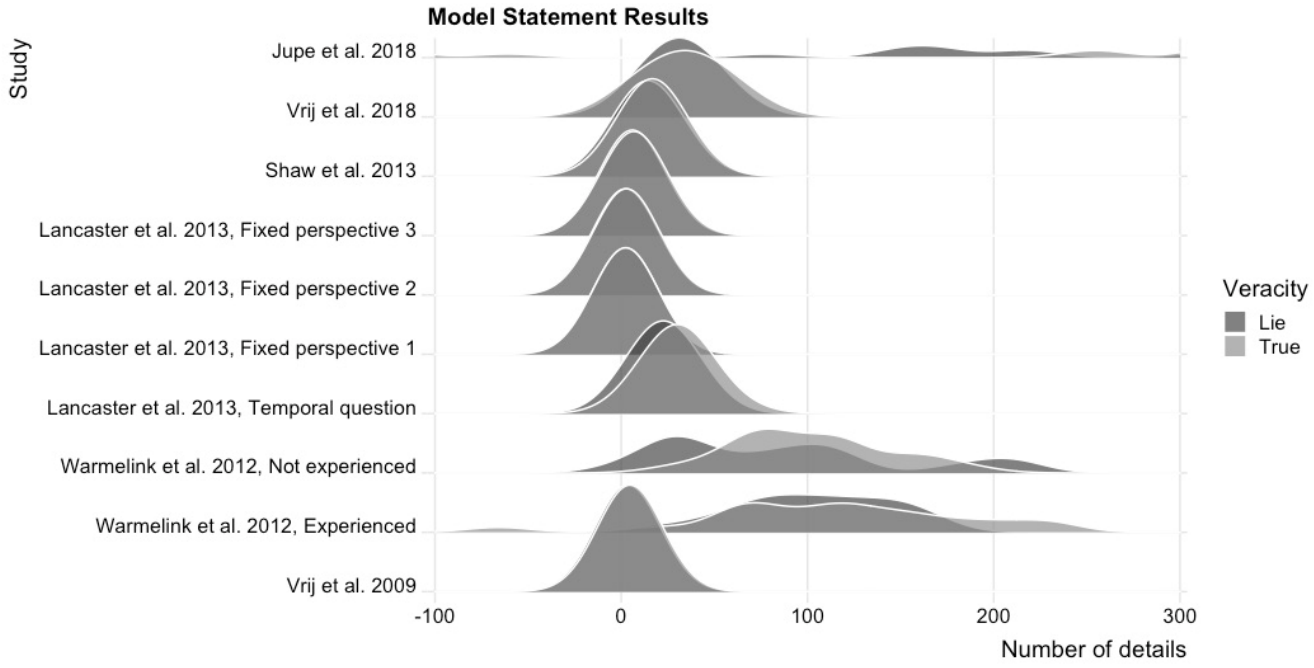

Figure 3: Theory - Protocol – Procedure paradigm for implementation of credibility assessment.