Examining the effects of evidence disclosure timing and strength on information inconsistencies and provision within investigative interviews

Sam Polman, Kirk Luther, Hannah de Almeida, Janina Eggers & Steven James Watson

To cite this article: Sam Polman, Kirk Luther, Hannah de Almeida, Janina Eggers & Steven James Watson (18 Jul 2024): Examining the effects of evidence disclosure timing and strength on information inconsistencies and provision within investigative interviews, Psychology, Crime & Law, DOI: 10.1080/1068316X.2024.2378061

To link to this article: https://doi.org/10.1080/1068316X.2024.2378061
Examining the effects of evidence disclosure timing and strength on information inconsistencies and provision within investigative interviews

Sam Polman\textsuperscript{a}, Kirk Luther \textsuperscript{b}, Hannah de Almeida\textsuperscript{b}, Janina Eggers\textsuperscript{a} and Steven James Watson\textsuperscript{a}

\textsuperscript{a}Department of Conflict, Risk and Safety, University of Twente, Enschede, The Netherlands; \textsuperscript{b}Department of Psychology, Carleton University, Ottawa, Canada

**ABSTRACT**

Late disclosure of evidence within investigative interviews with guilty suspects has been shown to increase statement-evidence and within-statement inconsistencies, which are indicators of deception. We experimentally tested whether such inconsistencies were influenced by the timing of evidence disclosure and strength of the evidence. We also tested whether evidence disclosure timing or strength had any effect on the provision of novel investigative information, or the rapport and trust between interviewer and interviewee. We employed a 2(Evidence disclosure timing: Early vs Late) x 2(Evidence strength: Weak vs Strong) between-participants design. Participants (\(N = 101\)) role-played a suspect guilty of theft and were interviewed via videoconference. Participants were instructed to convince the interviewer that they were innocent. Late disclosure of evidence led to more statement-evidence inconsistencies and within-statement inconsistencies than early evidence disclosure. Evidence disclosure timing did not affect rapport or the provision of novel investigative information. There were no clear indications of the impact of evidence strength, however, we observed that the manipulations of evidence proximity and reliability did not consistently impact perceptions of the evidence’s strength.

The Strategic Use of Evidence (SUE) technique was initially developed to improve interview diagnosticity – the extent to which behavioral cues can accurately and reliably distinguish between deceptive and truthful suspect statements (Hartwig et al., 2014). However, research into SUE has shifted in two ways (Oleszkiewicz & Watson, 2021). First, studies increasingly examine the role of gradual disclosure of evidence (where evidence is revealed incrementally across the interview) relative to the initial SUE paradigm of late evidence disclosure (where evidence is revealed only after eliciting a suspect’s account). Second, studies increasingly examine whether SUE can be used beyond a...
deception detection method; specifically, to facilitate information disclosure from suspects. A recent meta-analysis of experimental SUE studies revealed that the gradual disclosure of evidence has not yet been clearly shown to increase information disclosure (Oleszkiewicz and Watson 2021). However, the meta-analysis also indicated a lack of studies evaluating whether the classic SUE paradigm of late evidence disclosure has any effect upon information elicitation. There is also a lack of knowledge regarding whether SUE might inhibit the development of a trusting and rapport-based relationship with suspects.

The goal of the current study is to replicate and extend an early influential SUE study; Hartwig et al. (2005). Specifically, Hartwig and colleagues examined the timing of evidence disclosure on cues to deceit. Hartwig et al. found that late disclosure of evidence resulted in more cues to deceit compared to the early disclosure of evidence. We extended this previous work by also examining the impact of evidence disclosure timing on both obtaining novel investigative information and the interviewer-interviewee relationship (i.e. trust and rapport). We also tested whether changing how closely and reliably the disclosed evidence placed the suspect at the location of the crime affected each outcome.

**Suspect interview strategies**

It has been theorized that innocent and guilty suspects try to provide a credible account in different ways during investigative interviews (Granhag & Hartwig, 2008). Innocent suspects tend to adopt a forthcoming strategy. Conversely, guilty suspects tend to adopt an avoidant strategy (e.g. not mentioning case relevant information) to prevent revealing incriminating details, and adopt escape strategies when their account is directly challenged (e.g. silence, refusal to comment, or outright denial). However, avoidance and denial strategies are only credible when the police do not have evidence indicating the suspect is being evasive or deceitful. Consequently, it is critical for guilty suspects to anticipate what evidence the police might hold against them so that they can make sensible decisions about what incriminating details they can deny or avoid mentioning.

The concepts of information and decision control are key to understanding how suspects make decisions about what to admit, avoid, or deny (Granhag & Hartwig, 2008). Information and decision control are cognitive self-regulatory strategies that suspects use to manage the threats investigative interviews pose. Suspects gain information control by predicting what might happen in the interview so that they can prepare a counter strategy (Averill, 1973; Hartwig et al., 2014). Critical to information control is a prediction of what evidence the investigators may hold against the suspect. Decision control is achieved when a suspect decides how to act during an interview, and it is informed by the process of information control (Hartwig et al., 2014; Oleszkiewicz & Watson, 2021). That is, guilty suspects are assumed to make decisions about what to avoid, deny, or admit depending on what evidence they believe the investigators are likely to hold (Hartwig et al., 2007).

As a consequence of seeking information and decision control, guilty suspects are expected to only be forthcoming about details where they believe there is no harm in making an admission. Specifically, guilty suspects will disclose information that either does not implicate them in the crime, or because they believe the interviewer is likely to hold evidence that would make such evasions pointless or damage the credibility of their narrative.
Innocent suspects are expected to be more heavily informed by the cognitive biases belief in a just world and the illusion of transparency (Hartwig et al., 2007). Belief in a just world refers the tendency for people to believe that people get what they deserve (Lerner, 1965). Following this reasoning, an innocent suspect would think it unlikely that they would be punished when they have not committed a crime. The illusion of transparency refers to the belief that people are able to accurately infer others’ states (Gilovich et al., 1998), and so an innocent suspect would have (often unfounded) confidence that the interviewer will be able to perceive the suspect’s innocence. The consequence of these cognitive biases is that innocent suspects often adopt a forthcoming approach in the belief that disclosures will exonerate them and admissions will be interpreted by interviewers as a sign of honesty, and not culpability (Hartwig et al., 2014; Hartwig et al., 2007).

The net result of the different interview strategies employed by guilty and innocent suspects is that innocent suspects are expected to explain evidence against them before the evidence is presented to them; while guilty suspects are expected to omit incriminating details or deny them when directly challenged. Therefore, the evidence presented later in the interview should more often correspond to the account given by an innocent suspect, while it should contradict either directly or by omission, the account given by a guilty suspect.

Evidence disclosure timing and statement inconsistencies

SUE was designed to exploit the aforementioned differences between guilty and innocent suspect strategies to maximize cues to deception in the form of: (i) inconsistencies within guilty suspects’ statements, and (ii) between guilty suspects’ statements and the available evidence (Granthag & Hartwig, 2008, 2015; Hartwig et al., 2005). The most basic version of SUE involves disclosing the evidence held by investigators to the suspect only after the suspect’s account is taken. This late disclosure appears to maximize the differences between truth tellers and liars’ testimony (Granthag & Hartwig, 2015; Hartwig et al., 2005; Oleszkiewicz & Watson, 2021). Late evidence disclosure is argued to be superior to early disclosure because it prevents suspects from changing their accounts to explain already revealed evidence (Granthag & Hartwig, 2008). Specifically, when evidence is disclosed late (cf. early), guilty suspects produce more inconsistencies between their initial statement and the available evidence (statement-evidence inconsistencies). Guilty suspects are also more likely (cf. innocent suspects) to alter their description of events after evidence is presented (within-statement inconsistencies).

Due to the fallibility of memory, some inconsistencies are expected, and do occur, in innocent suspects’ accounts. Nonetheless, meta-analyses have shown that guilty suspects are more likely than innocent suspects to make statement-evidence inconsistencies, regardless of the timing of evidence disclosure (Hartwig et al., 2014; Oleszkiewicz & Watson, 2021). However, the number of these contradictions greatly increases when evidence is disclosed to guilty suspects late in the interview. However, there is no corresponding increase in inconsistencies when innocent suspects have evidence presented late versus early. These findings indicate that, at least in laboratory studies, the late disclosure of evidence only increases the number of inconsistencies significantly within the accounts of guilty suspects. Thus, when a guilty suspect is not aware of the available evidence against them, they are more likely to make statements that contradict the
Evidence against them than innocent suspects are (Hartwig et al., 2014). Therefore, the late disclosure of evidence appears to produce a reliable cue to deception.

Oleszkiewicz and Watson’s (2021) meta-analytic review also considered other methods of evidence disclosure, beyond early versus late. Broadly, there are three general approaches to disclosing evidence: (i) early – before an account is taken from a suspect, (ii) late – after a full account is taken from the suspect, or (iii) gradually – in a drip feed fashion depending on when investigators believe it is appropriate to release specific pieces of information (Oleszkiewicz & Watson, 2021). Police interviewers use all three approaches in practice, with a preference for gradual or late disclosure (Walsh et al., 2016); early disclosure is used by interviewers when they hope a suspect will then confess (Dando & Bull, 2011). Some more recent studies consider the effects of SUE beyond its utility as a deception detection method, and in particular there has been increased attention with regard to whether SUE can be used to facilitate information elicitation from suspects (Oleszkiewicz & Watson, 2021).

**Evidence disclosure timing and novel investigative information**

Novel investigative information has gained popularity in research regarding evidence disclosure because it is of practical use to practitioners (Tekin et al., 2016; Walsh & Bull, 2015). The elicitation of novel investigative information may also be affected by manipulating the timing of the evidence disclosure. However, the direction of any effects remains unclear. According to Oleszkiewicz and Watson (2021), no studies have directly compared the potential effect of late versus early disclosure on the elicitation of novel investigative information. This lack of data reflects that incorporating novel investigative information as an outcome emerged at a time when SUE studies began testing gradual disclosure of evidence to encourage suspects to shift their strategy from avoidant toward forthcoming by manipulating a suspect’s information control (Granhag & Luke, 2018). Briefly, the argument is that gradual disclosure of information encourages a suspect to re-evaluate their prior assumptions about how much information the interviewer holds, most commonly leading them to assume that the interviewer holds more evidence than was assumed before the interview. The result of this reappraisal is that suspects are more likely to provide more details about their involvement in a crime to better match any evidence the interviewer may hold, avoid contradicting evidence, and thus maintain credibility (Luke & Granhag, 2022).

The effect of late (versus early) disclosure of evidence on information elicitation is unknown, but we expect late disclosure of evidence is likely to elicit a more forthcoming account. Immediately disclosing evidence allows suspects to know exactly what information is held by an interviewer, and so we would expect a guilty suspect that does not intend to confess would likely only reveal information that accounts for the revealed evidence. With late disclosure, the evidence held by the interviewer is unknown; thus, suspects may assume the interviewer holds more information than they do and provide more details than would be necessary to explain their activities prior to evidence disclosure. Similarly, suspects might be expected to provide more information as a consequence of information disclosure. Tekin et al. (2016) found that suspects may become more forthcoming when they are confronted with statement-evidence inconsistencies to explain away contradictions and omissions. Since we expect the number of statement evidence
inconsistencies to be higher when evidence is disclosed late, it seems reasonable to predict that late disclosure of evidence may provoke greater information provision as suspects seek to explain away these inconsistencies. However, there is a need to directly test how late (versus early) disclosure of evidence might affect how much information is provided by suspects, especially since we know both methods are used in practice (Dando & Bull, 2011; Walsh et al., 2016).

Evidence disclosure and interviewer-interviewee relationships

Establishing a respectful and honest relationship between interviewer and suspect has repeatedly been shown to be critical to information disclosure (e.g. Alison et al., 2013; Collins et al., 2002; Gabbert et al., 2021). Clemens et al. (2020) surveyed German police officers regarding the SUE technique, and one concern raised was that SUE may interfere with the quality of interaction between interviewer and interviewee. That is, SUE may reduce rapport. Given the central importance of rapport in eliciting information (Gabbert et al., 2021), any reduction in rapport may be detrimental for interview outcomes. Moreover, suspects may feel deceived by late disclosure of evidence, and so late disclosure may also negatively affect how far the interviewee trusts the interviewer. Any violation of trust may also inhibit the willingness of suspects to provide information because trust is also associated with cooperation and information provision by suspects (Collins et al., 2002; Griffiths & Rachlew, 2018). Consequently, we test for any effects of evidence disclosure timing on rapport and trust.

Strength of evidence and novel investigative information

Data from investigative interviews revealed that stronger evidence against a suspect is associated with a greater number of admissions, especially for suspects with limited criminal histories (Cassell & Hayman, 1995; Moston et al., 1992). However, there have been surprisingly limited explorations of the impact of evidence strength on suspect behavior.

In one of the few current explorations of the impact of evidence strength on suspect behavior, Brimbal and Luke (2022) found that statements made by participants about mock crimes were more aligned with the evidence when the evidence was very incriminating. More specifically, they found that more reliable evidence (e.g. evidence from CCTV versus an eyewitness) generated more consistent accounts, while evidence proximity (e.g. placing a suspect in a specific location vs only the general area of a crime) did not affect consistency. Brimbal and Luke argued that more reliable evidence strongly links the suspect to the crime compared to less reliable evidence; consequently, suspects are less inclined to force their stories to account for weaker evidence and so are willing to generate a statement-evidence inconsistency.

Sellers and Kebbell (2009) previously considered the effect of both evidence strength and disclosure timing on suspect confessions. They found that late disclosure of evidence increased the rate of confessions, unless weak evidence was presented in which case confessions became less likely. In fact, late presentation of weak evidence led to some confessions being withdrawn. Sellers and Kebbell did not measure the number of inconsistencies and novel investigative information elicited from suspects. Nonetheless, their results help inform our predictions about the effect of evidence strength on
suspect behavior. Specifically, it seems likely that a suspect’s motivation to account for weak evidence is low given suspects willingness to avoid or even recant confessions within Seller’s and Kebbell’s study. Thus, we anticipate that strong evidence disclosed early in an interview should lead to fewer statement-evidence inconsistencies than weak evidence because we expect suspects are more motivated to provide an alternative account for strong evidence. In contrast, suspects may feel free to omit or deny weak evidence against them. Within-statement inconsistencies for evidence disclosed early should remain low regardless of evidence strength because there is no new evidence presented to provoke contradictions.

Late disclosure might be expected to generate statement evidence inconsistencies regardless of evidence strength because the suspect’s account has been contradicted before the evidence is revealed. However, following Sellers and Kebbell (2009), within-statement inconsistencies should be more likely to be provoked by stronger evidence because suspects may feel less inclined to explain away weak evidence.

Finally, we expect suspects would be more motivated to provide additional information in response to strong over weak evidence regardless of disclosure timing, because suspect motivation to account for evidence and provide additional information or a counter narrative would be expected to be higher when evidence against them is stronger over weaker (Sellers & Kebbell, 2009).

The current study

Oleszkiewicz and Watson (2021) stressed that there have been very few recent studies testing the original form of SUE; that is, evidence disclosed early versus late (as opposed to variations of gradual disclosure of evidence). In particular, there have been no studies that experimentally examine the effect of early versus late disclosure on information provision by suspects. We have also outlined a need for the effect of SUE on the relationship between interviewer and interviewee to be tested. Therefore, the current study aims to partially replicate the fundamental early versus late disclosure of evidence paradigm of Hartwig et al. (2005). Further, we extend this partial replication by examining whether the effects of early versus late disclosure differ depending upon the strength of the evidence that is disclosed.

Summary and hypotheses

All our hypotheses and predictions (pre-registered and available via AsPredicted: https://aspredicted.org/cn87k.pdf) refer only to the behavior of guilty suspects as we do not examine the behavior of innocent suspects in the current study.

First, we hypothesized that early disclosure would lead to fewer statement-evidence and within-statement inconsistencies than late disclosure. Our justification was that early disclosure provides guilty suspects an opportunity to integrate evidence they are aware the interviewer has into their account and arguably facilitates manufacturing a credible account.

Second, we hypothesized that strong evidence would lead to fewer statement-evidence inconsistencies than weak evidence in the case of early disclosure. Our justification is that disclosure of strong evidence has been found to lead to suspects making
statements more consistent with the evidence (Brimbal & Luke, 2022), and further evidence indicating suspects are likely to be more motivated to provide accounts compliant with stronger over weaker evidence (Sellers & Kebbell, 2009). For similar reasons, we also expect late disclosure of evidence to lead to the provision of more novel investigative information because suspects may be motivated to account for initial discrepancies in their story.

We further tested whether the disclosure of strong evidence provokes suspects to provide more novel investigative information than weak evidence, because guilty suspects are motivated to account for evidence that implicates them (Sellers & Kebbell, 2009). Finally, we exploratorily tested whether the evidence strength or disclosure timing affect the perceived rapport or trust with the interviewer from the suspects perspective.

**Methods**

**Participants**

Participants were recruited through the participant recruitment system of a Dutch University, and through the personal contacts of the researchers, with 101 participants in total. Our participants were 54.5% were female ($n = 55$) and 45.5% were male ($n = 46$), between 18 and 63 years of age ($M = 25.55, SD = 7.34$) and 44.6% were students at the host university ($n = 45$), 10.9% were a student somewhere a different university ($n = 11$), and 44.6% were not a student ($n = 45$).

**Design**

The study consisted of a 2 (Evidence Disclosure Timing: Early vs Late) x 2 (Evidence Disclosure Strength: Weak vs Strong) between-subjects design. Participants were randomly assigned to one of the four groups. There were 26 participants in the Early Disclosure of Weak Evidence group, and 25 in all other experimental groups. Sensitivity analysis indicates that with our 2 x 2 design and 101 participants, we would be able to reliably detect effects larger than $\eta^2_p = .07$, using an $\alpha$ of .05 and an 80% power threshold.

**Procedure and materials**

All our materials are available at the project Open Science Framework project page: https://osf.io/jy7pw/. This project took place entirely online using vignettes and video calls rather than mock crimes and in-person interviews because data collection took place during a covid-19 lockdown, which made in-person data collection or crime simulation impossible.

**Pre-interview.** The project received ethical approval from the host institution (Approval number: 210103). The first and fourth authors were directly involved in data collection and participant recruitment. After signing up to take part in the research, an instruction sheet was sent to participants informing them that they were required to portray suspects guilty of a theft and that their task was to convince the interviewing officer of their innocence. The crime in question concerned the theft of Apple AirPods.
from a bag at a library. The participants were asked to imagine that they were did not have much money. They entered the library to search for a new job when a wealthy-looking man next to them spilled his coffee over their notes and then left to go to the bathroom. The participants were told they then looked into his bag, saw a sealed, brand-new pair of Apple AirPods and took them. They then went to a store to sell them in return for a voucher.

To increase the richness of the information available to participants they were given detailed descriptions of the victims clothing and manner, including that they acted arrogantly and failed to apologize when they spilled the coffee cup. We included these details to help participants imagine the scene, and to provide potential motives and mitigation for their imagined criminal act. The spilled coffee also provides scope for a plausible cover story for the participants suspicious activities below the table. Participants were also told details of the scene during the crime. For example, they were informed that there were other people sitting close by, including a person opposite them and a few seats down the table; they were also told that that the theft occurred when the participant believed these people were not looking.

Finally, participants were told that they would be interviewed as a suspect in this crime. Further, they were told to assume that there is some evidence tying them to the crime, but that the interviewer could not yet be sure if they are guilty or not. They were asked to make their best effort to convince the interviewer of their innocence, and to try to imagine how they would feel and act in an actual interview situation. The participants were also told that they could instead confess and justify their behavior to the interviewer, but only if they felt it was not possible to convince the interviewer of their innocence. Participants were not incentivized for their participation.

**Interview.** Participants were invited to a video call with either the first or fourth author, who played one of two roles during the interview: they either played the role of researcher (in which they took care of procedural formalities such as preparing the participant and guiding them through the post hoc measures and debrief) or the role of interviewer. There were some instances in which the participant knew one of the two individuals playing the role of researcher/interviewer (e.g. because they were directly recruited by one of the researchers or because they knew each other from shared courses). In such instances, the role of the interviewer was played by the author that did not know the participant previously/in another capacity to prevent any transfer of rapport from the non-interview context to the interview (Weiher, 2020; Weiher et al., in press). During the interview, the researcher muted and turned off their camera so that only the interviewee (participant) and interviewer were active participants during the interview.

The interviews were based on an interrogation paradigm first introduced by Kassin and Fong (1999) and used by Hartwig et al. (2005) in the initial SUE experiment. The interviews in the Early Disclosure group began with the researcher introducing themselves as an officer and informing participants they were suspected of having stolen Apple AirPods from a table at the nearby library. Participants were then asked whether they had committed the crime. Next, participants were then presented with the evidence against them. The available evidence consisted of the following pieces: (i) in the Weak Evidence group, participants were told that they were logged into a computer in the library at the moment of the theft (low proximity). They were also told that their fingerprints were
found on the victim’s table (low proximity). Lastly, they were told that the receptionist thought she saw someone that looked like the suspect looking in the victim’s bag (low reliability); (ii) in the Strong Evidence group, participants were told that they were logged into the computer next to the victim’s table at the moment of the theft (high proximity). Moreover, they were told that their fingerprints were found on the victim’s bag (high proximity) and that there is camera footage of them going through the victim’s bag (high reliability). Participants were then requested to provide the interviewer with a full account of what they were doing during the time of the crime. Afterwards, they were asked whether they had been in the library, whether they had seen the Apple AirPods, and whether they had touched them. At the end of the interview, they were asked one last time whether they confess to the crime after which they were thanked for their participation. The participants in the Late Disclosure group were also presented with these questions, but the available evidence was only disclosed to the participants after they provided their initial account of their actions at the time of the crime and answered all questions. Evidence was presented by interviewers saying that they had evidence that implicated the participant as a suspect. The evidence was then directly listed to the suspect. After the evidence was disclosed, they were asked to comment on the evidence. After having an opportunity to explain themselves, they were asked again whether they confess to the crime, after which the interview ended.

Post-interview. The interviewer muted themselves and turned the camera off once the interviews ended. The researcher who welcomed the participants asked the participant to complete the online post hoc questionnaire via Qualtrics. The questionnaire included demographic questions regarding their age, occupation, and the interviewee version of the rapport scale for investigative interviews developed by Duke et al. (2018). This scale consists of 21 statements (e.g. ‘The interviewer respects my knowledge’ and ‘The interviewer was attentive to me’) to which the participants could state to what extent they agreed using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Scores are presented as mean item scores, and high scores indicate high rapport. The scale had a good internal consistency (α = 0.94).

Participants also completed a trust scale based on Mayer et al. (1995). The trust scale is composed of the three underlying trust subscales, which include ability (six items; e.g. ‘The interviewer seemed to be well qualified’), benevolence (five items; e.g. ‘I got the impression the interviewer would go out of their way to help me’), and integrity (six items; e.g. ‘Sound principles seemed to guide the interviewer’s behavior’). Trust was indicated on a 5-point Likert scale (anchored at 1 = strongly disagree, and 5 = strongly agree). Scores are presented as mean item scores and high scores indicated higher trust. The internal consistency of all test items was good (Cronbach’s α = .95).

As a manipulation check, participants were ultimately asked two questions about the three pieces of evidence they were confronted with. First, they were asked to indicate how strongly they felt like the pieces of evidence implicated them as being guilty. Second, they were asked how difficult they found it to come up with and explanation for the pieces of evidence. Participants indicated how strongly they felt the evidence implicated guilt on a 5-point Likert scale ranging from 1 (very weak) to 5 (very strong), and how difficult it was to explain the evidence on a 5-point scale ranging from 1 (very easy) to 5 (very hard).

Interview coding. All interviews were independently coded by both the first and third authors. Statement-evidence inconsistencies were coded if participants made a statement
that either contradicted or failed to account for one of the three pieces of evidence the interviewer held against the suspect (e.g. ‘I was not at the library’ is a statement-evidence inconsistency because the evidence showed they were logged onto a library computer during the time of the crime). There were three pieces of evidence held by the interviewer and therefore the possible range for statement-evidence inconsistencies is 0-3.

Within-statement inconsistencies were coded when participants made a remark that contradicted something they had stated previously in the interview, but only where this related to one of the three pieces of evidence. For example, the statement ‘Oh yeah, I dropped my pen into his bag so I guess I would have touched it getting it back’ is a within-statement inconsistency if they stated earlier that they had not touched the victim’s bag. Similar to statement-evidence inconsistencies, we expect the number of within-statement inconsistencies to be between 0–3 because three pieces of evidence are disclosed. Although it would be possible for participants to exceed this value if they volunteered additional changes to their story (for example, because of imperfect memory).

We define novel investigative information as details which improve understanding of the incident or could advance the investigation (Cyr & Lamb, 2009). Therefore, novel investigative information was coded whenever the participants disclosed any verifiable statements about what happened during the crime which were not simple agreement with the pieces of evidence held by the interviewer. For example, a participant saying that they were in the library would not provide new information. In other words, any statement the interviewer could plausibly check for veracity, but which was not already known to the interviewer based on the evidence already held, was coded as novel investigative information. The amount of novel investigative information that can be provided is therefore not fixed by the experimental design.

We assessed initial interrater agreement between the two coders on the number of statement-evidence and within-statement inconsistencies using Cohen’s Kappa. We found good agreement for the number of statement evidence inconsistencies (κ = .75), though agreement was lower for within-statement inconsistencies (κ = .66). Because agreement was suboptimal for both types of inconsistencies, the two coders reviewed the discrepancies with a third coder (the final author). For statement-evidence inconsistencies, we identified a systematic difference in coding between the two coders which accounted for disagreements. Where participants gave a cover story which simultaneously addressed two pieces of evidence (for example, saying they searched the victim’s bag addressed both having fingerprints on the victim’s bag and the evidence they looked into the bag) had been coded as addressing only one piece of evidence by one coder, rather than two. We corrected this issue for the final analysis and these agreed codes are used for all analyses.

For within-statement inconsistencies, discrepancies occurred either through error (a change was simply missed), or because there was some ambiguity in the wording used by participants. For example, one coder may interpret ‘I heard a noise coming from their bag and so searched the top of their bag’ as indicating a single change in the story (because it explains the bag being touched) while the second interpreted this as implying looking within the bag (thus explaining both the fingerprints on the bag and being seen looking through the bag). During the review, only when it was very clear that a change in story addressed multiple pieces of evidence were story changes coded as implying more than one change to a story. Otherwise, we took a cautious
approach of only counting changes that are directly mentioned. The agreed codes are used for analysis.

We assessed reliability of the rating for the number of novel investigative information with a Pearson correlation given that we had two raters, which indicated a strong association between scores of the two raters ($r = .78$). For robustness, we also estimated reliability using an Intraclass correlation coefficient with a two-way mixed model assessing consistency in agreement, which indicated good agreement ($ICC = .84$). We used the mean score of the two raters for analyses.

**Results**

Table 1 presents the means, standard deviations, and correlations between statement-evidence inconsistencies, within-statement inconsistencies, novel investigative information, rapport, and trust.

Statement-evidence inconsistencies and within-statement inconsistencies were positively correlated, $r_s(99) = .47$, $p < .001$, while statement-evidence inconsistencies and novel investigative information were negatively correlated, $r_s (99) = -.41$, $p = .001$. In contrast, within-statement inconsistencies and novel investigative information were positively correlated, $r_s (99) = .22$, $p = .029$. Rapport and novel investigative information were not correlated, $r_s (99) = .03$, $p = .756$.

Table 2 shows the number of statement-evidence inconsistencies that were provided by participants per experimental group. Only two people contradicted all three pieces of evidence by either stating they were not present at the library or by saying nothing at all.

For early disclosure, there was an apparent stark impact of evidence strength, with 68% of participants giving an account consistent with all disclosed evidence when evidence was strong, compared to only 27% when evidence was weak.

For late disclosure categories (weak and strong evidence), most participants made statements that contradicted two pieces of evidence. Almost all participants placed themselves at the scene of the crime by admitting being at the library, but failed to provide a plausible explanation for the other two pieces of evidence (their fingerprints being found on the victim’s table/bag and being spotted by the receptionist/CCTV camera looking into the victim’s bag).

Broadly, these findings support our prediction that early disclosure is associated with fewer statement-evidence inconsistencies, especially when there is strong evidence.

We also considered the type of statement-evidence inconsistencies participants made in our sample (omissions vs fabrications). Across all conditions, only 1.4% of statement-

---

**Table 1.** Descriptive statistics and correlations.

<table>
<thead>
<tr>
<th></th>
<th>$M$</th>
<th>$SD$</th>
<th>Rapport</th>
<th>Trust*</th>
<th>SEI*</th>
<th>WSI*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapport</td>
<td>3.90</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trust</td>
<td>3.29</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEI*</td>
<td>1.45</td>
<td>0.89</td>
<td>-.04</td>
<td>-.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSI*</td>
<td>0.60</td>
<td>0.81</td>
<td>.04</td>
<td>.05</td>
<td>.47</td>
<td></td>
</tr>
<tr>
<td>NII*</td>
<td>1.88</td>
<td>1.64</td>
<td>.03</td>
<td>.08</td>
<td>-.41</td>
<td>.22</td>
</tr>
</tbody>
</table>

$N = 101$, SEI = Statement-evidence inconsistency, WSI = within statement inconsistency, NII = Novel Investigative information. Rapport and Trust are measured via 1–5 Likert scales. SEI and WSI are bound between scores of 0-3.

*Note that correlations with these variables are Spearman’s Rho rather than Pearson correlations.

Coefficients in *italics* represent $p < .05$, those in *bold* $p < .01$, and *bold italics* $p < .001$. 
evidence inconsistencies consisted of fabrications (statements that directly contradicted the evidence). The remaining 98.6% of statement-evidence inconsistencies were omissions (statements where the suspect did not mention anything to contradict or account for the available evidence).

Table 3 shows the number of within-statement inconsistencies provided by the suspects as a function of evidence strength and disclosure timing. In both early disclosure categories (weak and strong evidence), within-statement inconsistencies were extremely rare. For late disclosure, most participants made one within-statement inconsistency with weak evidence compared to two with strong evidence. These results support our prediction that there are more within-statement inconsistencies with late versus early disclosure.

**Manipulation checks**

We tested whether participants in the different evidence strength groups did or did not perceive the pieces of evidence as stronger or weaker. Since the majority of the comparisons are for single item ordinal measures, we compared scores using Mann Whitney U tests. Table 4 indicates the extent participants thought each piece of evidence implicated them as being guilty, and how hard they thought each piece of evidence was to explain during the interview, as well as median (quartile) scores across the three pieces of evidence.

Participants mostly did not report different perceptions of evidence strength depending on the experimental condition. The only exception was that fingerprints on the victim’s bag were considered more implicating of guilt than fingerprints only on the table next to where the theft took place. While not statistically significant, surprisingly

**Table 2. Frequencies of statement-evidence inconsistencies.**

<table>
<thead>
<tr>
<th>Evidence Strength</th>
<th>No. Statement-Evidence Inconsistencies</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early Disclosure</td>
<td>Late Disclosure</td>
</tr>
<tr>
<td>Weak</td>
<td>0</td>
<td>7 (27%)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>6 (23%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>13 (50%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Strong</td>
<td>0</td>
<td>17 (68%)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 (4%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6 (24%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1 (4%)</td>
</tr>
</tbody>
</table>

**Table 3. Frequencies of within-statement inconsistencies.**

<table>
<thead>
<tr>
<th>Evidence Strength</th>
<th>No. Within-Statement Inconsistencies</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early Disclosure</td>
<td>Late Disclosure</td>
</tr>
<tr>
<td>Weak</td>
<td>0</td>
<td>24 (92%)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 (4%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1 (4%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Strong</td>
<td>0</td>
<td>23 (92%)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 (4%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1 (4%)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
participants also reported that it may have been harder to explain the (lower reliability) eyewitness testimony than the (higher reliability) camera footage.

Since we intended that the evidence would increase in how strongly it implicated participant’s guilt and become harder to explain, we also performed Friedman tests to see if the strength of evidence was perceived to increase from the first to the third piece of evidence disclosed. These tests showed a statistically significant effect regarding how much participants believed the evidence implicated them as guilty, \( \chi^2 (2) = 55.27, p < .001 \); and the extent to which evidence became harder to explain, \( \chi^2 (2) = 42.00, p < .001 \). Post-hoc comparisons with Bonferroni adjustment showed that the first piece of evidence differed with the second and third piece of evidence regarding both how much participants believed they implicated them as guilty and how hard they were to explain (all \( p < .001 \)). However, the second and third pieces of evidence were not considered to implicate guilt more strongly (\( p = .201 \)), nor to differ in how hard they were to explain (\( p = .690 \)).

### Hypothesis testing

We pre-registered that we would use a factorial ANOVA comparing Disclosure Timing (early versus late) and Strength of Evidence (weak versus strong) for each of our dependent variables. However, the statement-evidence and within-statement inconsistencies and novel investigative information dependent variables all violated the relevant parametric assumptions. Therefore, we instead analyzed these variables using Mann Whitney \( U \) tests to test the main effects of Disclosure Timing and Evidence Strength. We probed the interaction effects using a Kruskal Wallis test with Dunn post hoc tests using Bonferroni adjustment for multiple testing to make comparisons between all experimental groups. For full transparency, and to reflect the parameters reported in our sensitivity analysis, we also report the results of our comparisons using our pre-registered ANOVA. We present these analyses within Table 5.

Table 5 shows that while statement-evidence and within-statement inconsistencies are elicited to a statistically significantly greater extent with Late (cf. early) Disclosure, there was no clear effect of Evidence Strength. Similarly, while the Kruskal Wallis test across all four experimental groups was statistically significant for both statement-evidence and within-statement inconsistencies, all statistically significant

---

**Table 4.** Comparison of perceptions of how implicating and hard to explain evidence was between weak and strong evidence conditions.

<table>
<thead>
<tr>
<th>Evidence piece</th>
<th>Outcome</th>
<th>Weak evidence</th>
<th>Strong Evidence</th>
<th>U</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mdn Quartiles</td>
<td>Mdn Quartiles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logged into computer</td>
<td>Implicating</td>
<td>2.00, 1.00, 2.00</td>
<td>1.00, 1.00, 3.00</td>
<td>1234</td>
<td>.765</td>
</tr>
<tr>
<td></td>
<td>Hard to explain</td>
<td>1.00, 1.00, 2.00</td>
<td>1.00, 1.00, 2.00</td>
<td>1099.5</td>
<td>.166</td>
</tr>
<tr>
<td>Fingerprints</td>
<td>Implicating</td>
<td>4.00, 2.00, 4.00</td>
<td>4.00, 3.00, 5.00</td>
<td>844.5</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td>Hard to explain</td>
<td>3.00, 1.00, 3.00</td>
<td>3.00, 2.00, 4.00</td>
<td>1075.5</td>
<td>.166</td>
</tr>
<tr>
<td>Observed at bag</td>
<td>Implicating</td>
<td>4.00, 3.00, 4.00</td>
<td>4.00, 2.75, 4.00</td>
<td>1271</td>
<td>.977</td>
</tr>
<tr>
<td></td>
<td>Hard to explain</td>
<td>3.00, 2.00, 4.00</td>
<td>2.00, 2.00, 4.00</td>
<td>1001</td>
<td>.055</td>
</tr>
<tr>
<td>Total</td>
<td>Implicating</td>
<td>3.00, 2.33, 3.33</td>
<td>3.33, 2.67, 3.67</td>
<td>1022</td>
<td>.082</td>
</tr>
<tr>
<td></td>
<td>Hard to explain</td>
<td>2.33, 2.00, 3.00</td>
<td>2.33, 1.67, 3</td>
<td>1218</td>
<td>.697</td>
</tr>
</tbody>
</table>

All measures are on a 1–5 Likert scale, where higher scores indicate greater perceived implication of guilt from evidence, or that it was harder to explain.

Rows in **bold** represent group differences with \( p < .01 \).
comparisons were between groups which differed between Early and Late disclosure timing.2

Table 5 also shows that there was no evidence that either Disclosure Timing or Evidence Strength had any impact on the disclosure of novel investigative information.

**Rapport.** A factorial ANOVA tested for effects of Disclosure Timing (early vs late) and Strength of Evidence (weak vs strong) on Rapport. We observed no statistically significant main effects for Disclosure Timing, $F(1, 97) = 0.3, p = .606, \eta^2_p < .01$, or Evidence Strength, $F(1, 97) < 0.1, p = .929, \eta^2_p < .01$. An exploratory Bayesian ANOVA revealed moderate evidence in favor of the null hypotheses for Disclosure Timing ($BF_{01} = 4.20$) and Strength of Evidence ($BF_{01} = 4.76$). Therefore, we found no meaningful difference in the perceived rapport between those to whom evidence was presented Early vs Late ($M = 3.86, SD = 0.78$ vs. $M = 3.94, SD = 0.78, d = 0.11, 95%CI[−0.28, 0.50])$, nor Weak vs Strong evidence ($M = 3.90, SD = 0.75$ vs. $M = 3.89, SD = 0.73, d = 0.01, 95%CI[−0.38, 0.40]$).

We also observed no significant interaction effect between Disclosure Timing and Evidence Strength, $F(1, 97) = 1.7, p = .200$. An exploratory Bayesian ANOVA showed very strong evidence in favor of the null hypothesis $BF_{01} = 37.80$, which indicates there was no significant difference in the perceived rapport for suspects who were presented with Weak Evidence Early ($M = 3.77, SD = 0.76$), Strong Evidence Early ($M = 3.95, SD = 0.81$), Weak Evidence Late ($M = 4.04, SD = 0.74$) and Strong Evidence Late ($M = 3.83, SD = 0.65$).

In summary, we did not find evidence that timing of evidence disclosure or strength of evidence had any impact on participants’ self-reported rapport.

**Table 5.** Comparisons of Group Means for Evidence Disclosure Timing and Evidence Strength with regard to Statement-evidence Inconsistencies, Within-statement Inconsistencies and Novel Investigative Information.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEI</td>
</tr>
<tr>
<td>Disclosure Timing</td>
<td></td>
</tr>
<tr>
<td>Early</td>
<td>$M$</td>
</tr>
<tr>
<td></td>
<td>0.94</td>
</tr>
<tr>
<td>Late</td>
<td>1.96</td>
</tr>
<tr>
<td>Mann Whitney</td>
<td>$U = 556.5, p &lt; .001$</td>
</tr>
<tr>
<td>ANOVA</td>
<td>$F = 53.3, p &lt; .001, \eta^2_p = .36$</td>
</tr>
<tr>
<td>Evidence Strength</td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>$M$</td>
</tr>
<tr>
<td></td>
<td>1.61</td>
</tr>
<tr>
<td>Strong</td>
<td>1.29</td>
</tr>
<tr>
<td>Mann Whitney</td>
<td>$U = 1084.0, p = .122$</td>
</tr>
<tr>
<td>ANOVA</td>
<td>$F = 5.7, p = .019, \eta^2_p = .06$</td>
</tr>
<tr>
<td>Interaction Term</td>
<td></td>
</tr>
<tr>
<td>Early / Weak</td>
<td>$M$</td>
</tr>
<tr>
<td></td>
<td>1.23</td>
</tr>
<tr>
<td>Early / Strong</td>
<td>0.64</td>
</tr>
<tr>
<td>Late / Weak</td>
<td>2.00</td>
</tr>
<tr>
<td>Late / Strong</td>
<td>1.92</td>
</tr>
<tr>
<td>Kruskal Wallis</td>
<td>$H = 37.7, p &lt; .001$</td>
</tr>
<tr>
<td>ANOVA</td>
<td>$F = 3.3, p = .072, \eta^2_p = .03$</td>
</tr>
</tbody>
</table>

Note. SEI = Statement-Evidence inconsistencies, WSI = Within-Statement Inconsistencies, NII = Novel Investigative Information.

Effects in *italics* represent $p < .05$, those in **bold** $p < .01$, and **bold italics** $p < .001$. 

14 S. POLMAN ET AL.
Trust. We also observed no statistically significant main effects on Trust for either Disclosure Timing, $F(1, 97) = 0.3, p = .575, \eta^2_p < .01$, or Evidence Strength, $F(1, 97) < .1, p = .864, \eta^2_p < .01$. An exploratory Bayesian ANOVA revealed moderate evidence in favor of the null hypotheses for Disclosure Timing ($BF_{01} = 4.11$) or Strength of Evidence ($BF_{01} = 4.68$). Therefore, we found no difference in the perceived rapport between those to whom evidence was presented Early vs Late ($M = 3.24, SD = 0.99$ vs $M = 3.34, SD = 0.78, d = 0.11, 95\% CI [−0.28, 0.50]$); nor Weak vs Strong Evidence ($M = 3.27, SD = 0.92$ vs. $M = 3.30, SD = 0.86, d = −0.03, 95\% CI[−0.42, 0.36]$).

We did observe a significant interaction effect between Disclosure Timing and Evidence Strength, $F(1, 97) = 4.9, p = .030, \eta^2_p = .05$. Tests of simple effects revealed no statistically significant differences between any of the experimental groups. Trust was non-statistically significantly lower when Weak versus Strong evidence is revealed Early ($M = 3.03, SD = 0.97$ vs $M = 3.45, SD = 0.97, d = 0.43, 95\% CI[−0.83, − 0.04]$; $F(1,97) = 2.9, p = .094, \eta^2_p = .03$). There was slightly stronger evidence that Trust is higher when Weak evidence is released Early vs Late ($M = 3.03, SD = 0.97$ vs. $M = 3.51, SD = 0.81, d = 0.54, 95\% CI[0.14, 0.93]$; $F(1,97) = 3.9, p = .052, \eta^2_p = .04$). Trust did not differ significantly when Strong evidence was presented Late ($M = 3.16, SD = 0.72$) to any of the other experimental groups.

Overall, we did not find any clear evidence for effects of evidence disclosure timing or evidence strength on trust, but it may be the case that early disclosure of stronger evidence is associated with higher trust than early disclosure of weaker evidence. This effect may be reversed with late disclosure, where the late release of weak evidence may lead to reduced trust compared to the late disclosure of stronger evidence.

Discussion

We found clear support that late disclosure of evidence produces both more statement-evidence and within-statement inconsistencies. The proportion of statement-evidence inconsistencies appeared especially low when strong compared to weak evidence was revealed early in the interview. Nonetheless, statement-evidence inconsistencies were not found to differ depending on evidence strength when evidence is released early after correction for multiple testing. Neither evidence strength nor disclosure timing had any effect on the amount of novel investigative information revealed. We also found no evidence to suggest that evidence strength or disclosure timing had any negative impact upon rapport between interviewer and interviewee, with only weak indications of any effects on trust where early disclosure of weak evidence may be detrimental.

Statement-Evidence inconsistencies

As hypothesized, early evidence disclosure produced fewer statement-evidence inconsistencies than late evidence disclosure. This is consistent with the argument that early disclosure provides suspects with an opportunity to create a plausible account that covers each piece of evidence (Hartwig et al., 2014; Hartwig et al., 2005; Oleszkiewicz & Watson, 2021). Following Brimbal and Luke (2022), we reasoned that if this is why statement-evidence inconsistencies are lower with early disclosure, then suspects should be
especially motivated to make sure their story aligns with stronger (cf. weaker) evidence against them. We found that almost 75% of participants made at least one statement-evidence inconsistency when weak evidence was disclosed early, while only 32% made any statement evidence inconsistencies when strong evidence was released early. This difference in statement-evidence inconsistencies between weak and strong evidence tentatively supported our hypothesis that participants felt able to contradict, or avoid accounting for, evidence they considered weak. Nonetheless, a Dunn test indicated no statistically significant difference in the number of statement-evidence inconsistencies between early disclosure of either strong or weak evidence once a Bonferroni correction was applied. We conclude that the balance of probability is that evidence strength does not affect the number of statement-evidence inconsistencies with early disclosure, but also believe that a replication of these findings would provide more certainty. In particular, because our sample size was not large, and our necessitated shift to non-parametric statistics further reduced our statistical power.

One important observation was that most statement evidence inconsistencies represented omissions rather than direct contradictions of the existing evidence. These findings correspond with theoretical assumptions that suspects adopt an avoidant strategy prior to the disclosure of evidence (Granhag & Hartwig, 2008), and more general findings that liars tend to provide fewer details than truth tellers (DePaulo et al., 2003). However, the first piece of evidence (being logged onto a library computer) was rarely omitted. Based on the research of Hartwig et al. (2005), our participants were told that they were brought in for questioning because the police know they were ‘in the area’ on the day of the crime. Similarly, the description of the criminal act made it clear there would be multiple potential witnesses that the suspect was at the table, but that these witnesses were not looking when the theft occurred. Following the principles of information control, most participants could thereby be expected to predict interviewers would hold evidence that they were in the area; even in the late disclosure condition when this evidence was not directly disclosed. Consequently, we think it is likely that most participants felt very comfortable admitting to being in the area of the crime.

Overall, our findings provide good support to the arguments that late disclosure of evidence produces statement-evidence inconsistencies. We have more tentatively shown that participants may be especially motivated to explain evidence against them when the evidence is strong. However, this latter result requires further investigation.

Within-Statement inconsistencies

We also support previous findings that late disclosure of evidence leads to more within-statement inconsistencies than early disclosure (Hartwig et al., 2014; Oleszkiewicz & Watson, 2021). In the case of early disclosure, there were very few within-statement inconsistencies because participants could construct a narrative that accounted for all the evidence against them. With late disclosure, within-statement inconsistencies were more frequent because suspects adjusted their story to account for omissions to their story (or rarely, to address an outright lie or denial).

Contrary to our predictions, evidence strength had no bearing on the elicitation of within-statement inconsistencies. We expected participants to be more motivated to change their accounts when presented with strong evidence (compared to weak
evidence) late in the interview. However, it appears motivation to account for inconsistencies with evidence by changing an initial account were not dependent on evidence strength. Rather, the presentation of any evidence – weak or strong – prompted changes in suspects’ stories.

**Novel investigative information**

One critique of research into SUE has been a focus on statement inconsistencies as an outcome variable (Oleszkiewicz et al., 2023). In particular, the assumption that inconsistencies are diagnostic of deceit (Bagnall et al., 2023). This assumption is reasonable in laboratory settings where experimenters know the true narrative of events. Therefore, even if individual inconsistencies may represent memory errors by participants, across experimental groups differences in inconsistency frequency are very likely to represent deliberate counter interrogation strategies. Consequently, inconsistencies do provide a reasonable proxy measure of deception. Outside of the laboratory, this assumption becomes more problematic. Every suspect will have a unique experience, and individual differences in ability to recall accurate information under often stressful interview conditions cannot be accounted for through randomization. Therefore, interpreting inconsistencies as indicating deception could be problematic; especially for vulnerable suspects with executive functioning difficulties. For example, autistic suspects (Bagnall et al., 2023), and suspects with ADHD or other intellectual impairments (Gudjonsson et al., 2007) may be particularly vulnerable. In addition to memory errors and attempts to conceal and deceive, omissions also reflect suspect’s assumptions about what investigators will or will not find interesting or useful (Clemens & Grolig, 2019). Even if inconsistencies were always direct cues to deception, deception does not always indicate guilt. For example, suspects may lie or avoiding disclosures within investigative interviews to preserve their reputation by hiding unrelated unlawful or lawful but immoral behaviors (Clemens & Grolig, 2019).

Given the practical limitations of applying inconsistencies within genuine investigative interviews, and reflecting that the main goal in any suspect interview is the recovery of important crime-relevant information (Walsh & Bull, 2015), research on evidence disclosure strategies has moved from a primary focus on deception detection to how evidence disclosure can facilitate disclosures from suspects (Luke & Granhag, 2022; May et al., 2017; Tekin et al., 2016). Thus far, laboratory studies have not shown convincingly whether SUE does or does not help facilitate information elicitation (Oleszkiewicz & Watson, 2021). In line with the overall findings of the meta-analysis by Oleszkiewicz and Watson (2021), we found no significant effects of either disclosure timing or evidence strength on the number of pieces of novel investigative information suspects shared.

However, we would not yet suggest that SUE is ineffective for eliciting novel investigative information. The use of SUE as an information elicitation paradigm is still developing, and early indicators suggest that both the manner in which evidence is disclosed and how inconsistencies are challenged are likely to be very important in whether evidence disclosure leads to enhanced disclosure. For example, Luke and Granhag (2022) showed that challenging inconsistencies in a non-accusatory manner leads to more information disclosure more than when inconsistencies are not directly challenged. In other words, it is not enough to generate inconsistencies and hope that these will provoke a change in the suspect’s strategy; the interviewer should take an active role in facilitating this
desired behavioral change (May et al., 2017), which can be achieved without resorting to coercive or accusatory interviewing methods.

Research by Oleszkiewicz et al. (2023) adds more nuance to the interpretation of information provision and inconsistencies. We, in line with most prior SUE research, counted statement-evidence inconsistencies whether evidence was directly contradicted or merely omitted. We also indicated that novel information had been identified whenever any disclosure from a suspect provided verifiable investigative information not already known by the interviewer. Oleszkiewicz et al. (2023) showed that without the use of careful follow up questions, such inconsistencies and disclosures often lack practical utility because they are only partial and so fail to provide sufficient information to inform whether cases should be prosecuted or not.

The critiques from both Luke and Granhag (2022) and Oleszkiewicz et al. (2023) provide important context for interpreting our findings. We followed as closely we could the procedures outlined in (Hartwig et al., 2005). Our interview script began and ended with a request for the suspect to confess, and direct statements about having evidence against the suspect, which suspects are then asked to explain; this script is thereby quite accusatory. Plausibly then, developing interview scripts that align more closely with a non-accusatory interviewing approach may be more likely to facilitate provision of reliable novel information in response to evidence disclosure than the methods we have employed here.

Evidence disclosure and the interviewer-interviewee relationship

One concern about the use of SUE is that it might inhibit the quality of the relationship between the interviewer and suspect (Clemens et al., 2020). For example, by making interviewers conform to a rigid structure and thus inhibiting the quality of communication, or by making suspects feel the interviewer has behaved unfairly by withholding evidence. Promisingly, we did not find convincing evidence SUE harms the interviewer-interviewee relationship. Rapport was unaffected by the late disclosure of evidence. While we did find evidence for an interaction effect of disclosure timing and evidence strength on trust, the strength of that interaction effect was modest. Specifically, we found some evidence that trust was higher when weak evidence was presented late compared to early, and weaker evidence that trust was higher when strong evidence was presented early than when weak evidence is presented early. Common to both tentative effects was that the lowest level of trust was found for the early disclosure of weak evidence which could be problematic given the importance of developing a cooperative relationship for information disclosure. However, while the overall interaction effect between disclosure timing and evidence strength was statistically significant, the identified partial eta squared was lower than our sensitivity analysis indicated our sample size could reliably detect. The specific comparisons between groups assessed through tests of simple effects were also not quite statistically significant. Therefore, we place low confidence on these specific results. We would seek to replicate these results in the future, but believe future research could more generally target the impacts of evidence disclosure on trust. Following the earlier critique of Oleszkiewicz et al. (2023), it seems likely that whether evidence is disclosed in an accusatory or accusatory frame would impact the extent to which disclosure timing affects trust.
The observed null effects for rapport are important, given the centrality of rapport to securing the cooperation of suspects (Alison et al., 2013; Gabbert et al., 2021; Walsh & Bull, 2012). However, any such effects are likely to be small when interviews are brief, and cases are simple. Given these limitations, we think it is important for researchers testing SUE to consider the interpersonal dynamics of the interview participants, especially rapport, so that our promising but tentative findings can be confirmed in suspect samples and with more complex criminal cases.

‘Strength’ of evidence?

One important question raised by our research is what we mean by the term ‘strength of evidence’. We did not set out to test what aspects of evidence participants perceive as strong or weak. Rather, we intended to have evidence escalate in how much it implicated suspects following Hartwig et al. (2005), and varied the evidence either in terms of how closely it tied the suspects to the scene of the crime (proximity) or how reproducible the evidence was (reliability). Nonetheless, our manipulation checks showed that participants did not react to our evidence manipulation as expected.

While the first piece of evidence (i.e. placing suspects at the scene of the crime) was considered weaker than the later evidence, participants did not consider the second piece of proximity evidence (fingerprints on the victim’s table or victim’s bag) as stronger or weaker than the corresponding third piece of evidence (being observed interacting with the victim’s bag via CCTV or an eyewitness). The lack of difference between these two pieces of evidence may reflect that participants considered all forms of evidence to be reliable. Brimbal and Luke (2022) showed that reliability impacts perceptions of evidence strength more than proximity and fingerprints are considered to be highly reliable evidence by both experts and non-experts (Martire et al., 2019). Similarly, non-experts have been shown to be relatively unaware of reliability concerns with eye-witness testimony (Benton et al., 2006). Following these considerations, our non-expert sample may have considered the eye-witness, CCTV and fingerprint evidence as all representing reliable evidence. These observations will hopefully be useful to future researchers seeking to manipulate evidence strength, especially with non-expert samples who may need the relative reliability of different forms of evidence directly explained to them.

Broadly, our findings showing that the evidence related to reliability was considered more implicating than evidence that was purely related to proximity, supporting the findings of Brimbal and Luke (2022). A departure from Brimbal and Luke (2022) was that our participants did perceive the proximity of the fingerprint evidence to matter for their perception of evidence strength. That is, participants considered fingerprints on the victim’s bag to be more implicating of guilt than fingerprints on only the victim’s table. It remains possible then that evidence proximity can be an important aspect of how strong evidence is perceived to be by suspects, at least when the evidence presented is perceived to be reliable.

Limitations

Part of the motivation for this study was to address a call for more research exploring the effects of early versus late disclosure (Oleszkiewicz & Watson, 2021). We attempted to
replicate, as far as we could, the interview procedures outlined in (Hartwig et al., 2005). Based on our experiences of adopting this approach, our main concerns with the scripts were that they were short, accusatory, and did not direct interviewers to probe suspect’s initial narratives prior to evidence disclosure. Consequently, suspects almost never opted to directly lie in our sample. That is, less than 2% of the statement-evidence inconsistencies we observed contradicted, rather than omitted, evidence. It is noticeable that later studies investigating SUE have moved from testing simple crimes to more complex crimes, as well as incorporating more complex interviewing including the gradual disclosure of evidence throughout the interview (Oleszkiewicz & Watson, 2021). We believe that this shift to increased complexity is a sensible response to ensure that laboratory interviews more closely reflect what happens in practice, but also to more fully test the efficacy of evidence disclosure methods.

Another limitation of our study was a consequence of design constraints due to conducting the study during the covid-19 lockdown. Specifically, we were required to conduct our study online due to restrictions on in-person testing at the time. Conducting the study online necessitated the use of video conferencing to carry out the interviews and the use of vignettes rather than have participants take part in a simulated crime. Experiments assessing the use of online interviews in the context of witness interviews indicates that while it is possible to build rapport and maintain disclosures (Dion Larivière et al., 2023), however the impact on (mock) suspect interviews is currently unknown. The use of vignettes is potentially problematic because, in a traditional simulated crime scenario, participant’s experiences of committing a mock crime are (reasonably) consistent, and the experimenter can be sure that the participants have direct experience of all relevant activities and locations. With a vignette, experimenter control is significantly reduced. While we provided participants with a rich description of the environment and included a timeline of their activities before, during, and after the crime, we cannot be sure how closely participants attended to this information. Similarly, some participants may have been more able to imagine the events and location, and so provide novel investigative information. These concerns are certainly problematic; however, we do not feel they are likely to invalidate our results. That is, there is no reason to believe these features would systematically vary between our evidence strength experimental conditions. The early disclosure of evidence could give participants additional information to provide a coherent story that addresses all the evidence. However, early disclosure of evidence is proposed to reduce statement-evidence inconsistencies through precisely this mechanism. Similarly, we hypothesized that participants would be more or less motivated to provide novel investigative information depending on evidence strength and disclosure timing, not necessarily that the provided information would always be of high investigative value. We do believe this somewhat reflects genuine interviews where suspects’ attempts to explain away evidence and provide a counter narrative can also vary in quality. However, we depart from genuine interviews after this point because our interviewers, unlike actual investigators, would not probe to test the quality of the suspect’s narrative.

Despite the limitations, our study adds value to the literature being one of the only independent replications of the SUE technique (Oleszkiewicz & Watson, 2021). We hope that sharing the methodological difficulties we encountered will be of use to other researchers. Further, many of our results align with the published literature,
adding to a convergence of evidence on the benefits of various evidence disclosure strategies.

**Conclusion**

Overall, while we find support for the use of SUE as a deception detection tool when comparing across groups in laboratory settings, and tentative evidence SUE does not lower the quality of the interviewer-interviewee relationship, we do not find that the basic late disclosure version of SUE leads to suspects providing additional information to their interviewers.

**Notes**

1. Note that this scale includes several items which overlap with our measure of trust because they assess, in our opinion, the perceived competence and integrity of the interviewer. We also conducted all our analyses using the rapport measure with these items excluded; our results did not change from those we found using the full rapport scale. We include this measure of rapport that omits the items that overlap with trust in our open data. Omitted items were: ‘I think the Interviewer is generally honest with me.’, ‘The Interviewer did his/her job with skill during the interview.’, ‘The Interviewer performed expertly during the interview.’, ‘I think that the Interviewer can generally be trusted to keep his/her word.’, ‘I feel I can trust the Interviewer to keep his/her word to me.’, ‘The Interviewer made an effort to do a good job’, and ‘The Interviewer acted like a professional.’
2. The mean SEIs of the Strong Evidence Early Disclosure and Weak Evidence Early Disclosure groups do appear different, with fewer inconsistencies when evidence is strong. However, after applying a Bonferroni correction the difference between these groups was no longer statistically significant ($p = .295$).

**Open Scholarship**

This article has earned the Center for Open Science badges for Open Data, Open Materials and Pre-registered. The data and materials are openly accessible at https://osf.io/c72b6/, https://osf.io/c72b6/ and https://osf.io/c72b6/registrations.

**Acknowledgements**

We thank Mariëlle Stel for feedback on an earlier draft of this manuscript. We would also like to thanks two anonymous reviewers. This article greatly benefitted from excellent constructive feedback, and we extend our thanks.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Authors contribution**

Sam Polman contributed in Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing and Project administration.
Kirk Luther contributed in Formal Analysis, and Supervision, Writing – Review & Editing and Project administration. Hannah de Almeidar contributed in Formal Analysis and Writing – Review & Editing. Janina Eggers contributed in Methodology and Investigation. Steven J. Watson contributed in Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Data Curation and Data Curation.

**Data availability statement**

The data that support the findings of this study are openly available via the Open Science Framework at https://osf.io/jy7pw/. We also provide all our data collection materials (vignette, interview script) at this same location. This study was pre-registered and the (anonymized for peer review) pre-registration is available here: https://aspredicted.org/cn87k.pdf

**ORCID**

*Kirk Luther* http://orcid.org/0000-0002-2067-9888

**References**


