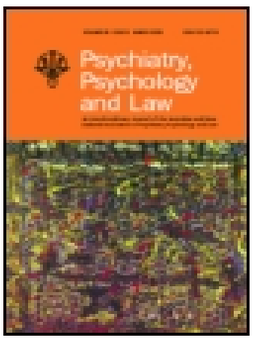

Psychiatry, Psychology and Law

ISSN: 1321-8719 (Print) 1934-1687 (Online) Journal homepage: http://www.tandfonline.com/loi/tppl20

# True and false alibis among prisoners and their detection by police detectives

Ricardo Nieuwkamp, Robert Horselenberg & Peter van Koppen

**To cite this article:** Ricardo Nieuwkamp, Robert Horselenberg & Peter van Koppen (2018): True and false alibis among prisoners and their detection by police detectives, Psychiatry, Psychology and Law, DOI: 10.1080/13218719.2018.1482570

**To link to this article:** https://doi.org/10.1080/13218719.2018.1482570



Published online: 26 Jun 2018.

*Psychiatry, Psychology and Law*, 2018
https://doi.org/10.1080/13218719.2018.1482570



# True and false alibis among prisoners and their detection by police detectives

Ricardo Nieuwkamp[a], Robert Horselenberg[a] and Peter van Koppen[a,b]

[a]*Department of Criminal Law and Criminology, Faculty of Law, Maastricht University, Maastricht, The Netherlands;* [b]*Department of Criminal Law and Criminology, Faculty of Law, VU University, Amsterdam, The Netherlands*

The present study was designed to determine whether differences exist between true and false alibis and how accurate police detectives and lay people are in determining the veracity of alibis. This article provides a replication of the research by Culhane et al. (2013) with more representative participants. In the first experiment, real suspects in a remand prison generated true or false alibis. In the second experiment, a subset of those alibis were written out and were provided to experienced police officers and students for alibi evaluation and discrimination. Our results show that differentiating between true and false alibis is difficult, and even when more representative materials and participants are included, the accuracy did not exceed 60%. Interestingly we found that students and police officers focus on other aspects during the alibi discrimination. Thus, research using student participant cannot be, directly, used in alibi discrimination studies.

**Key words:** alibis; police detectives; prison; students; true and false statements; veracity; verbal lie detection.

The research on discrimination between true and false alibis is scarce (Culhane et al., 2013; Nahari & Vrij, 2014; Porter & Yuille, 1996; Strömwall, Granhag, & Jonsson, 2003). This article describes two experiments aiming for more representative alibi discrimination research using real suspects and police detectives as a sample.

One of the most important limitations in all studies on alibis is the lack of representative samples of participants (e.g. Eastwood, Snook, & Au, 2016). The experiments in the present paper are inspired by the article of Culhane et al. (2013), in which the authors first asked students to generate either true or false alibis. A subset of these alibis was then video-recorded and was subsequently shown to another group of students who were asked to make two piles: those that were true and those that were false. In real life, police detectives are the first to evaluate the alibis. Given the fact that a wrongful classification of the veracity of a suspect's alibi has been an important factor in the review of wrongful conviction cases (e.g. Simon, 2012; Wells et al., 1998), there is an urgent need to conduct research on this topic with a representative sample of participants. This is the main aim of this article.

## *Alibi research*

Three broad domains can be identified in the research on alibis: generation, evaluation and

Correspondence: Ricardo Nieuwkamp, Department of Criminal Law and Criminology, Faculty of Law, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: ricardo.nieuwkamp@maastrichtuniversity.nl

discrimination of alibis (Burke, Turtle, & Olson, 2007; Culhane et al., 2013). In alibi generation studies, non-offenders are asked to generate an alibi and supportive evidence for a specific period in time. The veracity of the alibi is determined in the alibi discrimination domain. In alibi evaluation studies, often mock jurors determine the alibi believability based on the overall believability, strength and ease by which the evidence can be fabricated (e.g. Jung, Allison, & Bohn, 2013).

In the criminal justice system there appears to be an assumption that innocent people usually generate an accurate and believable alibi (Olson & Charman, 2012). This means that the alibi should be correct and supported by strong evidence. True alibis can, however, also be inaccurate because of various reasons (Crozier et al., 2017). It has been demonstrated that the larger the time gap between the alleged crime and the moment of interviewing, the more errors occur in the alibi statement of non-offenders (e.g. Olson & Charman, 2012). Furthermore, a suspect might be simply mistaken about their whereabouts when at the time of the alleged crime they were involved in a less common event (e.g. Strange, Dysart, & Loftus, 2014). Lastly, a suspect may also lie about an alibi to conceal his or her involvement in another, but shameful situation (Nieuwkamp, Horselenberg, & Van Koppen, 2016). The assumption that non-offenders are able to present strong supportive evidence for their alibis has proven to be incorrect (Culhane, Hosch, & Kehn, 2008; Culhane et al., 2013; Olson & Charman, 2012). In fact, only 3% of the non-offenders can present strong evidence for their alibis (Nieuwkamp, Horselenberg, & Van Koppen, 2017).

### *Supportive evidence*

The strength of the supportive evidence can be determined using the alibi taxonomy of Olson and Wells (2004). In the taxonomy, two types of evidence to support an alibi are distinguished: witness evidence and physical evidence. According to the taxonomy, the testimony of an alibi witness is strong supportive evidence when the relational distance between the witness and the suspect is large (e.g. the suspect's neighbour or other unmotivated witness; Olson & Wells, 2004). The witness's testimony is weaker evidence when the relational distance is small (e.g. the suspect's girlfriend; Culhane & Hosch, 2004), or another motivated familiar witnesses. The rationale behind determining the strength of the witness evidence based on relational closeness is that the closer the relational distance between the suspect and the witness, the more inclined the witness may be to have a motive to lie in favour of the suspect. Police officers, however, focus more on the number of witnesses supporting an alibi than on the relational distance between the witnesses and the suspect (Eastwood et al., 2016).

Physical evidence is considered strong supportive evidence when three criteria are met. The evidence needs to (a) be linked to the suspect; (b) contain time information; and (c) contain location information (Olson & Charman, 2012). That is, it has to locate the suspect at a specific place at a specific time. An example of strong evidence is CCTV-footage from a security camera (Olson & Wells, 2004). A receipt is an example of weak physical evidence, because a receipt usually lacks a link to the one who purchased the goods. The rationale behind the three criteria is that the more difficult it becomes for the suspect to fabricate the physical evidence or to obtain it from someone else, the stronger the evidence becomes. The impact of physical evidence cannot be underestimated because weak physical evidence (i.e. a receipt) has a greater positive influence on the alibi believability than strong witness evidence (i.e. an unmotivated familiar other witness; Olson & Wells, 2004).

A suspect can also have a combination of witness and physical evidence for his alibi (Culhane et al., 2013). The strength of a combination of supportive evidence can also be determined using the taxonomy of Olson and

Wells (2004). The strength of the combination depends on the strength of the witness and physical evidence separately. For example, weak witness evidence combined with weak physical evidence is equally as strong as having strong physical evidence without witness evidence (Olson & Wells, 2004).

### *The studies by Culhane et al. (2013)*

Having knowledge about a non-offender's alibis is fundamental to differentiate between true and false alibis as argued by Olson and Charman (2012, p. 464): 'The perception in the legal system that alibis have diagnostic value is critically dependent on the assumption that innocent people should be able to produce relatively strong and accurate alibis; otherwise, alibis would be useless as tools to differentiate the innocent from the guilty.' Besides the generated alibi that needs to be correct, the presence and strength of the evidence are the most important factors for a believable alibi (Pozzulo, Pettalia, Dempsey, & Gooden, 2015). As stated above, the research on alibi generation shows that non-offenders are more likely to present weak supportive evidence rather than strong supportive evidence (e.g. Culhane et al., 2013). That makes distinguishing between true and false alibis difficult, although by definition all non-offenders presented a true alibi. It leads scholars to conclude that a discrepancy exists between what alibis non-offenders can present and what alibis are found to be believable (Culhane et al., 2008).

The most important limitation of alibi research in general and in particular in the field of alibi generation studies is the use of undergraduate students (Eastwood et al., 2016) because older adults 'are more likely to have additional family members (spouse and children) [who could support their alibis] and less likely to spend the same amount of time with friends' (Culhane et al., 2013, p. 627). In the present study, we tested alibi generation with a sample of non-students – namely, recently arrested male suspects – generating true and false alibis. So, we tested whether the results found by Culhane et al. (2013) among undergraduate students also hold for criminals and police officers using their general methodology for the experiments in the present paper. In order to do so, a short overview of the study by Culhane et al. (2013) is provided, followed by the most important adaptations in the present paper. The rationale behind these decisions is explained in the experiment described below. In the paper of Culhane et al. (2013) two experiments are described, one on alibi generation and the other one on alibi discrimination. In the first experiment, students were asked to provide either a true or a false alibi for an evening either five or 12 days ago. All participants were subsequently asked to collect all the reported evidence for their alibis and to return 48 hours later. Inconsistencies in the alibis were observed in 6.5% of the cases, and truth tellers are most often at home with friends or family. Those who reported a false alibi all reported at least one friend as a witness. In addition, false alibis consisted of more words than true alibis. The delay had no effect on the results other than that liars reported more witnesses when the delay increased.

Six true and six false statements were subsequently video-recorded and used as materials in the second experiment. In the second experiment, a second group of students was shown the videotaped generated alibis and were asked to differentiate between the true and false alibis. A low accuracy (43.3%) was found, where true alibis were more often classified correctly (57.4%) than false alibis (29.1%). While males were better at detecting false alibis, females were better at detecting true alibis.

In the current experiments we wanted to increase the ecological validity by improving the selection of participants for the experiments. In the first experiment, in which alibi generation was studied, we asked recently arrested suspects to present a true or a false alibi. In the second experiment we compared the performance of police detectives to that

of students in accurately discriminating between true and false alibis. Police detectives were chosen as participants since their role is very important in the criminal pre-trial investigation. Depending on whether they consider the alibi to be believable or not, the police officers may invest more or less time and energy investigating the suspect's potential involvement in the alleged crime. By doing so, they influence the further procedure.

## Experiment 1: alibi generation

The research question we addressed in the first experiment was whether differences could be observed between true and false alibis generated by real suspects. We expected that truth-tellers would report to be more often at home than those who lie, similar to the results from Culhane et al. (2013). Second, we expected that liars would less often mention supportive evidence than truth tellers. Concurrently, when liars mention evidence, the evidence would more often be a testimony of a motivated witness than an unmotivated witness. Third, we expected that truth-tellers would report physical evidence more often than liars because it is more difficult for a liar to fabricate fake physical evidence.

### *Method*

#### *Participants*

Fifty male suspects from a Dutch remand prison participated in the study. Their ages varied between 20 and 60 years ($Mdn$ = 34.5). Most of them were single, were unemployed prior to their arrest and had Dutch nationality. The suspects had been recently arrested by the police and were in pre-trial detention awaiting further questioning by the police. Due to privacy regulations, we were only allowed to ask them for their basic demographics.

#### *Research design*

The participants were randomly assigned to a condition in which they were asked to generate either a true or a false alibi for two days prior to their arrest by the police. The time between the testing and their alibi varied between 3 and 14 days ($M = 10.1$, $SD = 2.9$).

#### *Procedure and materials*

The study was conducted in a Dutch remand prison that was visited on 11 occasions in the period between February 2014 and December 2014. The participants had to meet strict criteria in order to participate in the study. They had to be suspects in a criminal investigation awaiting further interviewing by the police, and they were not arrested longer than 14 days before participating. That time period was chosen to ensure that most participants could report an alibi (Culhane et al., 2013; Olson & Charman, 2012) without shifting the study to a memory experiment. The results from alibi generation research show that by increasing the time span between the critical event and the testing results more often in error that requires a change in the alibi and the supportive evidence (e.g. Olson & Charman, 2012). That finding is important but the main focus of the present study is to determine potential differences between true and false alibis. In addition, the only observed effect of an increased delay of up to 12 days resulted in liars reporting more witnesses (Culhane et al., 2013); we therefore did not include a difference in the delay.

Ideally we would visit our participants 48 hours after they generated their initial alibi and to come up with supportive evidence in order to verify their statement, but given the context of being in prison that was impossible. It can therefore be argued that the veracity of the alibis cannot be verified. However, inconsistent alibis were only observed in 6.5% of the generated alibis in the study of Culhane et al. (2013). In our opinion, the limitation of not being able to verify the

evidence does not outweigh the benefits of including suspects in the alibi generation research.

With the help of the Bureau Selection and Detention, in the remand prison, suspects were selected for participation based upon their arrival date in prison. To guarantee the participant's anonymity, we were only given the cell number when visiting the prison for the study. If the suspect wanted to participate, he was brought to a private room. After consent, the participants read a mock crime report, which was a case description of a robbery, in which the police believed the participant could be a suspect and wherein the suspect is asked for his alibi on the day and time of the alleged crime. The participant was subsequently asked, in an open-ended question, to describe in as much detail as possible where he had been. Half of the participants were instructed to lie about the alibi; the other half was asked to be truthful. We used the same instruction as Culhane et al. (2013) provided to the lying participants: they were asked to fabricate an alibi that they would tell to the police in order to convince them of their innocence. The other half of the participants were asked to describe their true alibi of where they had been when the alleged crime was committed.

After the participants wrote down their alibis, they were asked whether they had supportive evidence (both physical and witness evidence). The participants were asked to describe their physical evidence in an open-ended question, similar to the method used by Culhane et al. (2008). If a witness could support the alibi, they were asked in a cued-question what their relation to the witness was. They could choose from six possible witnesses: friend, co-worker, acquaintance, stranger, partner and family member. Afterwards the participants were asked in an open-ended question whether they had any other unreported evidence that could support their alibi. The participants were then thanked, received a written debriefing, were offered the possibility to receive the results of the study when the data were collected and were brought back to their cell.

## *Results*

### *Alibis*

The participants generated 24 false and 25 true alibis. One participant refused to report a false alibi because he would never talk to the police. The generated alibis were post hoc categorized into 11 categories. The participants reported most often to be at home (40.8%), at a friend's house; alone outside (e.g. jogging); or in a bar with friends (all scores: 12.2%). Table 1 provides an overview. No difference among the conditions was observed for which alibi location was reported, although five participants said they were in a bar with friends when telling the truth compared to one participant in the lie condition, $\chi^2(1, N = 49) = 2.86$, $p = .09$, Cramer's $V = 0.24$. In addition, there was no difference between the conditions in relation to the number of words in the suspect's statement ($M_{true} = 74.44$, $SD = 10.93$; $M_{false} = 71.36$, $SD = 16.51$), $F(1, 48) = 0.61$, $p = ns.$ No support was thus found for our first

Table 1. Reported alibis by truth tellers and liars.

| | Truth tellers | | Liars | | Total | |
|---|---|---|---|---|---|---|
| Location | *N* | % | *N* | % | *N* | % |
| At home | 10 | 50.0 | 10 | 50.0 | 20 | 40.8 |
| Shopping | 1 | 100.0 | 0 | 0.0 | 1 | 2.0 |
| In transit | 1 | 100.0 | 0 | 0.0 | 1 | 2.0 |
| With friends | 2 | 33.3 | 4 | 66.7 | 6 | 12.2 |
| Abroad | 1 | 100.0 | 0 | 0.0 | 1 | 2.0 |
| Alone outside | 2 | 33.3 | 4 | 66.7 | 6 | 12.2 |
| With family | 3 | 75.0 | 1 | 25.0 | 4 | 8.2 |
| In a bar with friends | 5 | 83.3 | 1 | 16.7 | 6 | 12.2 |
| With partner | 0 | 0.0 | 2 | 100.0 | 2 | 4.1 |
| At a party | 0 | 0.0 | 1 | 100.0 | 1 | 2.0 |
| At the cinema | 0 | 0.0 | 1 | 100.0 | 1 | 2.0 |
| Total | 25 | 51.0 | 24 | 49.0 | 49 | 100.0 |

hypothesis because both truth-tellers and liars reported equally often being at home.

*Supportive evidence*

Forty-four participants (about 90%) said they had supportive evidence for their alibis, and there was no difference between the conditions, $\chi^2(1, N = 49) = 0.27$, $p = .60$, Cramer's $V = 0.07$. That leads us to reject our second hypothesis that truth-tellers would more often report supportive evidence than would liars. The evidence consisted of only physical evidence (6.8%), only witness evidence (40.9%) or a combination of physical and witness evidence (52.3%). No difference between the conditions was observed if evidence was presented and regardless of the evidence that was presented, although 15 truth-tellers compared to eight liars reported a combination of evidence, $\chi^2(1, N = 44) = 3.24$, $p = .07$, Cramer's $V = 0.27$. Therefore, the individual categories of supportive evidence were examined.

Twenty-four participants (49.0%) reported having physical evidence for their alibis. No difference was observed between the truth-tellers (62.5%) and liars (37.5%) if they reported physical evidence, $\chi^2(1, N = 44) = 2.48$, $p = ns$, Cramer's $V = 0.23$, which leads us to reject our third hypothesis that truth-tellers would more often report physical evidence than would liars. They reported 30 items of physical evidence, and again there was no difference between the conditions and the number of items mentioned, $F(1, 22) = 0.08$, $p = ns$. The items of physical evidence were post hoc clustered into 15 categories; Table 2 provides an overview. No differences were observed between the categories and the conditions, largest $\chi^2(1, N = 24) = 0.83$, $p = .36$, Cramer's $V = 0.19$.

Witness evidence was reported by 41 participants (93.2%), and no difference was found between the conditions whether or not witness evidence was reported, $\chi^2(1, N = 44) = 0.46$, $p = ns$, Cramer's $V = 0.10$. Participants reported a total of 105 witnesses,[1] and again no difference was observed between the conditions and the number of witnesses reported, $F(1, 39) = 0.12$, $p = ns$. A friend was most often mentioned to be able to support the participant's alibi ($N = 44$) while no participant mentioned a co-worker as a witness. An acquaintance and a stranger can be classified as unmotivated witnesses, and a partner

Table 2. Reported physical evidence by truth tellers and liars.

| | Truth tellers | | Liars | | Total | |
|---|---|---|---|---|---|---|
| Item of physical evidence | *N* | % | *N* | % | *N* | % |
| Phone records | 2 | 50.0 | 2 | 50.0 | 4 | 13.3 |
| Bank transaction | 2 | 100.0 | 0 | 0.0 | 2 | 6.7 |
| Receipt | 4 | 50.0 | 4 | 50.0 | 8 | 26.7 |
| CCTV footage | 4 | 80.0 | 1 | 20.0 | 5 | 16.7 |
| Photos | 2 | 100.0 | 0 | 0.0 | 2 | 6.7 |
| Groceries | 1 | 50.0 | 1 | 50.0 | 2 | 6.7 |
| Fresh paint on the wall | 1 | 100.0 | 0 | 0.0 | 1 | 3.3 |
| Computer logs | 1 | 50.0 | 1 | 50.0 | 2 | 6.7 |
| Fresh tattoo | 1 | 100.0 | 0 | 0.0 | 1 | 3.3 |
| Used condom | 0 | 0.0 | 1 | 100.0 | 1 | 3.3 |
| Movie ticket | 0 | 0.0 | 1 | 100.0 | 1 | 3.3 |
| Laundry after playing sports | 0 | 0.0 | 1 | 100.0 | 1 | 3.3 |
| Total | 18 | 60.0 | 12 | 40.0 | 30 | 100.0 |

while a family member and friend can be classified as motivated witnesses. In total, 91 motivated witnesses (86.7%) were mentioned compared to 14 unmotivated witnesses (13.3%). No differences were found for the number of motivated witnesses between the conditions, $F(1, 39) = 0.14$, $p = ns$, which leads us also to reject the hypothesis that liars would more often report a witness with a potential motive to lie than would truth tellers.

Five participants in the truth condition reported additional supportive evidence compared to zero participants in the lie condition, $\chi^2(1, N = 49) = 5.35$, $p = .02$, Cramer's $V = 0.33$. Two participants mentioned physical evidence (unreported telephone records), one participant reported witness evidence (his neighbour), and two participants reported unclear evidence (e.g. dirty hands caused by working on a friend's scooter).

### *Discussion*

The aim of the first experiment was to determine whether differences can be observed between true and false alibis generated by actual suspects. However, no differences were observed between the true and false alibis. Additional evidence was the only factor that differentiated between true and false statements.

Two downsides may have caused the absence of differences between the groups. First, the participants wrote down their statements because we were not allowed to bring any recording devices into the facility. Audio recordings might have resulted in alibis that were richer in details. Second, the participants are actual suspects, and they might be familiar with generating alibis compared to inexperienced students. That may have taken away differences between true and false alibis.

The results show that alibi discrimination based on the type, quantity and quality of the supportive evidence makes discrimination very difficult. To determine whether professionals are better at discriminating between true and false alibis, a second experiment was conducted.

## Experiment 2: Alibi discrimination and evaluation

The proposition that innocent people can generate accurate alibis supported with strong evidence (Olson & Charman, 2012) raises the possibility that those who are not innocent cannot generate accurate alibis with strong supportive evidence. However, as described above, there are various reasons why non-offenders would not report an accurate alibi (Crozier et al., 2017). In addition, alibi generation research shows that the strength of the supportive evidence cannot be used on its own to determine the believability of an alibi because most non-offenders do not have such strong evidence although they are innocent (e.g. Nieuwkamp et al., 2017). It has led scholars more recently to study whether any differences can be distinguished between true and false alibis rather than only focusing on the strength of the evidence (Culhane et al., 2013). In alibi discrimination, evaluators try to differentiate between true and false statements. To understand the research on alibi discrimination, a short and to the point overview is provided of lie detection research.

### *Verbal and non-verbal cues for lie detection*

In general, one can either apply non-verbal techniques for deception detection or use verbal techniques for it. Using non-verbal techniques, more attention is paid to non-verbal behaviour (e.g. gaze, body movement) rather than the content of someone's statement. From lie detection research it is known that when using video-recorded statements, people tend to focus on non-verbal cues (e.g. movements and posture) rather than verbal cues (i.e. the speech content; e.g. Vrij, 2008). A meta-analysis on verbal and non-verbal cues of deception showed that the non-verbal cues of deception are especially weak and

unreliable (DePaulo et al., 2003). More support for the use of verbal cues can be found in the Undeutsch hypothesis (Stellar, 1989), which asserts that true statements can be distinguished from fabricated statements because they consist of different verbal characteristics (Van Koppen & Van Koppen, 2010). The results of research on the topic of discrimination between true and false statements show increased accuracy when evaluators only have access to written or audio statements and when only verbal cues are available compared to when non-verbal cues are also available (Bond & DePaulo, 2006; Burgoon, Blair, & Strom, 2008; DePaulo, Rosenthal, Green, & Rosenkrantz, 1982; Lindholm, 2008). More specifically, when the evaluator focuses on a cluster of verbal cues, the accuracy in the discrimination can increase by up to 80% (Vrij, 2008).

However, not all verbal cues are diagnostic to differentiate between true and false statements. The cues that have been found to be diagnostic are plausibility, level of detail, logical structure, length and the number of contradictions (DePaulo et al., 2003; Vrij, 2005, 2008). Furthermore, Hartwig and Bond (2011) found that the strongest cue for deception detection is the global impression of the statement. They argued therefore that a variety of cues can be better predictors than an evaluation on individual cue level (Hartwig & Bond, 2011).

Regardless of the research on lie detection, it is worth mentioning that during a police interview, interviewers tend to focus more attention on non-verbal cues rather than on verbal cues (Vrij, Granhag, Mann, & Leal, 2011). In order to differentiate better between true and false statements, the use of verbal cues is therefore preferable to the use of non-verbal cues (Vredeveldt, Van Koppen, & Granhag, 2014; Vrij, 2008), which, for example, can be achieved by providing a written statement. It is therefore advisable to focus more on *what* is being said rather than *how* it is being said (e.g. Vredeveldt et al., 2014).

### *Alibi discrimination*

In all studies on alibi discrimination, the participants were undergraduate students. To make a decision whether the alibis presented to them are true or false, students are typically shown a videotaped statement (Culhane et al., 2013; Strömwall et al., 2003) or statements written out on paper (Nahari & Vrij, 2014; Porter & Yuille, 1996). From that type of research it is known that true alibis are found to be more coherent (Porter & Yuille, 1996) and more often supported by evidence and details that can be verified than false alibis (Nahari & Vrij, 2014). False alibis more often consist of statements in which people report to be at home where a witness can support their alibi than do true alibis (Culhane et al., 2013). These researchers also found that untruthful alibis consist of more words than true alibis.

Students in the first experiment found the discrimination between true and false alibis difficult, with a modest accuracy of 66%. True alibis are more often classified correctly (about 71%) than false alibis (about 62%). The non-representative materials and participants in the research can, perhaps, explain the modest results. The present experiment aims at narrowing the gap between laboratory studies and fieldwork. Ten of the alibis generated in Experiment 1 were presented to experienced police detectives to determine whether they are better at discriminating between true and false alibis than students.

The research question that was investigated in the second experiment is whether police detectives can make a better decision on alibi discrimination than students. Second, we were interested in which criteria both groups of participants use to discriminate between the alibis. We expected that the overall accuracy would be around 66% (based on the results of previous research). Because the introduction of police detectives into the research on alibi discrimination is new, we had no concrete expectation on their performance but we expected that their

performance would be better than that of students. Second, we expected that true alibis would be evaluated as more believable, contain more details, have a more logical structure, be shorter and contain fewer contradictions than false alibis, based on previous findings in lie detection (DePaulo et al., 2003; Vrij, 2005, 2008). In addition, we know that both police detectives and students expect strong evidence for a true and believable alibi (Dysart & Strange, 2012; Jung et al., 2013); we therefore expected that for both groups the presence of strong (i.e. physical) evidence would be important when they discriminate between alibis. However, given their expertise, it might be that police detectives focus on additional aspects of alibis in the process of alibi generation. We explore the factors influencing the decision making about the alibis in both groups.

## Method

### Participants

A total of 46 police detectives (26 males, 20 females)[2] and 47 undergraduate students (11 males, 36 females) were recruited for the study. The detectives' ages varied between 24 and 63 years ($Mdn = 45.0$). On average, they had worked for more than 20 years for the police and had about 14 years of interview experience. Only five police detectives (11.4%)[3] had received some training in deception detection or on how to evaluate alibis. Most of the police detectives evaluated an alibi less than once a month in criminal cases they encountered as part of their work. The students' ages varied between 19 and 27 years ($Mdn = 21.0$). Two gift cards of 10 euros were raffled among each group of participants.

### Research design

The design of the study consisted of a 2 (police detective, student) × 2 (true, false alibi) mixed factorial design.

### Materials

From the 24 false alibis generated in prison, five alibis were randomly selected (using www.randomnumbergenerator.com), and five true alibis were selected from the 25 true alibis using the same procedure. The 10 selected alibis were written out in the format of an official police report (see attachment). The statements were written out because the results of lie detection research show that verbal cues of a statement are more effective for differentiating between true and false statements than non-verbal cues (Vredeveldt et al., 2014; Vrij, 2008). The participants were subsequently provided with a random set of five alibis out of the 10 alibis selected for the study. The presented set of five alibis could thus consist of five true alibis, five false alibis or five alibis of which some were true and some were false.

### Pilot study

Because the statements generated in Experiment 1 were quite short, and the number of details was quite low, we wanted to know whether the a priori selected statements would be suitable for the second experiment. To assess whether the generated alibis in prison were suitable for the present experiment, a researcher (different from the authors) asked 10 laypersons to provide either a true or a false alibi following the same instructions as those given to the participants in Experiment 1. When the new statements were collected, the participants were asked follow-up questions (e.g. what did you do prior to 8.30 pm?) and questions to generate more details (e.g. could you describe what you have watched on the television?) after their free recall. These statements were verbatim transcribed in question–answer format. The length of the laypeople's statements consisted of 496 words on average ($M_{true} = 441.6$, $M_{false} = 551.0$), and the statements from Experiment 1 consisted of 73 words on average ($M_{true} = 74.44$, $M_{false} = 71.36$). The 10 alibis selected a priori

for the second experiment and the 10 alibis generated by laypeople were provided to 10 scholars from the KU Leuven University in Belgium. They were asked to classify the statements into two categories: true and false. No difference was observed in their accuracy between the groups despite the longer and more detailed statements from laypeople. We therefore kept the alibis generated by the suspects to increase the ecological validity of the materials.

### *Procedure*

The student participants could register for participating in the experiment by applying via the SONA system or by contacting the researcher via contact information on posters throughout the faculty. The police detectives were contacted through personal contacts. The participants completed the questionnaire individually. A researcher was always present in case clarification was needed. In a minority of cases, a detective completed the questionnaire at a later moment after she or he was called away for an emergency.

The participants had to fill in seven separate forms. After completing the *informed consent* and *demographics* form they read a *crime report* containing the mock crime scenario identical to the one used in Experiment 1. In that report it was also stated that multiple suspects were arrested and that all of them presented an alibi. Because it was unclear whether the arrested suspects were involved in that crime, it could be the case that all alibis were true, that all alibis were false or that some alibis were true and some false, similar to the method used by Culhane et al. (2013). The participants were then presented with five alibis. After reading each alibi, the participants received an *evaluation form*.

On the evaluation form, the participants were first asked in a forced-choice question whether the alibi was true or false (i.e. alibi discrimination). The participants then evaluated the alibis (i.e. alibi evaluation) on criteria that have been proven to be effective in differentiating between true and false statements from the lie detection literature (e.g. the number of contradictions) and alibi evaluation (e.g. alibi believability). A variety of cues was presented because such variety can be a better predictor than single cue indicators (Hartwig & Bond, 2011; Vrij, 2008). In addition, the participants were asked to evaluate all the reported supportive evidence on three aspects per type of reported evidence: the difficulty of fabrication, the strength and the believability. The participants were also asked to score the verifiability of all presented evidence. The number of questions on the evaluation forms thus varied between seven and 17 questions depending on whether the alibi was supported with evidence and what type of evidence was reported. When an alibi was presented without supportive evidence, the participant answered seven questions. When the three types of supportive evidence were reported, the participants answered 17 questions.

Besides the forced-choice question at the beginning of the evaluation form, the participants were asked to mark their answers on a 100-mm visual analogue scale (VAS: 0 = not at all believable; 100 = completely believable; Luria, 1975).

Last, the participants were asked to divide 100 points over the 12 aspects of the evaluation form[4] indicating which of the aspects are most important for them to determine the veracity of each alibi. Using that method, we hoped to receive more information on which aspects are important to the participants in alibi discrimination. After the participants completed all forms, they received a written debriefing, were offered the possibility to receive the results of the experiment and were able to write down their email address to participate in the raffle of gift cards.

## ***Results***

### *Accuracy*

First, the proportion of the accuracy was calculated. The overall accuracy was moderate

(*Mdn* = .60); however, for the police officers, their accuracy scores for the true alibis were non-normally distributed, with skewness of –1.61 (*SE* = 0.43) and kurtosis of 2.23 (*SE* = 0.85). The police data for the false alibis were also non-normally distributed with skewness of 0.11 (*SE* = 0.12) and kurtosis of –1.67 (*SE* = 0.92). Therefore separate analyses were conducted to determine the accuracy of the true and false alibis. A Mann–Whitney test indicated that the detection of true alibis was higher for police detectives (*Mdn* = 1.00) than for students (*Mdn* = .67), *U* = 506.0, *p* = .020, *r* = .24. No such difference was observed for the false alibis, *U* = 983.0, *p* = .073, *r* = .20. Examining the proportions, a substantial number of all participants have either all correct or all non-correct. For example, when only one true alibi was presented, and the participant classified the alibi correctly, it can be concluded that all true alibis were classified correctly. To control for that, we selected the participants (29 police officers and 31 students) who at least evaluated either two true or two false alibis. In that case, police officers are still better at classifying true alibis (*Mdn* = 1.00) over students (*Mdn* = .67), *U* = 303.5, *p* = .025, *r* = .29. With regard to false alibis that were classified correctly, students performed better than police, although the median of the proportion of false alibis classified correctly was identical for both groups (*Mdn* = .50); officers: *U* = 621.5, *p* = .023, *r* = .29. When we examined the cases in which the participants were presented with at least two true and true false alibis, all significant differences disappeared, due to a very small sample of participants (15 police officers and 14 students). We can therefore conclude that police detectives are better at detecting true alibis than students, while students are better at detecting false alibis when at least two false alibis are presented. The performance of police officers was not influenced by whether or not they were more or less experienced within the police force or in conducting interviews, highest *U* = 257, *p* = .210, *r* = .20. The results leads us to carefully accept our hypothesis that the accuracy would be around 66% although important differences between police detectives and students are shown.

### *Alibi evaluation*

All the data on the alibi evaluation aspects were normally distributed, apart from the clearness of the alibi. The police officers' scores on that variable were non-normally distributed, with skewness of 0.75 (*SE* = 0.16) and kurtosis of –1.22 (*SE* = 0.32). The students' scores were also non-normally distributed, with skewness of –0.08 (*SE* = 0.16) and kurtosis of –1.03 (*SE* = 0.16). The results show that while no difference is observed between the true and the false alibis, a difference is observed between the groups of participants. For both true and false alibis, students found the alibis more clear (*Mdn* = 5.2) than the police officers (*Mdn* = 4.7), *U* = 31,174.0, *p* = .001, *r* = .16.

The other variables were analysed using multivariate analyses of variance. An alpha level of .05 was used in each of the analyses. All significant effects were examined applying a Bonferroni adjustment for multiple testing. The results revealed that police detectives came to their decision more rationally (*M* = 5.50, *SD* = 2.80) than students (*M* = 4.76, *SD* = 2.46), *F*(1, 455) = 9.17, *p* = .003, $\eta_p^2$ = .020, and police detectives found the alibis to contain fewer contradictions (*M* = 2.49, *SD* = 2.17) than did students (*M* = 2.99, *SD* = 2.38), *F*(1, 455) = 5.01, *p* = .011, $\eta_p^2$ = .011. Furthermore, true alibis were also found to be more difficult to fabricate (*M* = 3.61, *SD* = 2.77), believable (*M* = 5.48, *SD* = 2.30) and detailed (*M* = 3.81, *SD* = 2.74) and to contain fewer contradictions (*M* = 2.99, *SD* = 2.38) than false alibis.[5]

Although the true alibis were found to be more believable than false alibis, the mean scores are relatively low (5.48 on a 10-point scale). These low scores could possibly be explained by the modest overall accuracy (60%). Therefore, we examined whether the

average scores on these aspects were more related to the participants' evaluation (true or false) than the ground truth of the alibis, using generalized estimating equation (GEE) analyses (Hanley, Negassa, Edwardess, & Forrester, 2003).

*GEE analysis*

In the present experiment the participants rated the same aspects of all alibis. The five measurements of believability, for example, are dependent within-subjects factors and produce correlated data. To control for such a dependency, separate GEE analyses were conducted for each aspect per group of participants. In the analyses it was determined whether the score on each aspect was more related to the ground truth of the alibi (hereafter: ground truth) or more to the evaluation of the alibi (hereafter: evaluation). For example, for the students' score on the aspect 'the number of details', the intercept consists of 1.77 points, which is the expected score when a false alibi is evaluated as false. The beta score for ground truth is 0.81 points, thus when the alibi is true the expected score for this aspect increases with 0.81 points, and that increase is significant. When the alibi is evaluated as true, the score increases with 2.22 points, which is again a significant increase. Thus, when the alibi is true and is evaluated as true, the expected score increases significantly by 0.81 + 2.22 points = 4.80 points; see Table 3 for the test results.

For both groups of participants the results show that the scores on all dependent measures were significantly related to the *evaluation* of the alibi, apart from the police officers' score on the rationality of their decision. Besides the aspect 'contradiction', all aspects obtained a positive beta score indicating a higher score on the dependent measures when the alibi was evaluated as true than when it was evaluated as false (i.e. the value of the intercept). The aspect contradiction received a negative beta score for the evaluation. That is, when the alibi is evaluated as true, the number of contradictions decreases. For both groups, the results show that the scores on the number of contradictions were significantly related to the *ground truth* of the alibi. In the student data the ground truth of the number of details was also significantly related to their scores. All three aspects had a positive beta score indicating that when the alibi is true, the expected score on the dependent measures would increase. In sum, the scores on the dependent measures are more often related to the evaluation of the alibi than to the ground truth of the alibi. It can be concluded that when the alibi is evaluated as true, higher scores are observed for the difficulty of fabrication, its believability, the number of details and the clearness of the alibi, while the number of contradictions decreases. Based on the results, our second hypothesis is partly rejected. True alibis are found to be more difficult to fabricate, believable and detailed and to contain fewer contradictions, but no difference in the number of words has been observed.

*Supportive evidence*

Similar to the analyses conducted for the alibi evaluation, the analyses of the strength, ease of fabrication and believability of the supportive evidence resulted in counter-intuitive results. For example, the physical evidence for a false alibi is more believable than that for a true alibi. As stated above, given the modest accuracy in detecting the veracity of statements, it should be determined whether the scores on the dependent measures of the supportive evidence relate more to the decision of whether the alibi is true than its ground truth. Therefore linear regression analyses were conducted, and the results are displayed in Table 4.

Based on the participants' *decision*, it can be concluded that both groups of participants gave significant higher scores for alibis they evaluated to be true than for alibis they classified as false on all evidence aspects (i.e. strength, believability and difficulty of

Table 3. GEE analysis with the predicted scores on the alibi aspects.

| | | Police detectives | | | | | Students | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Alibi aspect* | | *B* | *SE* | Wald $\chi^2$ | *df* | *p* | *B* | *SE* | Wald $\chi^2$ | *df* | *p* |
| Difficulty of fabrication | Intercept[a] | 1.56 | 0.30 | 27.16 | 1 | <.001 | 1.86 | 0.22 | 74.84 | 1 | <.001 |
| | Ground truth | 0.50 | 0.32 | 2.40 | 1 | *ns* | 0.47 | 0.36 | 1.77 | 1 | *ns* |
| | Evaluation | 1.89 | 0.37 | 25.56 | 1 | <.001 | 1.88 | 0.33 | 32.33 | 1 | <.001 |
| Clear alibi | Intercept[a] | 2.65 | 0.52 | 26.12 | 1 | <.001 | 3.97 | 0.40 | 96.06 | 1 | <.001 |
| | Ground truth | 0.31 | 0.27 | 1.29 | 1 | *ns* | 0.04 | 0.30 | 0.02 | 1 | *ns* |
| | Evaluation | 2.04 | 0.46 | 20.03 | 1 | <.001 | 2.02 | 0.34 | 36.23 | 1 | <.001 |
| Believability | Intercept[a] | 3.02 | 0.29 | 109.81 | 1 | <.001 | 2.87 | 0.18 | 249.72 | 1 | <.001 |
| | Ground truth | 0.23 | 0.20 | 1.29 | 1 | *ns* | 0.41 | 0.22 | 3.47 | 1 | *ns* |
| | Evaluation | 2.45 | 0.33 | 53.64 | 1 | <.001 | 3.58 | 0.22 | 257.63 | 1 | <.001 |
| Detailed | Intercept[a] | 1.49 | 0.28 | 28.48 | 1 | <.001 | 1.77 | 0.21 | 73.57 | 1 | <.001 |
| | Ground truth | 0.58 | 0.33 | 3.17 | 1 | *ns* | 0.81 | 0.27 | 8.64 | 1 | .003 |
| | Evaluation | 1.80 | 0.32 | 32.76 | 1 | <.001 | 2.22 | 0.29 | 58.76 | 1 | <.001 |
| Contradiction | Intercept[a] | 2.93 | 0.33 | 82.11 | 1 | <.001 | 3.28 | 0.24 | 182.02 | 1 | <.001 |
| | Ground truth | 1.20 | 0.28 | 17.99 | 1 | <.001 | 1.11 | 0.25 | 20.43 | 1 | <.001 |
| | Evaluation | −1.22 | 33 | 13.46 | 1 | <.001 | −1.46 | 0.27 | 29.34 | 1 | <.001 |

*Note:* GEE = generalized estimating equation; *B* = beta score; *SE* = standard error; *df* = degrees of freedom.
[a]Intercept denotes the predicted score of the alibi aspect, given *a* participants who evaluate a false alibi as false.

fabrication), besides the strength of additional evidence by police officers. These scores are positive and imply that when an alibi is evaluated as true, the score increases compared to when the alibi is evaluated as false. For the *ground truth* of the alibis, most of the participants' scores were significantly lower than the constant in the model. In such cases the values were negative, which implies that if the alibi is true, lower scores on the aspects are observed.

### *Which aspects are important for alibi discrimination?*

After the participants evaluated the alibis, they were asked for each alibi to divide 100 points over 12 aspects to determine what aspects in the alibi were found to be important to decide whether the alibi was true or false. Out of the 465 evaluated alibis, the points of 409 alibis could be used. Other cases had to be removed either due to missing values ($N = 14$) or due to participants who allocated more than 100 scores to the 12 aspects ($N = 42$). The data were non-normally distributed, causing the average or median score not to be usable in the analyses due to higher number of cases in which 0 points were attributed to the aspects. We therefore created four levels of scores per aspect points to obtain a normal distribution of scores: 0; scores between 1 and 5; scores between 5 and 10; and scores more than 10. These ordinal categories of scores were compared between the groups. The results show that students more often attribute more than 10 points on whether or not physical evidence was reported ($N = 34$) than do police officers ($N = 11$), $\chi^2(2, N = 151) = 11.90$, $p = .003$, $\tau_b = .227$, $p = .003$. Police officers more often assigned 10 or more points on the length of the alibi ($N = 15$) than students ($N = 4$), $\chi^2(2, N = 108) = 11.79$, $p = .003$, $\tau_b = –.296$, $p = .002$. In addition, police officers more often assigned 10 or more points to the verifiability of the evidence ($N = 67$) than students ($N = 37$), $\chi^2(2, N = 250) = 33.17$,

Table 4. Evaluation of the supportive evidence by police detectives and students.

| | | Police detectives | | | | | | Students | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Evidence | | *B* | *SE* | *β* | *T* | *p* | $R^2$ | *B* | *SE* | *β* | *T* | *p* | $R^2$ |
| *Physical* | | | | | | | | | | | | | |
| Strength | Constant[a] | 2.68 | 0.93 | | 2.88 | .005 | .17 | 4.49 | 0.84 | | 5.82 | <.001 | .17 |
| | Decision | 2.90 | 0.81 | 0.37 | 3.60 | .001 | | 1.71 | 0.66 | 0.26 | 2.60 | .011 | |
| | Ground truth | −1.11 | 0.62 | −0.18 | −1.79 | *ns* | | −1.74 | 0.64 | −0.27 | −2.73 | .008 | |
| Believability | Constant[a] | 2.24 | 0.90 | | 2.49 | .015 | .20 | 5.19 | 0.71 | | 7.35 | <.001 | .22 |
| | Decision | 3.55 | 0.78 | 0.43 | 4.58 | <.001 | | 1.98 | 0.55 | 0.35 | 3.59 | .001 | |
| | Ground truth | −0.25 | 0.61 | −0.04 | −0.41 | *ns* | | −1.40 | 0.54 | −0.25 | −2.62 | .010 | |
| Difficulty of fabrication | Constant[a] | 3.56 | 1.00 | | 3.54 | .001 | .12 | 3.24 | 0.85 | | 3.81 | <.001 | .09 |
| | Decision | 2.16 | 8.7 | 2.6 | 2.48 | .015 | | 2.18 | 0.67 | 0.34 | 3.81 | <.001 | |
| | Ground truth | −1.44 | 0.67 | −0.23 | −2.15 | .034 | | 0.36 | 0.65 | 0.06 | 0.56 | *ns* | |
| *Witness* | | | | | | | | | | | | | |
| Strength | Constant[a] | 2.53 | 0.42 | | 5.96 | <.001 | .20 | 3.41 | 0.35 | | 9.62 | <.001 | .15 |
| | Decision | 2.85 | 0.43 | 0.46 | 6.64 | <.001 | | 2.11 | 0.36 | 0.40 | 5.95 | <.001 | |
| | Ground truth | −0.03 | 0.37 | −0.01 | −0.09 | *ns* | | −0.21 | 0.35 | −0.04 | −0.60 | *ns* | |
| Believability | Constant[a] | 3.00 | 0.38 | | 7.86 | <.001 | .20 | 3.28 | 0.31 | | 10.59 | <.001 | .19 |
| | Decision | 2.57 | 0.39 | 0.46 | 6.67 | <.001 | | 2.06 | 0.31 | 0.44 | 6.61 | <.001 | |
| | Ground truth | −0.38 | 0.33 | −0.78 | −1.14 | *ns* | | 0.40 | 0.41 | 0.09 | 1.30 | *ns* | |
| Difficulty of fabrication | Constant[a] | 2.22 | 0.44 | | 5.01 | <.001 | .11 | 2.71 | 0.38 | | 7.16 | <.001 | .06 |
| | Decision | 2.18 | 0.45 | 0.35 | 4.85 | <.001 | | 1.28 | 0.38 | 0.24 | 3.39 | .001 | |
| | Ground truth | −0.08 | 0.39 | −0.01 | −0.20 | *ns* | | 0.56 | 0.37 | 0.11 | 1.49 | *ns* | |
| *Additional* | | | | | | | | | | | | | |
| Strength | Constant[a] | 4.39 | 0.89 | | 4.91 | <.001 | .23 | 4.48 | 0.81 | | 5.55 | <.001 | .37 |
| | Decision | 1.48 | 0.79 | 0.21 | 1.87 | *ns* | | 2.96 | 0.70 | 0.42 | 4.25 | <.001 | |
| | Ground truth | −2.15 | 0.61 | −0.41 | −3.55 | .001 | | −1.82 | 0.52 | −0.35 | −3.51 | .001 | |
| Believability | Constant[a] | 3.87 | 0.87 | | 4.43 | <.001 | .21 | 5.18 | 0.86 | | 6.03 | <.001 | .24 |
| | Decision | 2.02 | 0.77 | 0.30 | 2.62 | .012 | | 2.13 | 0.74 | 0.32 | 2.87 | .005 | |
| | Ground truth | −1.60 | 0.59 | −0.31 | −2.70 | .009 | | −1.62 | 0.55 | −0.32 | −2.93 | .005 | |

(*continued*)

Table 4. (*Continued*)

| Evidence | | Police detectives | | | | | | Students | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *B* | *SE* | *β* | *T* | *p* | $R^2$ | *B* | *SE* | *β* | *T* | *p* | $R^2$ |
| Difficulty of fabrication | Constant[a] | 2.43 | 1.01 | | 2.41 | .019 | .10 | 3.92 | 0.94 | | 4.19 | <.001 | .29 |
| | Decision | 2.07 | 0.89 | 0.29 | 2.33 | .023 | | 2.69 | 0.81 | 0.35 | 3.34 | .001 | |
| | Ground truth | −0.86 | 0.68 | −0.16 | −1.27 | *ns* | | −1.90 | 0.60 | 0.34 | −3.16 | .002 | |
| Verifiability all evidence | Constant[a] | 5.20 | 0.43 | | 12.22 | <.001 | .08 | 4.58 | 3.37 | | 12.28 | <.001 | .01 |
| | Decision | −1.44 | 0.43 | −0.25 | −3.34 | .001 | | −0.78 | 0.37 | −0.15 | −2.09 | .038 | |
| | Ground truth | −0.84 | 0.37 | −0.17 | −2.25 | .026 | | −0.01 | 0.37 | −0.00 | −0.01 | *ns* | |

*Note:* $B$ = beta score; $SE$ = standard error; $\beta$ = standardized beta; $R^2$ = determination coefficient (percentage of dependent variable that is explained by the linear model with the independent variables (ground truth and evaluation).
[a]Constant denotes the predicted score of the aspect, given *a* participants who evaluate an actual false alibi (ground truth = false), and the participant correctly classifies the alibi as false (evaluation = false).

$p < .001$, $\tau_b = -.302$, $p < .001$. Comparing the column proportions with a $z$-test applied with a Bonferroni correction for multiple comparisons showed that all proportions were also significant different from each other ($p < .05$). It can therefore be concluded that when discriminating between alibis, students find the presence of physical evidence most important, while police officers more often focus on the length of the statement and in particular on how easy the alibi can be verified. The data lead us to reject the third hypothesis because only students found the presence of strong supportive evidence one of the most important aspects in the alibi discrimination process.

### *Discussion*

By providing written statements of real suspects to professionals, we were able to increase the accuracy rates to 60%. Although this percentage is slightly higher than the percentage reported by Culhane et al. (2013; 57.4%), the accuracy score in the present experiment is comparable to previous findings (Culhane et al., 2013; Granhag, Strömwall, & Jonsson, 2003; Nahari & Vrij, 2014; Porter & Yuille, 1996). We found that police officers are better at detecting true alibis and that students are better at detecting false alibis – that is, when their sample consisted of minimal two false statements. As expected, true alibis are found to be more difficult to fabricate, more believable and more detailed and to contain fewer contradiction than false alibis. The data show that police officers come to their decision more rationally than students when evaluating an alibi while students find the alibis to contain more inconsistencies than police officers. Most importantly, however, is that professionals and students focus on different aspects of the alibi when deciding whether the alibi is true or false. Whereas students focus on the length of the statement and the presence of physical evidence, police detectives pay more attention to whether or not the alibi can be verified

in general. The present experiment adds to the discussion that more research should be conducted with representative samples of participants (e.g. Eastwood et al., 2016), especially because students differently discriminate between statements compared to professionals.

One of the most important downsides of the experiment is the material used. The police officers expressed that they would have liked an opportunity to see the suspect when he told his alibi or to be provided with the possibility to ask follow-up questions to each suspect. That is a common limitation when case vignettes are used; however, it guarantees that all participants were provided with the same information because the follow-up question would vary a lot depending on the interest of the participant.

## General discussion

The goal of the present experiments was to apply the general methodology of Culhane et al. (2013) on generating and discriminating between true and false alibis with different groups of participants. Different groups of participants were included to increase the ecological validity of the research. Our results are nevertheless comparable with the findings reported by Culhane et al. (2013), although the overall accuracy in these alterations is higher. One of the most interesting findings in our results is the knowledge that police detectives focus on other aspects during the alibi discrimination process compared to students. That is an important finding for future research and adds to the argument that the results of alibi research using students as participants cannot be translated on a one-to-one basis in practice (e.g. Culhane et al., 2008; Nieuwkamp et al., 2017).

In general it can be concluded that differentiation between true and false alibis is a complex matter. Examining differences between true and false statements (e.g. the alibi, the presence of supportive evidence) did not result in any differences in Experiment 1; besides, the presence of additional evidence is more often reported by truth-tellers than by liars. In addition, the results of Experiment 2 show that the discrimination between true and false alibis does not appear to be easy, with a modest accuracy of 60%. Our results therefore add to the modest amount of research on alibi discrimination that establishes that discrimination is difficult (Culhane et al., 2013; Granhag et al., 2003; Nahari & Vrij, 2014; Porter & Yuille, 1996). We expected the professionals to outperform the students based on their rich experience in investigating crimes, but this was not the case. Although the accuracy scores in the present study are slightly better than those of Culhane et al. (2013), they do not approach the accuracy rates in the verbal lie detection domain where up to 80% accuracy scores are reported (Vrij, 2008). However, an alibi is a different type of statement compared to those normally used in (verbal) lie detection. Alibis often go hand-in-hand with reported supportive evidence and the presence and strength of such evidence. Both strongly influence the alibi believability (e.g. Jung et al., 2013). Based on the relatively weak evidence that non-offenders can present in general (e.g. Nieuwkamp et al., 2017; Olson & Charman, 2012), and likewise observed in Experiment 1, the weak evidence might have affected the overall impression of the alibi. Knowing that global impression of a statement is the strongest predictor of its veracity (Hartwig & Bond, 2011), the alibis might have been negatively affected by that weak evidence.

Although the accuracy rates in the present study do not differ much from previous findings from alibi discrimination research, it is interesting to know that police officers are better at detecting true alibis than are students. Although it has been argued that police officers display a guilty bias towards suspects (e.g. Meissner & Kassin, 2002), it could be the case that the opposite happens with alibis. That is, police officers believe the suspect's alibi. During the debriefing, the police officers also expressed that, although they had limited information, until the alibi can be

verified they tended to believe the suspect bearing in mind the presumption of innocence.

The students were better at detecting false alibis than police officers. That result is unexpected because previous studies found out that students are better at detecting true alibis than false alibis (Culhane et al., 2013). The current result might be explained by two factors. First, normally only students are included in the research on alibi discrimination. In the present study we compared two different groups of participants to each other. The variation in the comparison group might explain the unexpected results. Second, the participants were provided with alibis generated by real suspects rather than fellow students. One could argue that the social life and daily activities of students differ from those of potential criminals, and therefore differences occur in alibi supportive evidence. Future research is needed to answer that question.

Most of our knowledge on alibis originates from research on undergraduate students. Reporting an alibi without strong supportive evidence for it is one of the most stable findings in all alibi research (e.g. Jung et al., 2013; Pozzulo et al., 2015). Such strong evidence most often consists of physical evidence. In fact, even the weakest type of physical evidence is still stronger evidence than the strongest type of witness evidence (Olson & Wells, 2004). Not only do students find the presence and the strength of physical evidence a diagnostic factor in alibi evaluation, our results also show that students focus on the presence of physical evidence in discriminating between true and false alibis. If professionals would focus on the same aspects as students, then we would have gained important knowledge over the past years of research in the field of alibis. Unfortunately that is not the case because professionals do not focus most on whether or not physical evidence is reported. Rather, they are interested in the verifiability of the alibi. Research on the verifiability approach is growing in various kinds of research domains – for example, whether an insurance claim is truthful (Harvey, Vrij, Nahari, & Ludwig, 2017; Nahari, Leal, Vrij, Warmelink, & Vernham, 2014; Vrij, Nahari, Isitt, & Leal, 2016), whether one is honest about his or her occupation (Jupe, Vrij, Nahari, & Leal, 2016), in lie detection in general (Nahari & Vrij, 2015; Nahari, Vrij, & Fisher, 2014a, 2014b) but also in alibi discrimination (Nahari & Vrij, 2014). The approach is based on two assumptions: (a) truth-tellers will include more details in their statement than liars, and (b) liars will avoid including details in their statement and will give vague statements instead (Nahari et al., 2014a). For example, a liar is more likely to state 'I saw a black Audi' rather than 'I called my friend Kris at 10:30 am this morning'. The details of the latter statement can be more easily verified than those of the first statement. Using the verifiability technique, an accuracy of up to 88% in lie detection can be achieved. Our data show that the verifiability of the alibis is most important for professionals. Future research in the domain of the verifiability approach on alibi discrimination is therefore strongly recommended.

This article offers new insights in the domain of alibi generation and discrimination using more representative samples of participants. Although the results are comparable with those of previous studies, this paper shows that police detectives perceive an alibi as true until proven wrong and offers a new scope on the guilty bias of police officers. Furthermore, the data show that the research on alibis conducted with students cannot be translated on a one-to-one basis to professionals. That is an important finding for future alibi research.

**Acknowledgements**

We would like to thank Eline Klein Hofmeijer and Laura Dilling for their help in collecting the data.

**Funding**

This research was financially supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO) [grant number 404-10-349].

## Declaration of conflicts of interest

Ricardo Nieuwkamp has declared no conflicts of interest

Robert Horselenberg has declared no conflicts of interest

Peter van Koppen has declared no conflicts of interest

## Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

## Informed consent

Informed consent was obtained from all individual participants included in the study.

## Notes

1. The participants could choose between six categories to indicate the number of witnesses: none, one, two, three, four, five or more. In cases where the latter option was ticked, we used the lower bound (i.e. five) to calculate the number of witnesses.
2. Four participants did not report their gender.
3. Two missing values; only valid percentage is reported.
4. For the alibi: the ease by which it can be fabricated, the clearness, the believability, the number of details, the length and the number of contradictions. For the evidence: the presence/absence of physical, witness and additional evidence, the ease by which the evidence can be fabricated, the believability and the verifiability.
5. Respectively: ($M = 4.98$, $SD$=2.32), $F(1, 455) = 6.39$, $p = .012$, $\eta_p^2 = .014$; ($M = 2.76$, $SD = 2.66$), $F(1, 454) = 11.20$, $p = .001$, $\eta_p^2 = .024$; ($M = 4.65$, $SD = 2.34$), $F(1, 455) = 13.85$, $p < .001$, $\eta_p^2 = .030$; ($M = 2.69$, $SD = 2.59$), $F(1, 455) = 19.73$, $p < .001$, $\eta_p^2 = .042$; and ($M = 2.49$, $SD = 2.17$), $F(1, 455) = 16.48$, $p < .001$, $\eta_p^2 = .035$.

## References

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality & social psychology Review, 10*, 214–234. doi:10.1207/s15327957pspr1003_2

Burgoon, J. K., Blair, J. P., & Strom, R. E. (2008). Cognitive biases and nonverbal cue availability in detecting deception. *Human Communication Research, 34*, 572–599. doi:10.1111/j.1468-2958.2008.00333.x

Burke, T. M., Turtle, J. W., & Olson, E. A. (2007). Alibis in criminal investigations and trials. In M. P. Toglia, J. D. Read, D. F. Ross, & R. C. Lindsay (Eds.), *The handbook of eyewitness psychology: Memory for events* (Vol. *1*, pp. 157–192). Mahwah, NJ: Lawrence Erlbaum Associates.

Crozier, W. E., Strange, D., & Loftus, E. F. (2017). Memory Errors in Alibi Generation: How an alibi can turn against us. *Behavioral Sciences and the Law, 35*(1), 6–17.

Culhane, S., & Hosch, H. (2004). An alibi witness' influence on mock jurors' verdicts. *Journal of Applied Social Psychology, 34*, 1604–1616.

Culhane, S. E., Hosch, H. M., & Kehn, A. (2008). Alibi generation: Data from U.S. Hispanics and U.S. non-Hispanic whites. *Journal of Ethnicity in Criminal Justice, 6*, 177–199. doi:10.1080/15377930802243395

Culhane, S. E., Kehn, A., Horgan, A. J., Meissner, C. A., Hosch, H. M., & Wodahl, E. J. (2013). Generation and detection of true and false alibi statements. *Psychiatry, Psychology and Law, 20*, 619–638. doi:10.1080/13218719.2012.729018

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

DePaulo, B. M., Rosenthal, R., Green, C. R., & Rosenkrantz, J. (1982). Diagnosing deceptive and mixed messages from verbal and nonverbal cues. *Journal of Experimental Social Psychology, 18*, 433–446. doi:10.1016/0022-1031(82)90064-6

Dysart, J. E., & Strange, D. (2012). Beliefs about alibis and alibi investigations: A survey of law enforcement. *Psychology, Crime and Law, 18*, 11–25. doi:10.1080/1068316X.2011.562867

Eastwood, J., Snook, B., & Au, D. (2016). Safety in numbers: A policy-capturing study of the alibi assessment process. *Applied Cognitive Psychology, 30*, 260–269.

Granhag, P. A., Strömwall, L. A., & Jonsson, A. C. (2003). Partners in crime: How liars in collusion betray themselves. *Journal of Applied Social Psychology, 33*, 848–868.

Hanley, J. A., Negassa, A., Edwardess, M. D. B., & Forrester, J. E. (2003). Statistical analysis of

correlated data using generalized estimating equattions: An orientation. *American Journal of Epidemiology, 157*, 364–375. doi:10.1093/aje/kwf215

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgement. *Psychological Bulletin, 137*, 643–659. doi:10.1037/a0023589

Harvey, A. C., Vrij, A., Nahari, G., & Ludwig, K. (2017). Applying the Verifiability Approach to insurance claims settings: Exploring the effect of the information protocol. *Legal and Criminological Psychology, 22*, 47–59. doi:10.1111/lcrp.12092

Jung, S., Allison, M., & Bohn, L. (2013). Legal decision-making on crimes involving an alibi. *Applied Psychology in Criminal Justice, 9*, 45–58. doi:10.1037/t04698-000

Jupe, L., Vrij, A., Nahari, G., & Leal, S. (2016). The lies we live: Using the Verifiability Approach to detect lying about occupation. *Journal of Articles in Support of the Null Hypothesis, 13*, 1–14.

Lindholm, T. (2008). Who can judge the accuracy of eyewitness statements? A comparison of professionals and lay−persons. *Applied Cognitive Psychology, 22*, 1301–1314. doi:10.1002/acp.1439

Luria, R. E. (1975). The validity and reliability of the visual analogue mood scale. *Journal of Psychiatric Research*, *12*, 51–57.

Meissner, C. A., & Kassin, S. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26*, 469–480. doi:10.1023/A:1020278620751

Nahari, G., Leal, S., Vrij, A., Warmelink, L., & Vernham, Z. (2014). Did somebody see it? Applying the Verifiability Approach to insurance claim interviews. *Journal of Investigative Psychology and Offender Profiling, 11*, 237–243. doi:10.1002/jip.1417

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. doi:10.1016/j.jarmac.2014.04.005

Nahari, G., & Vrij, A. (2015). Can someone fabricate verifiable details when planning in advance? It all depends on the crime scenario. *Psychology, Crime & Law, 21*, 987–999. doi:10.1080/1068316X.2015.1077248

Nahari, G., Vrij, A., & Fisher, R. P. (2014a). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology, 19*, 227–239. doi: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. *Applied Cognitive Psychology, 28*, 122–128. doi:10.1002/acp.2974

Nieuwkamp, R., Horselenberg, R., & Van Koppen, P. J. (2016). A lie and a mistress: On increasing the believability of your alibi. *Psychiatry, Psychology and Law, 23*, 733–745. doi:10.1080/13218719.2016.1142934

Nieuwkamp, R., Horselenberg, R., & Van Koppen, P. J. (2017). The illusion of the perfect alibi: Establishing the base rate of non-offenders' alibis. *Journal of Investigative Psychology and Offender Profiling, 14*, 23–42. doi:10.1002/jip.1470

Olson, E. A., & Charman, S. D. (2012). 'But can you prove it?' - Examining the quality of innocent suspects' alibis. *Psychology, Crime and Law, 18*, 453–741. doi:10.1080/1068316x.2010.505567

Olson, E. A., & Wells, G. L. (2004). What makes a good alibi? A proposed taxonomy. *Law and Human Behavior, 28*, 157–176. doi:10.1023/B:LAHU.0000022320.47112.d3

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal cues to deception in the interrogation context. *Law and Human Behavior, 20*, 443–458. doi:10.1007/BF01498980

Pozzulo, J. D., Pettalia, J. L., Dempsey, J. L., & Gooden, A. (2015). Juvenile offenders on trial: Does alibi corroboration evidence and defendant age interact to influence jurors' perceptions and verdicts? *Psychiatry, Psychology and Law, 22*, 224–234. doi:10.1080/13218719.2014.937518

Simon, D. (2012). *In doubt: The psychology of the criminal justice process*. London: Harvard University Press.

Stellar, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135–154). Dordrecht, the Netherlands: Kluwer.

Strange, D., Dysart, J. E., & Loftus, E. F. (2014). Why errors in alibis are not necessarily evidence of guilt. *Zeitschrift für Psychologie, 222*, 82–89. doi:10.1027/2151-2604/a000169

Strömwall, L. A., Granhag, P. A., & Jonsson, A. C. (2003). Deception among pairs: 'Let's say we had lunch and hope they will swallow it!'. *Psychology, Crime & Law, 9*, 109–124. doi:10.1080/1068316031000116238

Van Koppen, M. V., & Van Koppen, P. J. (2010). Alibi's [Alibis]. In P. J. Van Koppen, H. Merckelbach, M. Jelicic, & J. W. De Keijser (Eds.), *Reizen met mijn rechter: Psychologie*

*van het Recht [Traveling with my judge: Psychology of the Law]* (pp. 439–448). Deventer: Kluwer.

Vredeveldt, A., Van Koppen, P. J., & Granhag, P. A. (2014). The inconsistent suspect: A systematic review of different types of consistency in truth tellers and liars. In R. Bull (Ed.), *Investigative interviewing* (pp. 183–202). New York: Springer.

Vrij, A. (2005). Criteria-Based Content Analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy and Law, 11*, 3–41. doi:10.1037/1076-8971.11.1.3

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

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science, 20*, 28–32. doi:10.1177/0963721410391245

Vrij, A., Nahari, G., Isitt, R., & Leal, S. (2016). Using the verifiability lie detection Approach in an insurance claim setting. *Journal of Investigative Psychology and Offender Profiling, 3*. doi:10.1002/jip.1458

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. (1998). Eyewitness identification procedures: Recommendations for lineups and photo-spreads. *Law and Human Behavior, 22*, 603–647.