UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

ORIGINAL ARTICLE

PSYCHOPHYSIOLOGY SPR | WILEY

Countering information leakage in the Concealed Information Test: The effects of item detailedness

Linda Marjoleine Geven[1,2,3] 🔟 | Bruno Verschuere[1] | Merel Kindt[1] |
Shani Vaknine[2] | Gershon Ben-Shakhar[2]

[1]Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Psychology, Hebrew University of Jerusalem, Jerusalem, Israel

[3]Institute of Criminal Law and Criminology, Leiden University, Leiden, The Netherlands

Correspondence
Linda Marjoleine Geven, Institute of Criminal Law and Criminology, Leiden University, Postbus 9520, 2300 RA Leiden, The Netherlands.
Email: l.m.geven@law.leidenuniv.nl

Abstract

Concealed Information Tests (CIT) are administered to verify whether suspects recognize certain features from a crime. Whenever it is presumed that innocent suspects were contaminated with critical information (e.g., the perpetrator had a knife), the examiner may ask more detailed questions (e.g., specific types of knives) to prevent false positives. However, this may increase the number of false negatives if the true perpetrator fails to discern specific details from its plausible irrelevant controls, or because detailed crime-scene information may be forgotten. We examined whether presenting items at the exemplar level protects against contamination, and whether it compromises the sensitivity in a physiological CIT. Participants ($N = 142$) planned a mock-robbery, with critical items encoded either at the category or at the exemplar level. The CIT was administered immediately or after a 1-week-delay, with questions phrased at the categorical or exemplar level. There were no effects of time delay. Results revealed that when item detailedness was congruent at encoding and testing, the SCR, HR, and RLL showed larger differential responses, as compared with incongruent conditions. Participants contaminated with crime knowledge at the categorical level did not show a CIT-effect for crime details at the exemplar level, suggesting detailed questions may counter the leakage problem. Asking questions at the exemplar level did not reduce the CIT detection efficiency as compared to asking questions at the categorical level. The importance of congruency between encoding and testing provides examiners with a challenge, as it is difficult to estimate how details are naturally encoded.

KEYWORDS

content/topics, lie detection, content/topics, memory, content/topics, heart rate, deception detection, methods, methods, respiration, methods, skin conductance

1 | INTRODUCTION

The Concealed Information Test (CIT; initially called the Guilty Knowledge Test; Lykken, 1959; Verschuere et al., 2011) has been intensively investigated and laboratory research revealed that it is a highly valid method of memory detection (e.g., Meijer et al., 2014). It is designed to detect concealed knowledge rather than deception and is constructed as a multiple-choice test, with several questions, each having one critical, crime-related item (e.g., the perpetrator used a *knife*) intermixed with several equally plausible, yet incorrect irrelevant items (e.g., *gun, stick,* or *taser*). The CIT assesses whether the examinee recognizes the critical item, assumed to be known only to individuals involved (i.e., the perpetrator of a crime and the investigative team). When a suspect systematically shows stronger physiological responses to critical items than to the irrelevant items, labeled as the CIT-effect, knowledge is inferred. As innocent participants are unknowledgeable regarding the crime, they cannot tell the critical crime items from the irrelevant items, resulting in unsystematic responsivity across all items. Guilty suspects, on the contrary, show an increased skin conductance response (SCR), a deceleration of heart rate (HR) and respiratory suppression (usually defined as shortening of the total respiration line length; RLL) upon recognition of critical items. This pattern has been explained by an increased orienting response toward the recognized significant item and the deliberate attempt to inhibit the physiological arousal experienced when confronted with critical crime details (klein Selle et al. 2016, 2017, 2019).

Obviously, a successful implementation of the CIT depends on the underlying assumption that only guilty suspects have knowledge of the crime under investigation. Unfortunately, this assumption is often violated in real-life cases, where rumors, media, or police interviewing practices (Alceste et al., 2020; Garrett, 2015) may contaminate innocent suspects with intimate crime details (i.e., non-public details about the crime under investigation assumed to be known only to the perpetrator; Ofshe & Leo, 1997). The applicability of the CIT is restricted in such cases (Podlesny, 1993, 2003), as it hampers discernment between guilty and innocent suspects and increases the risk of false-positive outcomes. To illustrate, while participants who received verbal information about a mock-crime revealed lower SCRs in comparison to those who personally executed the theft (Meijer et al., 2010), a large difference remained compared with uninformed innocents. Findings of several other studies suggest the same pattern: recognition of information—whether it is through enacting the crime or through an innocent source—is sufficient to elicit a CIT-effect (see e.g., Bradley & Rettinger, 1992;

Bradley & Warfield, 1984; Gamer et al., 2010; Nahari & Ben-Shakhar, 2011).

This highlights the need to deal with the information leakage problem in the CIT. A potential solution has been proposed by field examiners from Japan (Osugi, 2011, 2014, 2018), the only country currently implementing the CIT on a regular basis. Whenever it is suspected that individuals have been exposed to critical information by other means than involvement in the crime, the examiner may probe for more detailed information. Rather than presenting items at a category level (e.g., the perpetrator used a *knife*), items are presented at the exemplar level (e.g., the perpetrator used a *switchblade*, in comparison to a *swiss knife, dagger,* or *machete*). Presenting items at the exemplar level provides an opportunity to counter the information leakage problem, by exonerating individuals contaminated with categorical information (i.e., preventing false-positives[1]).

When presenting items at the exemplar level, guilty suspects must remain able to distinguish the critical items from the irrelevant items. Yet, if the question is too detailed, it may result in non-recognition for guilty suspects (i.e., leading to false negatives). This issue has been referred to as the "distinguishability" of items in the CIT (Osugi, 2014, 2018), defined as the ease with which examinees can distinguish different answer options in a question. Only few studies have examined the possible detrimental effects of exemplar-level items. For example, Osugi (2014) reviewed field data from 30 criminal cases and found stronger physiological CIT-effects for high distinguishable items in comparison to low distinguishable items. More recently, relying on response times (RT) to indicate knowledge of critical items, Geven et al. (2019) found no differences in detection efficiency between questions phrased at the category (e.g., What type of weapon did the perpetrator use?) and exemplar-level questions (e.g., Which specific knife did the perpetrator use?). Yet, as most CIT examiners rely on autonomic measures as an indication of guilty knowledge, it is important to further investigate the influence of item detailedness at the encoding and testing phases on the detection accuracy of the autonomic CIT in a controlled design.

Reduced recognition of items at the exemplar level (leading to lower sensitivity), may be aggravated by the passage of time. Perpetrators may forget details from the crime after a time interval between encoding and

---

[1]While we consider detecting recognition in an informed innocent examinee to be a false positive, we realize that a different view point is possible. Regarding the CIT as a memory test, one could argue that correctly recognized items in the informed innocents can be viewed as a hit (see Ogawa et al., 2015). However, here we take the perspective adopted by most CIT researchers that the CIT serves to assess crime involvement.
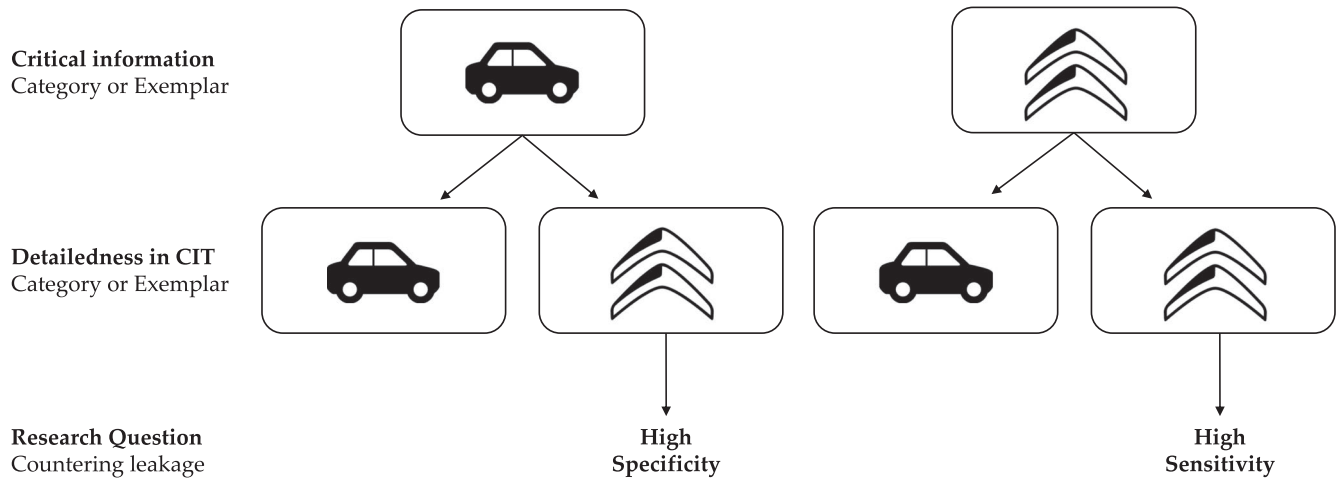
Critical information
Category or Exemplar

Detailedness in CIT
Category or Exemplar

Research Question
Countering leakage

High Specificity

High Sensitivity

FIGURE 1 Manipulated item detailedness at encoding and in the Concealed Information Tests

retrieval. Indeed, the detection efficiency in the CIT is usually diminished after a time delay, but mostly for less important, peripheral items (Gamer et al., 2010; Nahari & Ben-Shakhar, 2011; Peth et al., 2012). Memory research suggests that precise information is forgotten more rapidly than coarse information (see e.g., Christiaansen, 1980; Koriat et al., 2003). Following this basic-level convergence effect (Pansky & Koriat, 2004), it can be expected that exemplar-level memories are more likely to be stored at the categorical level after a delay. Critically, if perpetrators do not recognize highly detailed features from the crime after a time interval, the exemplar-level CIT may ultimately render ineffective to detect guilty knowledge.

To investigate the optimal item level in memory detection, the goal of the present study was twofold. First, we examined whether presenting items at the exemplar level may protect against false-positives expected to arise when categorical information is leaked to innocents. Second, we investigated to what extent the use of exemplar-level items reduces sensitivity in the CIT, particularly after a delay. Participants planned a mock-robbery, with the critical items encoded either at the category (e.g., you will flee the crime scene by *car*) or at the exemplar (e.g., you will flee the crime scene in a *Citroën*) level. Similarly, in the CIT items were presented either at the category or at the exemplar level in a crossed design (see Figure 1 for a visual depiction). To further disentangle the influence of memory deterioration over time, half of the participants were administered the CIT immediately, whereas the other participants completed the test after a 1-week delay.

2 | METHOD

The study was approved by the ethical committee of the Faculty of Social Sciences of the Hebrew University of

Jerusalem. This study was pre-registered: https://osf.io/z43md/. Task scripts, data, and other materials are publicly available on https://osf.io/gm4zj/. All participants provided consent before taking part in the study.

2.1 | Participants

A total of 142 participants (65.5% female)[2] were recruited for this study through a portal of the Hebrew University of Jerusalem. Their average age was 23.80 years ($SD = 2.66$, range from 18 to 34 years). Participants received course credits or a monetary compensation (40 ILS, equivalent to €10). Two participants were excluded from this initial sample: One participant received the wrong version of the crime scenario in the second session, and another participant completed the CIT after an interval of only 5 days. Seventy-one participants (69.0% female; $M_{age} = 24.03$, $SD_{age} = 2.93$) were randomly assigned to the immediate condition, and 69 participants (62.3% female; $M_{age} = 23.55$, $SD_{age} = 2.35$) were randomly assigned to the delayed condition, completing the CIT after a one-week interval (± 1 day). There were no significant differences between the immediate and the delayed conditions in age, $t(138) = 1.06$, $p = .290$, $d = 0.18$, 95% CI [−0.15;0.51] or gender, $X^2(1) = 0.70$, $p = .404$, $\varphi_c = 0.07$.

---

[2]As G*Power does not allow for power calculation of the within-between mixed ANOVA design, a power analysis was done for the follow-up *t*-tests. Since these tests require more participants in general, this will automatically result in sufficient power for the main ANOVA. Calculating for a medium effect size of d = 0.5, with a power of 0.80, the required participant number is 128. Since it was expected that around 10% of the participants were to be excluded, or participants would not complete the study in the delayed condition, a total sample size of 141 participants was required.

2.2 | Procedure

2.2.1 | Encoding

Similar to the procedure described in Geven et al. (2019), participants came to the laboratory in pairs and studied the plan of a bank robbery together. When only one participant showed up ($n = 9$), the experimenter took the place of the partner in crime while planning the mock crime. Participants planned a mock robbery with the other participant of the pair. The crime scenario consisted of a coherent story featuring eight critical items, of which four were presented at a categorical level and four at an exemplar form: Participants encoded a scenario according to which they met each other in a *sports center* (exemplar: *swimming pool*) and planned to rob a *bank* (exemplar: *Hapoalim bank*) in their residence area *Gush Dan* (exemplar: *Hulon*) in *May* (exemplar: *May 26th*). Because they might not be able to flee the scene without a fight, they would bring a *pointed weapon* (exemplar: *flick knife*). The partners in crime plan to steal expensive *jewelry* (exemplar: *ring*) and hide it at *home* (exemplar: *attic*). Lastly, they planned to flee the crime scene *by car* (exemplar: *Citroën*). Two versions of the crime scenario were randomized between participant pairs, such that the items varied in Item Type at encoding (category vs. exemplar).

The experimenter instructed participants to study the items from the crime extensively and to visualize committing the robbery. The experimenter first read the plan for the robbery out loud, with its eight critical items presented in either categorical or exemplar level. Participants then wrote down the words, read the plan out loud, and probed each other for the information. During this encoding phase, the experimenter stayed in the room to intervene when participants would accidentally fill in an exemplar-level item when aligning their story (e.g., inventing a specific car brand if they had to encode the categorical stimulus). Finally, participants completed a gapped text of the story on a paper from which the critical items had been removed, followed by a free recall of the items. After this encoding phase, one participant of the pair was randomly assigned to the immediate testing condition, whereas the other participant was assigned to complete the CIT after a 1-week delay ($\pm$1 day). Participants were explicitly instructed not to discuss details of the experiment with each other in the one week between the encoding phase and the second session.

2.2.2 | CIT

In the second phase of the experiment, a second experimenter requested participants to wash their hands in preparation for the physiological task, attached the RLL belts as well as the SCR and HR electrodes, and conducted the CIT procedure. The experimenter explained that the participants will take a polygraph test. Participants were instructed to hide all information about the planned crime and aim for a test result indicating innocence. Upon successful concealment, the participant would receive an additional course credit compensation or an equivalent monetary bonus (10 ILS, the equivalent of €2.50).

For each of the eight critical items encoded in the crime scenario, the CIT included the correct answer, a buffer item, and four incorrect answers serving as irrelevant options (ratio 1:1:4). For instance, if *car* was the critical stimulus, the buffer item was *bicycle*, and the irrelevant stimuli were *motorbike*, *train*, *bus*, and *helicopter* (categorical Item Type), and if *Citroën* was the critical stimulus, the buffer was *Volkswagen*, and the irrelevant stimuli were *Opel*, *Fiat*, *Ford*, and *Peugeot* (exemplar Item Type). In total, participants were presented with eight questions, each consisting of seven items (i.e., one buffer, one critical, four irrelevant, and one catch item), totaling 56 items. The order of the eight questions and their answering alternatives were randomly determined, with a short break between two blocks, each containing four questions, to maintain participant's attention.

Items were presented either at the categorical or exemplar level: two stimuli were encoded at the category level and were also presented at the category level (e.g., encoded as *car*, tested as *car,* congruent with encoding*)* and two stimuli were encoded at the exemplar level and were also presented at the exemplar level (e.g., encoded as *Citroën*, tested as *Citroën,* congruent with encoding). In two other instances, the stimuli encoded at the category level were replaced by the corresponding test stimulus in its exemplar form (e.g., encoded as *car*, tested as *Citroën;* incongruent with encoding, as no exemplar-level information was made available at encoding) and two of the stimuli encoded on the exemplar level were replaced by the corresponding test stimulus in its category form (e.g., encoded as *Citroën*, tested as *car*; incongruent with encoding, as only exemplar-level information was made available at encoding). This randomization resulted in four versions of the CIT, such that every two items were varied in encoding (category vs. exemplar) and testing (category vs. exemplar) between the scripts (see also Figure 1).

During the CIT all participants were explicitly instructed to conceal their knowledge of the planned robbery and respond with a verbal "no" to all presented words. Lastly, catch items (i.e., random numbers) were inserted to ensure that examinees paid attention to all items. Participants were instructed to say the numbers out loud as soon as they appeared on the screen.

All questions and alternatives were presented on the computer screen in written words. Simultaneously, participants heard the spoken version through their headphones. Audio files (with a mean duration of 3 s for the questions, and 1 s for the items) were pre-recorded by a third party who was blind to the procedure. The question remained on the screen for 10 s, followed by the answering alternatives that were presented for 5 s each, with a mean inter-stimulus interval of 18 s (range 16–20). The first answering alternative following the question was always a buffer-item, designed to absorb the initial OR, followed by the critical item, 4 irrelevant items, and a catch item in random order.

After the CIT procedure, participants completed the follow-up questionnaire. Participants were asked to report, on 5-point Likert scales, on their ability to focus on the computer screen during the CIT, general involvement in the experiment, memory for the planned robbery, and their effort to conceal knowledge by suppressing or enhancing physiological responding. Finally, participants were allowed to freely elaborate on strategies to avoid detection.

Then, participants were told that they did not have to hide any information anymore. To investigate the extent of forgetting during the time interval between encoding and the CIT, participants completed a free recall and then a recognition memory test. For free recall, the number of correctly recalled items were counted, leading to a total recall score between 0 and 8. For the multiple-choice recognition task, participants had to pick the critical item studied during the encoding phase from four irrelevant options. For each question, correct identification of the item resulted in a score of 1, leading to a total recognition score between 0 and 8. Finally, participants were thanked and debriefed.

2.3 | Data acquisition and reduction

Electrodermal activity was recorded using a constant voltage system (0.5 V ASR Atlas Researches, Hod Hasharon, Israel) and an A/D (NB-MIO-12) converter with a sampling rate of 50 Hz. Two Ag/AgCl electrodes (0.8 cm diameter) filled with a 0.05 M NaCl electrolyte (TD-246, Discount Disposables) were placed on the distal phalanges of the left index and left ring finger. The SCR was measured from 1 to 5 s after stimulus onset and defined as the maximal increase in conductance during this time window.

The ECG measure was acquired by placing three Ag/AgCl electrodes filled with an electrode paste in a standard Einthoven lead I configuration: one electrode attached to the distal phalange of the left index finger (i.e., one of the SCR electrodes), one electrode attached to the right wrist

and the ground electrode attached to the left wrist. The ECG signal was sampled at 500 Hz, digitized at 12-bit resolution, and filtered using a bandpass of 1–35 Hz. Before analysis, the inter-beat intervals were converted to HR in beats per minute (bpm) per real-time epoch (1 s). The second-by-second post-stimulus HR values were baseline-corrected by subtracting the average HR value in the 3 s preceding stimulus onset (i.e., the pre-stimulus baseline value), resulting in 15 post-stimulus difference scores (ΔHR). The average of these 15 scores was used as the HR deceleration dependent measure.

Respiration was recorded with a respiratory band positioned around the thoracic area. Respiration responses were defined based on the total RLL, which is a composite measure of respiratory amplitude (depth of breathing) and respiratory cycle (rate of breathing), during a 15-s interval following stimulus onset. Following Elaad et al. (1992), we defined each response as the mean of ten length measures (0.1 s after stimulus onset through 15.1 s after stimulus onset, 0.2 s through 15.2 s after stimulus onset, etc.). The RLL was defined as the mean of the ten length measures computed for the ten windows.

CIT-scores were calculated as individual Z-scores, reflecting relative responses to the critical item compared to the irrelevant items, computed within each block, separately for each of the three physiological measures. These CIT-scores are calculated by subtracting the mean response across all items from the response to the critical items, divided by the respective standard deviation (see Ben-Shakhar & Elaad, 2002; Elaad & Ben-Shakhar, 1997). Buffer and catch items are excluded from the standardization procedure (see klein Selle et al., 2016, 2017). For each participant and each physiological measure, a CIT-score was created by averaging the respective Z-scores of all critical items. Scores for HR and RLL are multiplied by −1 prior to analysis, hence a positive CIT-score is indicative of concealed information for all three measures.

2.3.1 | Exclusion criteria

On the participant level, individuals who showed a standard deviation of the raw SCR scores below 0.01 during the entire procedure ($n = 4$) were considered to be skin conductance non-responders and their data were eliminated from all SCR analyses.

For each of the dependent measures, item-specific responses were removed if the standardized score was smaller than −5 or larger than 5, reflecting outliers. When an excessive movement coincided with a positive standardized score (for SCR), the item was discarded from analyses (see also Geven et al., 2018). In the current dataset, a total of 15 item-specific SCR responses were removed ($n = 14$).

Further exclusions were performed when participants showed a standard deviation of the raw SCR scores below 0.01 throughout the presentation of a block (i.e., 4 questions). In these cases, all SCR measurements from that block were discarded from further analyses due to non-responsiveness ($n = 13$).

2.4 | Deviations from the preregistration

We recruited 142 participants instead of the targeted and preregistered 141 participants, since participation was in pairs.

Two participants were excluded from the original sample. One participant received the wrong version of the crime scenario in the second session, and another participant completed the CIT after an interval of only 5 days. We did not account for these exclusions and thereby deviate from the preregistered exclusion criteria.

Three individuals were excluded from SCR analyses due to technical errors with the electrodes. Data from the HR measure were eliminated from analyses based on the occurrence of frequent extrasystoles ($n = 5$) and when technical errors occurred ($n = 12$). RLL data were excluded due to technical errors ($n = 2$).

When an excessive movement coincided with a positive standardized score larger than 2 or lower than -2 (for HR and RLL), the item was discarded from analyses (see also Geven et al., 2018).

No ROC curves were computed, as the final design did not include an innocent condition. While an innocent condition could be simulated, this was not considered worthwhile for the current research question.

JZS Bayes factors were not evaluated using Jeffreys' (1961) criteria. As such criteria are arbitrary, we refrain from referring to any specified benchmark, neither for effect sizes nor for Bayes factors.

3 | RESULTS

All analyses used an alpha level of 0.05. Effect sizes for the ANOVA are reported using Cohen's f. Effect sizes for the independent samples t-tests are reported using Cohen's d. In addition, JZS Bayes factors (BF) were computed using JASP software version 0.8.4 (JASP Team, 2018), representing numerical values quantifying the odds ratio between the null and the alternative hypothesis given the data. BF_{01} annotates how much more likely the null hypothesis is, compared to the alternative hypothesis, given the data, and BF_{10} annotates how much more likely the alternative hypothesis is compared to the null hypothesis, given the data. For one-tailed testing, Bayes factors are reported as

either predicting the null (BF_{0+}) or the alternative hypothesis (BF_{+0}). It should be noted that values close to 1 fail to support either hypothesis. JZS prior with scaling factor $r = 0.707$ was used for the alternative hypothesis (see Rouder et al., 2009).

3.1 | Confirmatory analyses

3.1.1 | SCR

The main analysis consisted of a 2 (Delay: immediate vs. 1-week-delayed CIT, between-participants) by 2 (Item detailedness at encoding: category level vs. exemplar level, within-participants) by 2 (Item detailedness in the CIT: category level vs. exemplar level, within-participants) mixed ANOVA on the Z-scores of the critical items.

The mixed ANOVA revealed a significant main effect of Item detailedness at encoding, $F(1, 119) = 21.60$, $p < .001$, $f = 0.42$, and a significant main effect of Item detailedness in the CIT, $F(1, 119) = 8.15$, $p = .005$, $f = 0.26$, that subsumed under the significant interaction between these factors, $F(1, 119) = 58.19$, $p < .001$, $f = 0.69$. This interaction originates from the larger Z-scores when the item detailedness was congruent at encoding and the CIT (i.e., both exemplar or both categorical) compared to items for which there was an incongruency in detailedness between encoding and the CIT (i.e., encoded at exemplar level and tested categorically, or encoded at the categorical level and tested at the exemplar level).

There was no main effect of Delay, $F(1, 119) = 0.75$, $p = .389$, $f = 0.08$, and no interaction between Item detailedness at encoding and Delay, $F(1, 119) = 1.11$, $p = .293$, $f = 0.09$, or between Item detailedness in the CIT and Delay, $F(1, 119) = 1.06$, $p = .306$, $f = 0.09$. Also, the three-way interaction was not statistically significant, $F(1, 119) = 2.53$, $p = .114$, $f = 0.12$.

To narrow down the interaction between Item detailedness at encoding and Item detailedness in the CIT, planned contrasts were conducted with Item Type as fixed factors (i.e., Category[encoding]-Category[CIT], Category[encoding]-Exemplar[CIT], Exemplar[encoding]-Category[CIT], and Exemplar[encoding]-Exemplar[CIT]). As no significant influence of a time delay on CIT performance emerged, the reported contrasts were conducted across the immediate and delayed CIT conditions.

A first planned contrast compared the mean Z-score of the Category-Exemplar Item Type with the three other Item Types to test the hypothesis that participants with categorical information only do not show recognition of the exemplar-level stimuli. The one-tailed paired-samples t-test revealed a significantly lower mean Z-score in the

Category-Exemplar Item Type ($M = -0.19$, $SD = 0.64$) compared to the three other Item Types in which knowledge existed (i.e., Category-Category, Exemplar-Category, and Exemplar-Exemplar; $M = 0.66$, $SD = 0.97$, $t(128) = -10.76$, $p < .001$, $d = -0.95$). A Bayesian one-tailed paired-samples t-test revealed evidence for the alternative hypothesis ($BF_{-0} = 7.94e+16$).

A second planned contrast examined the congruency effect by comparing the mean Z-score of the Exemplar-Category Item Type with the two Item Types in which the item detailedness was congruent between encoding and the CIT. The one-tailed paired-samples t-test revealed a significantly lower CIT-effect in the Exemplar-Category Item Type ($M = 0.41$, $SD = 0.89$) than for the other two Item Types (i.e., Category-Category and Exemplar-Exemplar; $M = 0.78$, $SD = 0.81$, $t(129) = -3.16$, $p < .001$, $d = -0.27$). A Bayesian one-tailed paired-samples t-test revealed evidence for the alternative hypothesis ($BF_{-0} = 21.85$), suggesting higher detection efficiency when the detailedness in the CIT was congruent with encoding. Note that the Category-Exemplar Item Type was not included in these contrasts since participants are not expected to have a larger response to the critical compared to irrelevant items in this condition.

An additional comparison was performed on the Category-Category versus Exemplar-Exemplar Item Types, to examine whether questions in the CIT are best asked on category or the exemplar level, given congruency between encoding and the CIT. The two-tailed paired-samples t-test revealed that the mean Z-score for the Category-Category ($M = 0.65$, $SD = 0.98$) and the Exemplar-Exemplar Item Type ($M = 0.87$, $SD = 1.03$) did not significantly differ, $t(124) = -1.68$, $p = .097$, $d = -0.15$. A Bayesian two-tailed paired-samples t-test revealed evidence for the null hypothesis ($BF_{01} = 2.59$), suggesting that the highly detailed CIT does not hamper detection efficiency.

Moreover, a one-sample Bayesian t-test was performed on the mean Z-scores in each condition to investigate whether detection efficiency was above chance. Table 1 shows the mean scores for each cell of the design. For both

the immediate and delayed conditions, the results revealed the expected CIT-effect for the Item Types Category-Category and Exemplar-Exemplar, reflected by evidence that recognition of the critical item results in large Z-scores. For the Item Type Exemplar-Category there was also a significant CIT-effect, reflected by evidence for the alternative hypothesis showing generalization from the exemplar to the category level. For the Category-Exemplar Item Type, participants encoded the critical information at the categorical level and hence were not expected to distinguish the critical item from the irrelevant items in the exemplar-level CIT. Two-tailed Bayesian analysis revealed evidence for the alternative hypothesis in the immediate condition, yet the mean Z-score was negative and therefore not indicating recognition. For the delay condition, Bayesian statistics revealed evidence for the null hypothesis. Figure 2 shows the mean raw SCR scores for each Item Type across conditions.

3.1.2 | HR

The same $2 \times 2 \times 2$ ANOVA described above for the SCR was conducted on the HR Z-scores. The mixed ANOVA revealed a significant interaction between Item detailedness at encoding and Item detailedness in the CIT, $F(1, 120) = 9.77$, $p = .002$, $f = 0.16$. This interaction originates from the larger Z-scores when encoding and the CIT were on a congruent level compared to incongruent Item Types.

No significant main effect was found for Delay, $F(1, 120) = 0.38$, $p = .539$, $f = 0.03$. Additionally, there were no significant main effects of Item detailedness at encoding, $F(1, 120) = 3.08$, $p = .082$, $f = 0.07$, Item detailedness in the CIT, $F(1, 120) = 0.51$, $p = .472$, $f = 0.03$, and no interaction between Item detailedness at encoding and Delay, $F(1, 120) = 0.44$, $p = .508$, $f = 0.03$, and Item detailedness in the CIT and Delay, $F(1, 120) = 0.03$, $p = .868$, $f = 0.01$. Lastly, the three-way interaction did not reveal a significant effect, $F(1, 120) = 0.55$, $p = .462$, $f = 0.03$.

TABLE 1 Skin conductance Z-scores of the critical items per condition and Item Type

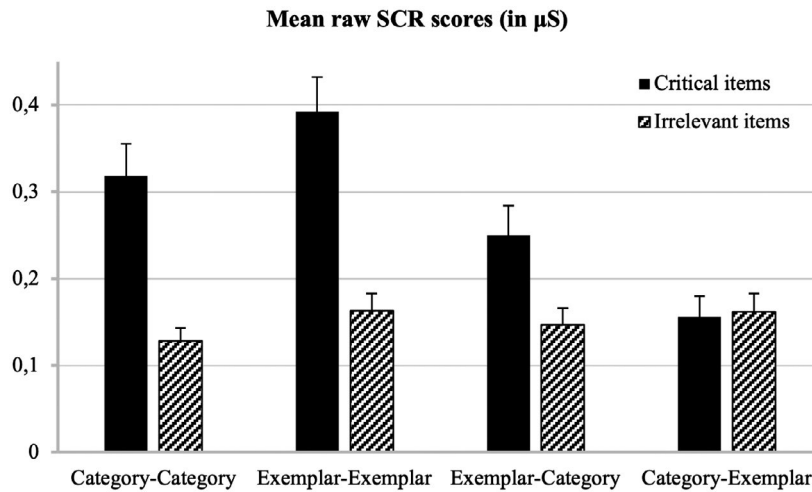| Item Type	Immediate condition			Delayed condition		
	M (SD)	d (95% CI)	BF	M (SD)	d (95% CI)	BF
Category Category	0.82 (0.96)	0.85 [0.61; ∞]	$BF_{+0} = 1.19e+7$	0.47 (0.97)	0.48 [0.26; ∞]	$BF_{+0} = 130.66$
Exemplar Exemplar	0.92 (1.12)	0.83 [0.59; ∞]	$BF_{+0} = 7.47e+6$	0.80 (0.94)	0.86 [0.61; ∞]	$BF_{+0} = 2.88e+6$
Exemplar Category	0.38 (0.92)	0.41 [0.20; ∞]	$BF_{+0} = 43.68$	0.45 (0.85)	0.53 [0.30; ∞]	$BF_{+0} = 400.32$
Category Exemplar	−0.21 (0.67)	−0.32 [−0.56; −0.07]	$BF_{10} = 3.11$	−0.16 (0.62)	−0.26 [−0.52; −0.01]	$BF_{01} = 1.00$

FIGURE 2 Mean raw skin conductance response scores (in μS) for all Item Types across condition

TABLE 2 Heart rate Z-scores of the critical items per condition and Item Type

Item Type	Immediate condition			Delayed condition		
	M (SD)	*d* (95% CI)	BF	*M (SD)*	*d* (95% CI)	BF
Category Category	0.25 (0.51)	0.48 [0.26; ∞]	$BF_{+0} = 139.48$	0.28 (0.51)	0.54 [0.31; ∞]	$BF_{+0} = 479.41$
Exemplar Exemplar	0.31 (0.42)	0.74 [0.50; ∞]	$BF_{+0} = 114,721.26$	0.30 (0.59)	0.51 [0.28; ∞]	$BF_{+0} = 219.72$
Exemplar Category	0.23 (0.48)	0.47 [0.25; ∞]	$BF_{+0} = 96.98$	0.13 (0.50)	0.25 [0.04; ∞]	$BF_{+0} = 1.60$
Category Exemplar	0.09 (0.46)	0.19 [−0.07; 0.44]	$BF_{01} = 2.60$	0.06 (0.49)	0.12 [−0.14; 0.37]	$BF_{01} = 4.85$

To narrow down the predicted interaction between Item detailedness at encoding and Item detailedness in the CIT, the same planned contrasts described above for the SCR were conducted for the HR. A first planned contrast compared the mean Z-score between the Category-Exemplar Item Type and the three other Item Types. The one-tailed paired-samples *t*-test revealed a significantly smaller Z-score in the Category-Exemplar Item Type ($M = 0.07$, $SD = 0.47$) compared to the three other Item Types in which knowledge existed (i.e., Category-Category, Exemplar-Category, and Exemplar-Exemplar; $M = 0.25$, $SD = 0.30$, $t(121) = -3.20$, $p < .001$, $d = -0.29$). A Bayesian one-tailed paired-samples *t*-test revealed evidence for the alternative hypothesis ($BF_{-0} = 24.91$).

A second planned contrast compared the mean Z-score of the Exemplar-Category Item Type with the two Item Types in which the abstractness level was congruent for encoding and CIT. The one-tailed paired-samples *t*-test revealed that the CIT-effect was significantly lower in the Exemplar-Category Item Type ($M = 0.18$, $SD = 0.50$) than for the other two Item Types (i.e., Category-Category and Exemplar-Exemplar; $M = 0.28$, $SD = 0.39$, $t(121) = -1.83$, $p = .035$, $d = -0.17$). A Bayesian one-tailed paired-samples *t*-test revealed inconclusive evidence for the alternative hypothesis ($BF_{-0} = 0.97$). Note that the Category-Exemplar Item Type was not included in these

contrasts since participants are not expected to have larger responses to the critical compared to irrelevant items.

An additional comparison was performed on the Category-Category versus Exemplar-Exemplar Item Types, to examine whether items in the CIT are best presented on category or the exemplar level given congruency between encoding and the CIT. The two-tailed paired-samples *t*-test revealed that the mean Z-score for the Category-Category ($M = 0.26$, $SD = 0.51$) and the Exemplar-Exemplar Item Type ($M = 0.30$, $SD = 0.51$) did not significantly differ, $t(121) = -0.72$, $p = .471$, $d = -0.07$. A Bayesian two-tailed paired-samples *t*-test revealed evidence for the null hypothesis ($BF_{01} = 7.71$), suggesting similar detection efficiency in the exemplar-level as in the categorical-level CIT. This implies that the highly detailed CIT may reduce false-positives due to information contamination, with no increase in false-negatives.

Moreover, a one-sample Bayesian *t*-test was performed on the mean Z-scores in each condition to investigate whether detection efficiency was above chance. Table 2 shows the mean scores for each cell of the design. For both the immediate and delayed conditions, the results revealed the expected CIT-effect for the Item Types Category-Category and Exemplar-Exemplar, reflected by evidence that recognition of the critical item results in positive Z-scores. For the

Item Type Exemplar-Category there was also a CIT-effect, reflected by evidence for the alternative hypothesis showing generalization from the exemplar to the category level, albeit less strong in the one-week delayed condition. For the Category-Exemplar Item Type, participants encoded the critical information at the categorical level and hence were not expected to distinguish the correct exemplar from the irrelevant exemplars in the CIT. Two-tailed Bayesian analysis revealed evidence for the null hypothesis in both conditions. Figure 3 shows the mean second-by-second ΔHR scores for each Item Type across condition.

3.1.3 | RLL

The same 2 × 2 × 2 ANOVA described above for the SCR and HR was conducted on the RLL Z-scores. The mixed ANOVA revealed a significant interaction between Item detailedness at encoding and Item detailedness in the CIT, $F(1, 135) = 16.14, p < .001, f = 0.35$. This interaction originates from the larger Z-scores when encoding and the CIT were on a congruent level compared to incongruent Item Types. Moreover, a significant interaction was revealed between Item detailedness at encoding and Delay, $F(1, 135) = 5.44, p = .021, f = 0.20$.

No significant main effect was found for Delay, $F(1, 135) = 1.55, p = .216, f = 0.11$. Additionally, there were no significant main effects of Item detailedness at encoding,

$F(1, 135) = 0.11, p = .746, f = 0.03$, Item detailedness in the CIT, $F(1, 135) = 1.87, p = .174, f = 0.12$, and no interaction between Item detailedness in the CIT and Delay, $F(1, 135) = 0.57, p = .452, f = 0.06$. Lastly, the three-way interaction did not reveal a significant effect, $F(1, 135) = 0.05, p = .824, f = 0.00$.

To narrow down the predicted interaction between Item detailedness at encoding and Item detailedness in the CIT, the same planned contrasts described above for the SCR and HR were conducted for the RLL. A first planned contrast compared the mean Z-score between the Category-Exemplar Item Type and the three other Item Types. The one-tailed paired-samples t-test revealed a significantly smaller Z-score in the Category-Exemplar Item Type ($M = 0.10, SD = 0.69$) compared to the three other Item Types in which knowledge existed (i.e., Category-Category, Exemplar-Category, and Exemplar-Exemplar; $M = 0.34, SD = 0.44, t(136) = -3.28, p < .001, d = -0.28$). A Bayesian one-tailed paired-samples t-test revealed evidence for the alternative hypothesis (BF$_{-0}$ = 30.38).

A second planned contrast compared the mean Z-score of the Exemplar-Category Item Type with the two Item Types in which the abstractness level was congruent for encoding and in the CIT. The one-tailed paired-samples t-test revealed that the CIT-effect was significantly lower in the Exemplar-Category Item Type ($M = 0.20, SD = 0.69$) than for the other two Item Types (i.e., Category-Category and Exemplar-Exemplar;
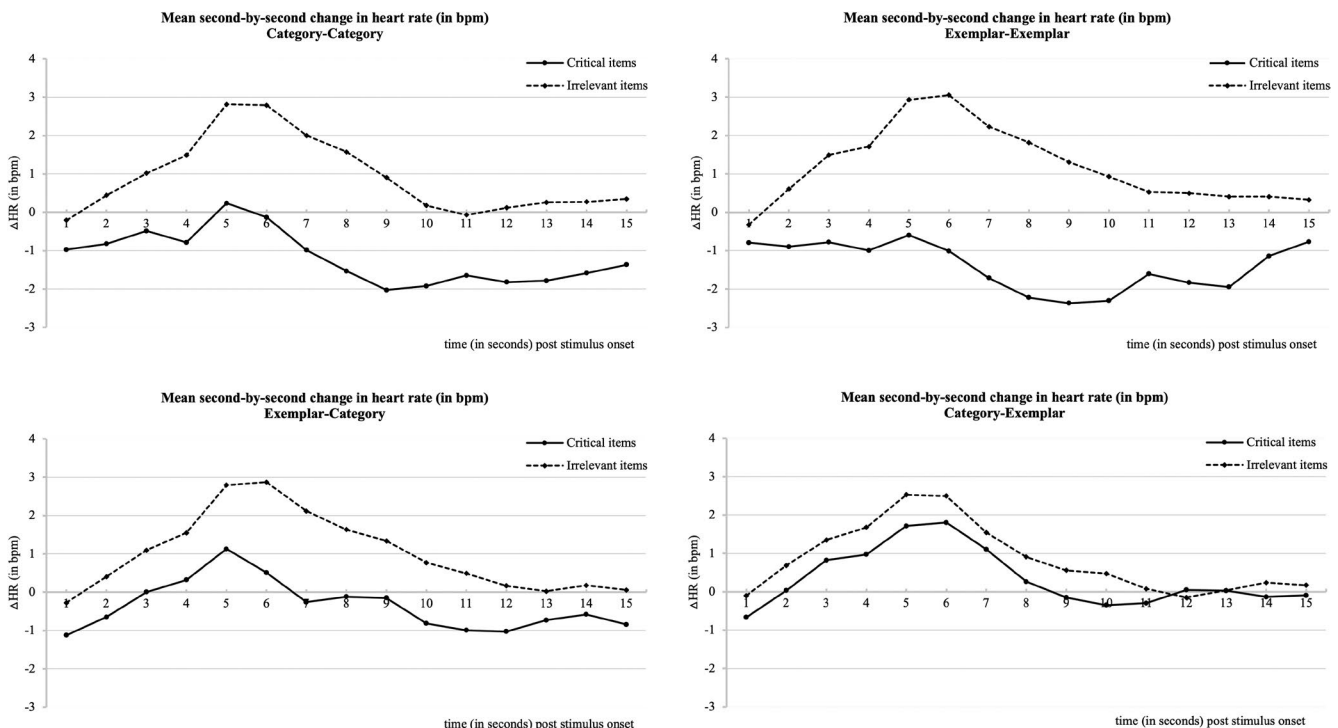

FIGURE 3 Mean second-by-second ΔHR scores (in bpm) for each Item Type across condition

$M = 0.42$, $SD = 0.55$, $t(136) = -2.94$, $p = .002$, $d = -0.25$). A Bayesian one-tailed paired-samples t-test revealed evidence for the alternative hypothesis ($BF_{-0} = 11.45$), highlighting the importance of congruency between encoding and testing. Note that the Category-Exemplar Item Type was not included in these contrasts since participants are not expected to have larger responses to the critical compared to irrelevant items.

An additional comparison was performed on the Category-Category versus Exemplar-Exemplar Item Types, to examine whether items in the CIT are best presented on category or the exemplar level given congruency between encoding and the CIT. The two-tailed paired-samples t-test revealed that the mean Z-score for the Category-Category ($M = 0.45$, $SD = 0.74$) and the Exemplar-Exemplar Item Type ($M = 0.38$, $SD = 0.71$) did not significantly differ, $t(136) = 0.83$, $p = .408$, $d = 0.07$. A Bayesian two-tailed paired-samples t-test revealed evidence for the null hypothesis ($BF_{01} = 7.51$), suggesting similar detection efficiency in the exemplar-level as in the categorical-level CIT. This implies that the highly detailed CIT may reduce false-positives due to information contamination, with no increase in false-negatives.

Moreover, a one-sample Bayesian t-test was performed on the mean Z-scores in each condition to investigate whether detection efficiency was above chance. Table 3 shows the mean scores for each cell of the design. For both the immediate and delayed conditions, the results revealed the expected CIT-effect for the Item Types Category-Category and Exemplar-Exemplar, reflected by evidence that recognition of the critical item results in positive Z-scores. For the Item Type Exemplar-Category there was also a CIT-effect, reflected by evidence for the alternative hypothesis showing generalization from the exemplar to the category level. For the Category-Exemplar Item Type, participants encoded the critical information at the categorical level and hence were not expected to distinguish the correct exemplar from the irrelevant exemplars in the CIT. Two-tailed Bayesian analysis revealed evidence for the null hypothesis in the one-week delayed condition, while Bayesian analysis was inconclusive for the

immediate condition. Figure 4 shows the mean raw RLL scores for each Item Type across condition and Figure 5 shows the mean Z-scores of the critical items across condition for each physiological measure.

3.2 | Exploratory analyses

3.2.1 | Self-report ratings

Independent-samples t-tests revealed that participants in the immediate condition reported having a better memory for items of the crime scenario than participants who were tested after a 1-week delay. Participants scored high on their reported focus, involvement, and effort to hide the critical information, with no significant differences between conditions. Table 4 shows the mean scores for each cell of the design.

Most participants reported imagining other items to be presented as to distract their thoughts (44%) as a countermeasure strategy. Twenty-nine participants reported not having a strategy (20%) or tried to remain calm throughout the experiment (12%). Fourteen participants reported to deliberately respond equally to all items (10%) and 20 participants (14%) mentioned increasing their responses specifically to the irrelevant items (evoking physiological arousal, biting their tongue), albeit sometimes possibly evoking an even stronger CIT-effect (i.e., slowing down breathing upon presentation of the critical item).

3.2.2 | Memory

A 2 (Delay: immediate vs. 1-week-delayed CIT, between-participants) by 2 (Item detailedness at encoding: category level vs. exemplar level, within-participants) ANOVA was performed on the recall and recognition scores separately. Table 5 shows the mean recall and recognition scores.

For recall, the mixed ANOVA revealed a significant effect of Delay, $F(1, 137) = 33.97$, $p < .001$, $f = 0.39$. There

TABLE 3 Respiration Z-scores of the critical items per condition and Item Type

Item type	Immediate condition			Delayed condition		
	M (SD)	d (95% CI)	BF	M (SD)	d (95% CI)	BF
Category Category	0.58 (0.65)	0.89 [0.65; ∞]	$BF_{+0} = 9.57e+7$	0.31 (0.80)	0.39 [0.18; ∞]	$BF_{+0} = 27.46$
Exemplar Exemplar	0.34 (0.73)	0.47 [0.26; ∞]	$BF_{+0} = 206.70$	0.42 (0.68)	0.62 [0.40; ∞]	$BF_{+0} = 10,219.85$
Exemplar Category	0.19 (0.61)	0.31 [0.11; ∞]	$BF_{+0} = 5.72$	0.21 (0.76)	0.28 [0.07; ∞]	$BF_{+0} = 2.79$
Category Exemplar	0.17 (0.70)	0.25 [0.01; 0.49]	$BF_{01} = 1.02$	0.03 (0.68)	0.04 [−0.20; 0.28]	$BF_{01} = 7.03$

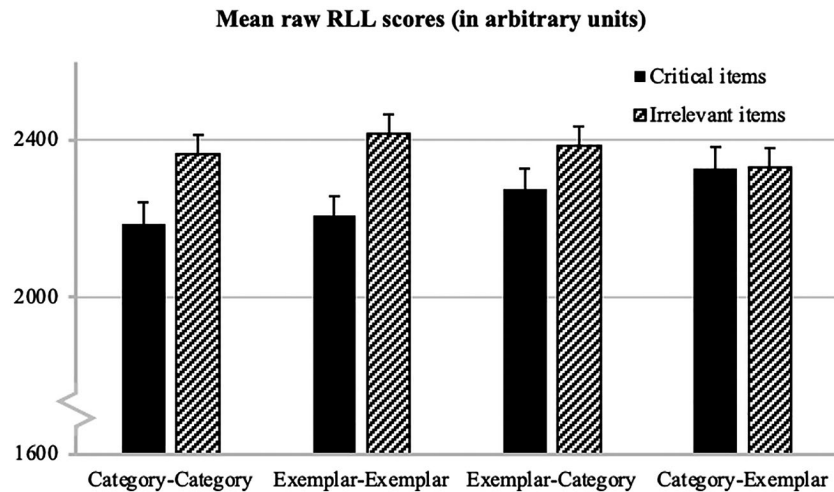**FIGURE 4** Mean raw respiration line length scores (in arbitrary units) for each Item Type across condition

Mean raw RLL scores (in arbitrary units)

FIGURE 5 Mean Z-scores of the critical items across conditions for each physiological measure

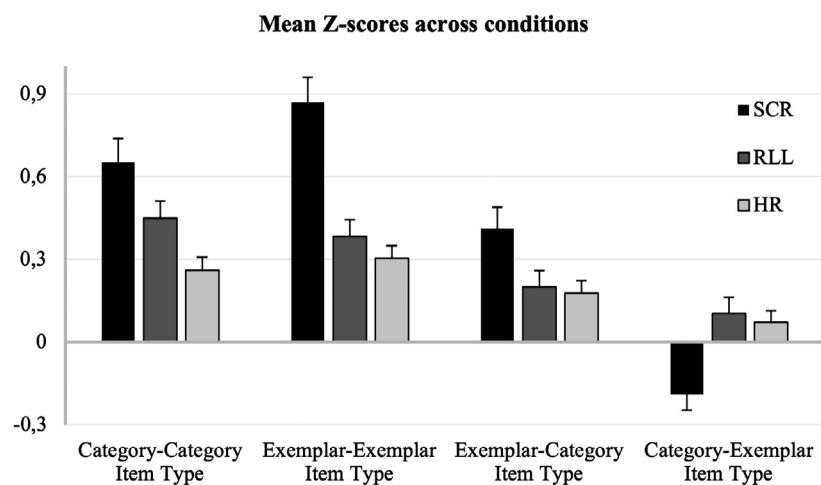Mean Z-scores across conditions

TABLE 4 Mean scores on the follow-up questionnaire (5-point Likert scale)

Question	M (SD)	M (SD)	t	df	p	$d_{between}$ (95% CI)	BF
	Immediate condition	Delayed condition					
Focus	3.87 (0.76)	3.90 (0.81)	0.20	137	0.839	−0.04 [−0.37; 0.30]	$BF_{01} = 5.39$
Involvement	4.33 (0.74)	4.52 (0.70)	1.59	137	0.115	−0.27 [−0.60; 0.07]	$BF_{01} = 1.75$
Memory for robbery	4.89 (0.36)	4.55 (0.61)	3.94	110.08	<0.001	0.67 [0.32; 1.01]	$BF_{10} = 178.86$
Effort to conceal knowledge	3.93 (1.00)	3.93 (0.90)	0.01	137	0.995	0.00 [−0.33; 0.33]	$BF_{01} = 5.50$

was no main effect of Item detailedness at encoding, $F(1, 137) = 1.13$, $p = .290$, $f = 0.05$, and no interaction between Item detailedness in the CIT and Delay, $F(1, 137) = 0.34$, $p = .560$, $f = 0.00$. To follow up, one-tailed independent-samples t-tests were conducted to evaluate whether the memory of participants in the immediate condition was significantly higher than in the delayed condition. Participants in the immediate condition generally

TABLE 5 Mean scores (*SD*s in parentheses) on memory recall and recognition (range 0–8)

Question	*M (SD)* Immediate condition	*M (SD)* Delayed condition
Recall	7.71 (0.57)	6.86 (1.17)
Recognition	7.93 (0.26)	7.88 (0.37)

recalled more items than participants in the delayed condition, $t(97.06) = 5.86$, $p < 001$, $d = 0.99$.

For recognition, the mixed ANOVA revealed no significant effect of Delay, $F(1, 138) = 0.73$, $p = .395$, $f = 0.05$, no main effect of Item detailedness at encoding, $F(1, 138) = 2.25$, $p = .136$, $f = 0.08$, and no interaction between Item detailedness in the CIT and Delay, $F(1, 138) = 2.25$, $p = .136$, $f = 0.08$. In conclusion, no differences emerged on recognition scores (note they were close to the maximum score of 8 in both conditions).

4 | DISCUSSION

The CIT, administered to verify whether suspects recognize critical details from the crime-scene, is used on a daily basis in criminal investigations in Japan (Osugi, 2011). Laboratory studies have validated the use of memory detection using psychophysiological measures (see Meijer et al., 2014), yet the assumption that *only* guilty suspects possess crime-related information may be violated in real-life applications. If innocents are inadvertently contaminated with perpetrator knowledge, the diagnostic value of the CIT could be compromised (Bradley & Rettinger, 1992). A possible solution to this problem may be provided by the use of highly specific, exemplar items, that are less likely to be leaked. The current study therefore further investigated the optimal item level for memory detection purposes. More specifically, whether the use of more specific (i.e., exemplar-level) items may counter false-positive results (i.e., specificity), while maintaining high sensitivity (i.e., true positives), as proposed by Osugi (2011, 2018). Results indicate that congruency between how details are initially encoded and subsequently tested is an important moderator for optimal memory detection (also see Geven et al., 2019). Despite near-perfect recollection of the critical details, a significant reduction in the CIT-effect was found in the incongruent conditions.

4.1 | Countering leakage in the CIT

Innocent suspects may be exposed to critical crime details due to rumors or media. Moreover, qualitative analyses of

confessions in DNA exoneration cases in the United States reveal that the majority of false confessors reported accurate, non-public details from the crime (Garrett, 2015), suggesting information contamination. Yet, the knowledge of these contaminated innocents is not expected to be as rich and detailed as the memory of the true perpetrator, who physically committed the crime. To counter possible effects of leakage in the CIT an examiner may probe for more detailed, exemplar-level information (Osugi, 2011). In the current study, we operationalized leakage by revealing categorical information (e.g., *knife*) to participants. When these participants were faced with several exemplar-level options (e.g., several types of knives) in the CIT, they could be correctly classified as unknowledgeable. Hence, presenting items at the exemplar level did not result in increased physiological responding when participants were contaminated with categorical information. Thus, presenting items at the exemplar-level may be a promising pathway to protect against information leakage.

While using exemplar-level items in the CIT is likely to result in increased specificity, it is equally important to correctly detect knowledgeable examinees. The current findings indicate that guilty individuals may still be correctly detected when the CIT involves highly detailed questions at the exemplar level. That is, participants who knew the specific type of knife used in the crime showed strong physiological responding to the correct option *switchblade* in comparison to *swiss knife* or *dagger,* even when the CIT was administered after a 1-week-delay. The results thereby reveal that participants correctly recognized the exemplar-level critical item embedded in equally detailed irrelevant controls, indicating they could sufficiently distinguish between exemplar-level details. Previously, Osugi (2014) revealed successful CIT detection for both high and low distinguishable items using physiological measures, although high distinguishable were associated with a larger CIT-effect, in comparison to low distinguishable items. In the current study, the sensitivity of the exemplar-level CIT was equivalent to the sensitivity obtained with questions phrased at a categorical level (see also Geven et al., 2019). This indicates that the use of highly detailed questions did not compromise CIT validity in our sample. Z-scores for the SCR and RLL measures revealed similar detection efficiency, substantiated by Bayesian evidence for the null, when the questions were asked on category or exemplar level, given congruency between encoding and testing. The difference between the current results and those obtained in the field study by Osugi (2014) could be attributed to differences in experimental design and power, as well as the detailedness of the exemplars. Whereas we moved from a general car to a specific brand, one could even ask about a specific model or other unique features.

In the current study, sensitivity and specificity could be maintained by using highly detailed questions in the CIT.

4.2 | Optimal level of detailedness in the CIT

Is it preferable to present CIT items at the exemplar level? While the current findings suggest that items could be accurately distinguished at both the category and exemplar level, it is important to investigate how a time delay between encoding and the CIT influences recognition. Subjective reports in the follow-up questionnaire as well as memory recall tests indicated a significant difference in recall ability between participants in the immediate and delayed conditions, albeit mean recall rate remained fairly close to the maximum score of eight. No significant differences emerged between participants who were tested directly after encoding the mock crime and participants completing the CIT after a delay in recognition scores and CIT detection scores. While acknowledging that the ceiling effects may be limited to the laboratory set-up, our study is not the only one that found the CIT to detect memories after a time interval (Carmel et al., 2003; Gamer et al., 2010; Geven et al., 2019; Nahari & Ben-Shakhar, 2011). In a preliminary study based on a very small sample size (*n* = 9), Hira et al. (2001, 2002) report to be able to distinguish guilty from innocent participants after even longer time delays (e.g., one month or a year after encoding). However, the effect may perhaps be due to the centrality of the critical items, as reduction in memory and physiological responding may be strongest for peripheral information (Carmel et al., 2003; Gamer et al., 2010; Nahari & Ben-Shakhar, 2011).

While CIT-results were optimal when items were presented on the same level of detailedness as the original encoded stimulus (see also Ben-Shakhar & Gati, 1987; Ben-Shakhar et al., 1995; Geven et al., 2019), significant—yet attenuated—effects were revealed upon presentation of category items (e.g., *knife*) when exemplar-level items (e.g., *switchblade*) were initially encoded. These findings thereby replicate previous studies suggesting that presentation of synonyms of the critical item (e.g., table–desk), pictorial presentation of verbally encoded words, as well as subordinate words (e.g., table–furniture) are sufficient to elicit physiological responses (Ben-Shakhar et al., 1996).

Under realistic circumstances, it is difficult to estimate how details are encoded by the perpetrator. Field examiners rarely know how deeply and at which level of detailedness crime details were originally encoded. To allow for congruency between encoding and testing in the CIT, it is therefore important to firstly investigate which details

perpetrators remember from the crime, how those details are stored, and whether a time delay affects these memory processes. This challenges CIT examiners to select optimal items for memory detection.

4.3 | Limitations

There are several limitations to this study. First, to further investigate the true effectiveness of an exemplar-level CIT to counter the information leakage problem, it is important to validate the present results under less pristine conditions. While in the current laboratory study both the recognition and recall rate revealed ceiling effects, it cannot be expected that perpetrators perfectly encode and retain all crime information in the field. As participants in the current experiment practiced the eight critical items until perfection, the encoding phase is deemed artificial. It is recommended to manipulate the memory factor by diminishing the encoding phase or by elongating the delay. To further enhance external validity in the current paradigm, it would be interesting to more closely mimic the environment expected in the field, for example by using visual stimuli at encoding, or a mock-crime paradigm. The study of Carmel et al. (2003) is an example for such attempt. This study manipulated the type of mock crime and compared the standard mock crime (where encoding of the critical items is guaranteed) to a more realistic mock crime. Indeed, both memory for the critical details and detection efficiency were attenuated under the more realistic circumstances. Yet, this reduction was mediated by the type of question used, indicating more robust memory and detection efficiency when central items were used.

Second, to allow a fully crossed within-between subject design using four different item sets, crime-related items were encoded during the planning of a mock-crime. It should be noted that the mock-crime was thus not executed, but studied verbally. This focus on verbal encoding may explain the reduced CIT-effect for the incongruent Exemplar-Category Item Type. Pre-viewing all test questions and items before the start of the CIT could result in more activation of the category instances, thereby increasing the generalization from exemplar to category, without affecting the validity of the CIT (Verschuere & Crombez, 2008).

Third, it might be important to further investigate possible differences in the level of detail. In the current study, we operationalized categorical and exemplar items, such as *car* and *Citroën*, respectively. However, more than two levels of detailedness could be possible, ranging from subordinate (e.g., *sports car*) to basic level (e.g., *car*) and superordinate terms (e.g., *vehicle*; Pansky & Koriat, 2004). In the CIT, Osugi (2014) found higher CIT-scores for basic

level items in comparison to various subordinates. Given the inverted U-shape of the basic level (i.e., basic level convergence effect; Pansky & Koriat, 2004), the optimal level of detail could be further investigated. For example, it remains to be tested whether perpetrators respond to items that are either too global (e.g., *vehicle*) or too specific (e.g., *blue Citroën C5*).

5 | CONCLUSIONS

The current study demonstrated that the effect of information leakage may be successfully countered by asking more detailed questions. Moreover, the use of exemplar-level items did not lead to lower recognition and physiological responding in the CIT. The findings suggest that exemplar-level CITs may lead to high sensitivity and specificity. Considering the extensive encoding employed in this study, it is recommended to validate our results using a more realistic design, and to explore other ways (e.g., using visual stimulus material) to counter leakage in the CIT.

CONFLICT OF INTEREST
The authors have no known conflict of interest to disclose.

AUTHOR CONTRIBUTIONS
Linda Marjoleine Geven: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Resources; Validation; Visualization; Writing-original draft; Writing-review & editing. **Bruno Verschuere:** Conceptualization; Methodology; Supervision; Validation; Writing-review & editing. **Merel Kindt:** Conceptualization; Supervision; Writing-review & editing. **Shani Vaknine:** Investigation; Project administration; Software. **Gershon Ben-Shakhar:** Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing-review & editing.

ORCID
Linda Marjoleine Geven [ID] https://orcid.org/0000-0001-5075-5223

REFERENCES
Alceste, F., Jones, K. A., & Kassin, S. M. (2020). Facts only the perpetrator could have known? A study of contamination in mock crime interrogations. *Law and Human Behavior*, *44*(2), 128–142. https://doi.org/10.1037/lhb0000367

Ben-Shakhar, G., & Elaad, E. (2002). Effects of questions' repetition and variation on the efficiency of the Guilty Knowledge Test: A reexamination. *Journal of Applied Psychology*, *87*(5), 972–977. https://doi.org/10.1037/0021-9010.87.5.972

Ben-Shakhar, G., Frost, R., Gati, I., & Kresh, Y. (1996). Is an apple a fruit? Semantic relatedness as reflected by psychophysiological responsivity. *Psychophysiology*, *33*(6), 671–679. https://doi.org/10.1111/j.1469-8986.1996.tb02363.x

Ben-Shakhar, G., & Gati, I. (1987). Common and distinctive features of verbal and pictorial stimuli as determinants of psychophysiological responsivity. *Journal of Experimental Psychology: General*, *116*(2), 91–105. https://doi.org/10.1037/0096-3445.116.2.91

Ben-Shakhar, G., Gati, I., & Salamon, N. (1995). Generalization of the orienting response to significant stimuli: The roles of common and distinctive stimulus components. *Psychophysiology*, *32*(1), 36–42. https://doi.org/10.1111/j.1469-8986.1995.tb03403.x

Bradley, M. T., & Rettinger, J. (1992). Awareness of crime-relevant information and the Guilty Knowledge Test. *Journal of Applied Psychology*, *77*(1), 55–59. https://doi.org/10.1037/0021-9010.77.1.55

Bradley, M. T., & Warfield, J. F. (1984). Innocence, information, and the Guilty Knowledge Test in the detection of deception. *Psychophysiology*, *21*(6), 683–689. https://doi.org/10.1111/j.1469-8986.1984.tb00257.x

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the Guilty Knowledge Test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied*, *9*(4), 261–269. https://doi.org/10.1037/1076-898X.9.4.261

Christiaansen, R. E. (1980). Prose memory: Forgetting rates for memory codes. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 611–619.

Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the Guilty Knowledge Test. *Psychophysiology*, *34*(5), 587–596. https://doi.org/10.1111/j.1469-8986.1997.tb01745.x

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, *77*(5), 757–767. https://doi.org/10.1037/0021-9010.77.5.757

Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the Guilty Actions Test. *Biological Psychology*, *83*(2), 101–107. https://doi.org/10.1016/j.biopsycho.2009.11.005

Garrett, B. L. (2015). Contaminated confessions revisited. *Virginia Law Review*, *101*(2), 395–454.

Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2019). It's a match!? Appropriate item selection in the Concealed Information Test. *Cognitive Research: Principles and Implications*, *4*(1), 11. https://doi.org/10.1186/s41235-019-0161-8

Geven, L. M., klein Selle, N., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2018). Self-initiated versus instructed cheating in the physiological Concealed Information Test. *Biological Psychology*, *138*, 146–155. https://doi.org/10.1016/j.biopsycho.2018.09.005

Hira, S., Sasaki, M., Matsuda, T., Furumitsu, I., & Furedy, J. J. (2001). Pz-recorded P300 is highly accurate and sensitive to a memorial manipulation in an objective laboratory Guilty Knowledge

Test. *Psychophysiology*, *38*, S50. https://doi.org/10.1111/j.1469-8986.2001.tb00003.x

Hira, S., Sasaki, M., Matsuda, T., Furumitsu, I., & Furedy, J. J. (2002). A year after the commission of a mock crime, the P300 amplitudes, but not reaction time, are sensitive guilty knowledge test indicators. *Psychophysiology*, *39*, S42. https://doi.org/10.1111/j.1469-8986.2002.tb00008.x

JASP Team. (2018). JASP (Version 0.8.4) [Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

klein Selle, N., Agari, N., & Ben-Shakhar, G. (2019). Hide or seek? Physiological responses reflect both the decision and the attempt to conceal information. *Psychological Science*, *30*(10), 1424–1433. https://doi.org/10.1177/0956797619864598

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2016). Orienting versus inhibition in the Concealed Information Test: Different cognitive processes drive different physiological measures. *Psychophysiology*, *53*(4), 579–590. https://doi.org/10.1111/psyp.12583

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2017). Unraveling the roles of orienting and inhibition in the Concealed Information Test. *Psychophysiology*, *54*(4), 628–639. https://doi.org/10.1111/psyp.12825

Koriat, A., Levy-Sadot, R., Edry, E., & de Marcas, S. (2003). What do we know about what we cannot remember? Accessing the semantic attributes of words that cannot be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1095–1105. https://doi.org/10.1037/0278-7393.29.6.1095

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385–388. https://doi.org/10.1037/h0046060

Meijer, E. H., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta-analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, *51*(9), 879–904. https://doi.org/10.1111/psyp.12239

Meijer, E. H., Verschuere, B., & Merckelbach, H. (2010). Detecting criminal intent with the Concealed Information Test. *The Open Criminology Journal*, *3*, 44–47. https://doi.org/10.2174/1874917801003010044

Nahari, G., & Ben-Shakhar, G. (2011). Psychophysiological and behavioral measures for detecting concealed information: The role of memory for crime details. *Psychophysiology*, *48*(6), 733–744. https://doi.org/10.1111/j.1469-8986.2010.01148.x

Ofshe, R. J., & Leo, R. A. (1997). The social psychology of police interrogation: The theory and classification of true and false confessions. *Studies in Law Politics and Society*, *16*, 189–254.

Osugi, A. (2011). Daily application of the concealed information test: Japan. In B. Verschuere, E. Meijer, & G. Ben-Shakhar (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 253–275). Cambridge University Press. https://doi.org/10.1017/CBO9780511975196.015

Osugi, A. (2014). Review and analysis of the practical data conducted in Japanese criminal investigation. *International Journal of Psychophysiology*, *2*(94), 131. https://doi.org/10.1016/j.ijpsycho.2014.08.617

Osugi, A. (2018). Field findings from the Concealed Information Test in Japan. In J. P. Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 97–121). Academic Press. https://doi.org/10.1016/B978-0-12-812729-2.00005-7

Pansky, A., & Koriat, A. (2004). The basic-level convergence effect in memory distortions. *Psychological Science*, *15*(1), 52–59. https://doi.org/10.1111/j.0963-7214.2004.01501009.x

Peth, J., Vossel, G., & Gamer, M. (2012). Emotional arousal modulates the encoding of crime-related details and corresponding physiological responses in the Concealed Information Test. *Psychophysiology*, *49*(3), 381–390. https://doi.org/10.1111/j.1469-8986.2011.01313.x

Podlesny, J. A. (1993). Is the guilty knowledge polygraph technique applicable in criminal investigations: Q review of FBI case records. *Crime Laboratory Digest*, *20*(3), 57–61.

Podlesny, J. A. (2003). A paucity of operable case facts restricts applicability of the guilty knowledge technique in FBI criminal polygraph examinations. *Forensic Science Communications*, *5*(3).

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (Eds.). (2011). *Memory detection: Theory and application of the Concealed Information Test*. Cambridge University Press.

Verschuere, B., & Crombez, G. (2008). Déjà vu! The effect of previewing test items on the validity of the Concealed Information Polygraph Test. *Psychology, Crime and Law*, *14*(4), 287–297. https://doi.org/10.1080/10683160701786407