The European Journal of Psychology Applied to Legal Context 7 (2015) 73–79

The European Journal of Psychology Applied to Legal Context

www.elsevier.es/ejpal

How good are future lawyers in judging the accuracy of reminiscent details? The estimation-observation gap in real eyewitness accounts

Aileen Oeberst*

Knowledge Media Research Center, Tuebingen, Germany

ARTICLE INFO

Article history:
Received 14 July 2014
Accepted 5 March 2015
Available online 19 June 2015

Keywords:
Eyewitness memory
Reminiscence
Implicit theories
Credibility
Judgment

ABSTRACT

Research has shown a discrepancy between estimated and actually observed accuracy of reminiscent details in eyewitness accounts. This estimation-observation gap is of particular relevance with regard to the evaluation of eyewitnesses' accounts in the legal context. To date it has only been demonstrated in non-naturalistic settings, however. In addition, it is not known whether this gap extends to other tasks routinely employed in real-world trials, for instance person-identification tasks. In this study, law students witnessed a staged event and were asked to either recall the event and perform a person identification task or estimate the accuracy of the others' performance. Additionally, external estimations were obtained from students who had not witnessed the event, but received a written summary instead. The estimation-observation gap was replicated for reminiscent details under naturalistic encoding conditions. This gap was more pronounced when compared to forgotten details, but not significantly so when compared to consistent details. In contrast, accuracy on the person-identification task was not consistently underestimated. The results are discussed in light of their implications for real-world trials and future research.

Habilidad de los futuros abogados para valorar la precisión de detalles evocados. La brecha entre estimación y observación en el relato real de testigos oculares

Palabras clave:
Recuerdo en testigos oculares
Evocación
Teorías implícitas
Credibilidad
Valoración

RESUMEN

La investigación ha revelado que hay diferencias entre la precisión estimada y la observada realmente en los detalles evocados en los relatos de testigos oculares. La brecha entre estimación y observación es especialmente importante en la evaluación de los relatos de testigos oculares en el contexto legal. Sin embargo, hasta la fecha solo se ha demostrado en contextos no naturales. Además, no se sabe si esta brecha es extensible a otras tareas habituales en pruebas en el mundo real, como las de identificación de personas. En este estudio, estudiantes de Derecho presenciaron un montaje y se les pidió que lo recordaran y llevaran a cabo una tarea de identificación de personas o bien que estimaran la precisión de la actuación de los demás. Además se obtuvieron estimaciones externas de los estudiantes que no habían presenciado el montaje, recibiendo un resumen escrito en su lugar. La brecha entre estimación y observación se replicó para detalles evocados en condiciones de codificación naturales. La brecha era más pronunciada cuando se comparaban con detalles olvidados, aunque no significativa cuando se comparaban con detalles congruentes. Por el contrario, no fue infravalorada de un modo coherente la precisión de la tarea de identificación de personas. Se comentan los resultados desde el punto de vista de sus implicaciones para los ensayos en el mundo real y la investigación futura.

* Corresponding author. Knowledge Construction Lab. Knowledge Media Research Center. Schleichstr. 6. D-72076 Tuebingen, Germany.
E-mail address: a.oeberst@iwm-kmrc.de (A. Oeberst)

http://dx.doi.org/10.1016/j.ejpal.2015.03.002

Consider the case of a witness who has been questioned twice by the police and who reports some details only at the second interrogation, one week later. Would you trust such novel recollections?

Research indicates that reminiscent details—details that have not been previously reported (Ballard, 1913) —are perceived to be less credible than details that have been consistently reported in both interrogations (Berman & Cutler, 1996). More importantly, they are perceived to be less credible than they actually are. In a recent study, Oeberst (2012) asked students to encode two different types of stimuli (pictures in Experiment 1 and a film in Experiment 2). Directly after encoding as well as one week later, they were asked to remember as many details as they could. Crucially, another group was asked to estimate their fellow students' accuracy on this task. Accuracy of reminiscent items was tremendously underestimated: while, after one week, only 19% of novel recollections were *expected* to be accurate, 84% were *observed* to be accurate (Oeberst, 2012; Exp. 2). Moreover, even though an estimation-observation gap was also found for forgotten as well as for consistently recalled items, it was most pronounced for reminiscent items. These findings are of particular relevance when it comes to eyewitness testimony and its evaluation in the legal context. After all, a discrepancy between actual and assumed accuracy can result in momentous consequences for the involved persons' lives. But does this striking pattern extend to more complex and dynamic real-world events? The current study aimed at answering this question by having participants witness a staged event. In addition, it was examined whether an estimation-observation gap would also be found in person-identification tasks, which are often used in real-world trials.

Presumably based on informal observations of one's own memory for everyday experiences, individuals commonly hold the implicit assumption that memory for an event is best immediately after that event, and that it subsequently decreases with the passage of time (Ballard, 1913; Gilbert & Fisher, 2006; Magnussen et al., 2008; Oeberst, 2012)[1]. Although this is true with respect to net memory performance over extended time intervals (e.g., Ebbinghaus, 1885), forgetting does not necessarily preclude reminiscence of items, which were not previously recollected (e.g., Buschke, 1974) —it only implies that forgetting exceeds reminiscence (Erdelyi, 2010). However, the pattern of forgetting is much more consistent with one's expectations (Fisher, Brewer, & Mitchell, 2009; Gilbert & Fisher, 2006). In contrast, the frequent occurrence of reminiscence as well as its reliability (Baugerud, Magnussen, & Melinder, 2014; Bluck, Levine, & Laulhere, 1999; Brock, Fisher, & Cutler, 1999; Dunning & Stern, 1992; Erdelyi, 2010; Gilbert & Fisher, 2006; Oeberst, 2012) is rather unknown.

These considerations gain particular importance with regard to the legal system. After all, decision-makers in this system are laypersons when it comes to memory functioning (Fisher et al., 2009; Wise & Safer, 2003). Thus, empirical evidence stands in stark contrast to what these laypersons might expect. Expectations, however, guide the evaluation of eyewitness evidence (Leippe & Romanzcyk, 1989). Moreover, some jury instructions even explicitly recommend consideration of the (in)consistency of a witnesses' statement made on various occasions (e.g., Florida Supreme Court Standard Jury Instructions 3d, 2009). Reminiscence falls under the umbrella of such inconsistencies since the term 'inconsistencies' is referred to in a rather general way (e.g., Sixth Circuit Criminal Pattern Jury Instructions, No. 107, 2005) thereby conflating different types of inconsistencies (e.g., reminiscence, forgetting, contradictions). Logical and empirical aspects argue against such a conflation, however (Berman & Cutler, 1996; Brock et al., 1999; Fisher et al., 2009; Gilbert & Fisher, 2006). After all, only contradictions involve at least one false statement[2]. Details, in contrast, which were recollected only once, but not another time, could very well be accurate. That is, neither details, which were forgotten thereafter, nor recollections that were reported only at a later date (i.e., reminiscence) are necessarily inaccurate. However, whereas the pattern of forgetting seems in line, reminiscence seems at odds with one's expectations (Fisher et al., 2009; Gilbert & Fisher, 2006). Furthermore, doubts in the reliability of reminiscent recollections may be nourished by lawyers, who are trained to provoke such inconsistencies (e.g., Prager, Moran, & Sanchez, 1996) in order to discredit vulnerable eyewitnesses (Ellison, 2001).

Despite the sizable gap between estimated and observed accuracy of reminiscent details found by Oeberst (2012) it remains unclear whether the results generalize to naturalistic settings. In that study, participants' attention was explicitly drawn to the to-be-remembered materials because of the research setting (i.e., participants were explicitly asked to watch a video or view pictures), which is usually not the case in real-world settings. Moreover, events in the real world differ from pictures and films in various ways. Beyond differences in scaling (screen-size vs. life-size) and dimensionality (two- vs. three-dimensional, e.g., Schmitt & Anderson, 2002), witnesses in real-world settings not only view the event from their unique perspective, but are also involved to some extent. All in all, encoding pictures or films is not comparable to real-world situations and hence, generalizability cannot be taken for granted (e.g., Fariña, Arce, & Real, 1994; Ihlebæk, Løve, Eilertsen, & Magnussen, 2003). Despite this insight and previously raised concerns regarding ecological validity (e.g., McCloskey & Egeth, 1983; Yuille & Wells, 1991) hardly anything is known for adult witnesses about the actual accuracy of reminiscent items under natural encoding conditions since research on the accuracy of reminiscence usually employed videos (e.g., Brock et al., 1999; Gilbert & Fisher, 2006; Scrivner & Safer, 1988; Turtle & Yuille, 1994) and studies investigating memory of naturally encoded events (e.g., autobiographical memory) often lack the possibility to assess accuracy reliably (e.g., Campbell, Nadel, Duke, & Ryan, 2011; Nadel, Campbell, & Ryan, 2007) or the possibility to identify genuine reminiscences due to media coverage (Yuille & Cutshall, 1986). However, should the estimation-observation gap be of any relevance for real trials, it is necessary to show that it occurs in more naturalistic settings as well. The main objective of the present study is therefore to examine whether the large discrepancy between expected and observed memory accuracy would replicate under naturalistic conditions.

A second goal of the present study was to examine whether the estimation-observation gap extends to identification tasks. To date, a majority of wrongful convictions may be attributed to errors in this process (Innocence Project, 2012; Scheck, Neufeld, & Dwyer, 2000). This suggests the opposite of what has been found for reminiscent items, namely an *over*estimation of what eyewitnesses are actually capable of. Moreover, research conducted under natural encoding conditions hints towards a rather low actual performance (e.g., Behrman & Davey, 2001; Fariña

[1] Note that there are also two studies arriving at the opposite conclusion, namely that the forgetting curve does not represent a common assumption among (potential) jurors, judges, and law enforcement (Benton, Ross, Bradshaw, Tomas, & Bradshaw, 2006; Wise & Safer, 2003). I believe, however, that this may be due to its operationalization. Both studies assessed (dis)agreement to the statement "The rate of memory loss for an event is greatest right after an event and then levels off over time" (Wise & Safer, 2003; p. 11), which represents a rather complicated wording and might thus be difficult to understand. Simple visualizations of memory performance over time as used by Oeberst (2012), in contrast, should be less prone to misunderstandings.

[2] Note that there are some cases such as when continuous information (e.g., age) is involved, in which two different statements could both count as correct – if one assumes a certain range of correct answers (e.g., 22-25 years).

et al., 1994; Pozzulo, Crescini, & Panton, 2008; Read, Tollestrup, Hammersley, McFadzen, & Christensen, 1990; Tollestrup, Turtle, & Yuille, 1994), which might preclude a gap between observations and estimations. A direct comparison, however, has not been conducted yet. Because eyewitness-identification tasks are routinely employed (Wells & Olson, 2002), it is important to know whether performance on these tasks is well estimated or rather over- or underestimated by authorities within the legal system. To this end, a photo-lineup procedure was included in the present study.

The present study thus set out to investigate whether the estimation-observation gap would be obtained under real-world conditions as well. To this end, law-students witnessed a staged event. Immediately after this event (t_1) and again one week later (t_2), half of them reported their recollections (observation group), and the other half estimated the amount and accuracy of their fellow student's accounts (online estimation group). Additionally, estimations from participants who had not witnessed the event themselves (external estimation group) were collected. Following Oeberst (2012), I expected actual accuracy of reminiscent items to be significantly underestimated by both estimation groups. Moreover, I hypothesized that this difference between estimated and observed accuracy would be higher for reminiscent compared to forgotten as well as consistently recollected items. In addition, a person-identification task at t_2 was included to explore whether accuracy in this case is under-estimated as well—or whether it is even overestimated, as suggested by previous indirect evidence.

Method

Participants

Altogether, 103 undergraduate law students (63 female) from four different classes of a German university took part in both sessions of this study. Mean age was 20.05 ($SD = 1.56$) (see Table 1). Observation and estimation conditions did not differ with respect to age or sex, p's > .1. All participants received candy for compensation.

Table 1
Cell Sample Sizes (N) per Condition.

	Observation group	Online estimation group	External estimation group
Target present	11	10	25
Target absent	16	18	23
Total sample	27	28	48

Design

This study comprised three independent variables. First, participants were assigned to one of three groups. One group had to provide their own recollections of the event (observation group) whereas the other two groups had to provide estimations of their peers' memory performance. Two forms of estimation groups existed: one group provided estimations after having witnessed the same event and at the same time as the observation group (online estimation group) and another group provided estimations only on the basis of a short description of the event (external estimation group). The second independent variable was point in time and it varied within participants but only for the observation and online estimation groups. Specifically, their recollections and estimations were assessed once immediately after the staged event (t_1) as well as one week later (t_2). Third, presence of the target in the photo lineup of the identification task varied quasi-experimentally between participants in all groups.

Materials

Observation and online-estimation group. The to-be-remembered stimuli consisted of a staged event, which was witnessed at the very beginning of two introductory classes for law students (Course 1, 2). The incident involved a young man who entered the room, walked to the projector in the front, turned it on (light was projected on one wall), turned it off again, unplugged it, coiled the cord and pushed the projector towards the door. In Course 2, the man actually left the room and took the projector with him. In Course 1, the room was too crowded so that he had to abandon the projector after appearing to try to take it with him before he left. His attempt was obvious, however, as indicated by the witnesses' accounts. The event lasted 0:55 min and 0:50 min in Course 1 and 2, respectively. In order to enable accuracy analyses, the staged event was covertly videotaped by a confederate (a law student who recorded the incident with a small camera hidden in his folder).

External-estimation group. To ensure external validity, there was an external-estimation group. In this group, participants of two different introductory classes for law students (Course 3, 4) received a short description of the staged event, which was based on the recollections obtained from the observation group. After all, this is the kind of information decision-makers in the legal context are provided with. To this end, I determined the average amount of details ($n = 8$) recollected by the observation group participants and then chose the eight most frequently reported details. The passage read: "A man entered the room, walked to the projector, tampered with it, unplugged it, coiled the cord and pushed the projector out of the room. The man was described as relatively tall (180-195 cm), in his twenties with dark blond hair. He was wearing dark glasses, a jeans, and a sweater."

Procedure

Observation and online-estimation group. The event was staged shortly before the tutorial was supposed to start. Most of the students were already present whilst the lecturer was not. Immediately after the "thief" had left the room, the lecturer and the experimenter entered the room and told the students that the incident was part of an eyewitness study. They were briefly informed that the study was about what witnesses actually recollect and in what people believe about eyewitnesses' recollections and that the study therefore consisted of two tasks that differed between participants. The experimenter then distributed the instructions and answer-sheets for the two groups (observation vs. estimation) in an alternating order, rendering assignment to condition pseudo-random.

Participants in the observation group were asked to recall the incident and describe the person as accurately as possible. Participants in the online-estimation group estimated the average accuracy of their fellow students' memory reports ("What do you think, how many of the details your peer students remember are accurate", 0% = *all details are inaccurate* to *100% all details are accurate*). Finally, all participants indicated their sex and age and generated a code in order to be able to match their data between both sessions. Participants were neither informed of this purpose nor of the second session, however. Rather, the experimenter led them to believe that the study consisted of one session only by thanking them and distributing candy for compensation.

One week later, the experimenter returned and briefly informed the students that the study actually consisted of two sessions. The task in each condition was identical to t_1, with estimations referring to memory performance at t_2. Additionally, participants were asked to estimate accuracy for the following item types: details recollected both times (consistent items), details reported only in the first session (forgotten items), and details recalled only in

Table 2
Estimated and Observed Accuracy of the Different Item Types (Mean Percentage of Correct Items, Standard Deviations, Medians, and Interquartile Range).

	Reminiscent		Forgotten		Consistent	
	M (SD)	Median (IQR)	*M* (SD)	Median (IQR)	*M* (SD)	Median (IQR)
Observation [a]	70.65 (39.88)	100.00 (55.00)	79.13 (30.00)	100.00 (45.83)	92.17 (6.52)	90.00 (13.33)
Online estimation	21.79 (14.80)	20.00 (18.75)	51.96 (22.66)	50.00 (30.00)	53.39 (25.50)	60.00 (45.00)
External estimation	25.88 (17.93)	20.00 (26.00)	57.23 (21.60)	60.00 (23.75)	56.00 (22.03)	60.00 (35.00)
O-E-Gap	46.28		23.84		37.15	

Note. For calculation of the observation-estimation gap (i.e., the difference between observed and estimated accuracy), estimations were collapsed across both estimation groups. Means reported in the paper may slightly deviate in the observation group due to missing values (of reminiscent or forgotten details) in within-subjects comparisons.

[a] Percentage of confabulations for each item type can be inferred from the difference to 100%. That is, 29.35% of the reminiscent, 21.87% of the forgotten and 7.83% of the consistently reported details were *inaccurate*.

the second session (reminiscent items). In order to avoid misunderstandings, each item type was described explicitly before prompting a response (e.g., for the reminiscent items: "Consider the case that a detail was reported today, in the second questioning, but not last week, in the first questioning. What do you think is the percentage of such details being accurate?").

Subsequently, participants engaged in an identification task. As recommended (Brewer & Wells, 2006; Wells, Memon, & Penrod, 2006; Wells et al., 1998), the specific number of photos to be viewed (9) was not revealed and it was emphasized that the target person might or might not be among them. Following a sequential lineup procedure, participants in the observation condition were asked to decide for each photo, whether it depicted the target person or not (dichotomously) and to indicate their confidence. Participants in the estimation condition were asked to estimate the percentage of correct identifications after they had watched the whole series of pictures if they thought that the target had been present as well as the percentage of mistaken identifications.

Distractor selection was based on the witnesses' verbal description of the target person and all photos were taken under similar conditions (i.e., frontal; white plain background, neutral facial expression; cf. Wells et al., 2006; Wells et al., 1998). None of them was affiliated with the law school where testing occurred. The distractor who resembled the target person most, as determined by a pretest, was substituted for the target in the target absent condition (Course 1). In Course 2, in contrast, the target was present. In order to minimize experimenter effects (Wells et al., 1998; see also Greathouse & Kovera, 2009), the order of pictures had been arranged by another person and the experimenter did not see the presentation. Finally, the experimenter thanked all participants and assured them that there would be no further sessions. Again, they received candy for compensation.

External estimation group. Participants in this group received a brief description of the incident (see above) and were then asked to estimate average accuracy of the recollections at t_1 and t_2 (measurement was in principle identical to the online estimation group only without the temporal references). They were then informed of the fact that the witnesses had recollected details either at both dates, or only at t_1 and t_2 respectively and were asked to estimate the percentage of correct details for these different item types. Finally, they read about the identification task employed (including instructions) and were presented with the photos participants had seen. Participants in Course 3 received the target-present condition and were informed that the target was among them[3]. They were then asked to indicate the percentage of students who had picked one of the persons, and among these the rate of correct identifications. Participants in Course 4 were informed of the target being absent and only indicated the percentage of students who had picked one of the persons (i.e., had committed a misidentification).

Data Analysis

Recollections were split into single information units (e.g., "The man coiled up the cord" was separated into "man", "coil up", "cord"). Recollections from t_2 were classified as either consistent (details recollected at t_1 and t_2: cord / cord), forgotten (details reported only at t_1 but not at t_2: cord / no description of cord), or reminiscent (details recalled only at t_2: no description of cord / cord). Changes that did not affect the content (e.g., synonyms) were treated as consistent information. Contradictions occurred only in 5 instances, therefore not allowing for reliable statistical analyses. For this reason this item-type was excluded. Note, however, that the pattern of results was identical when a more conservative test was applied—when contradictions were treated as two different units of information such that the first resembling a forgotten item and the second falling into the category of a reminiscent item (see Oeberst, 2012). Unverifiable statements such as suspicions about intentions of actors were not included. All units of information were scored for accuracy.

Results

The analyses reported below were also run with course as another between subjects factor. However, this factor yielded no main effects or interactions, *p*'s > .18. For the sake of brevity and clarity, this factor was omitted. Since the distribution of accuracy of reminiscent details was extremely left-skewed leading to high standard deviations, Table 2 also displays medians and non-parametrical tests assured statistical validity.

Planned Analyses

Reminiscence was operationalized as the number of details that had been mentioned at t_2 but not at t_1. Reminiscence occurred frequently (M = 2.42, SD = 1.84, range: 0-7) and was documented

[3] The rationale for this procedure was that decision makers in the legal context as well would know whether the suspect had been among the persons in the identification task. Hence, it is not a question of whether the suspect was the actual culprit, but rather, whether the accused person in court had been in the line-up (and whether he had been identified, or not).

for the overwhelming majority of participants (22 out of 27).

First, the hypothesis that accuracy of reminiscent items is generally underestimated was tested. An ANOVA with group (observation, online estimation, external estimation) as between-subjects factor on (actual and estimated) accuracy of reminiscent items indeed yielded a significant main effect of group, $F(2, 94) = 31.61$, $p < .001$, $\eta_p^2 = .40$. Post-hoc comparisons (Bonferroni corrected *p*s) revealed that the observation group differed significantly from both estimation groups, *p*'s < .001, whereas the estimation groups did not differ from one another, $p = 1$. As can be seen in Table 2, actual accuracy of reminiscent items was significantly higher than had been expected by either estimation group. Thus, the estimation-observation gap was replicated for reminiscent items under naturalistic encoding conditions.

In order to test whether this gap is particularly pronounced for reminiscent items, a mixed ANOVA with item type (reminiscent, forgotten, consistent) as within-subjects factor, and group (observation, online estimation, external estimation) as between subjects factor was run. It revealed a significant main effect of group, $F(2, 91) = 36.09$, $p < .001$, $\eta_p^2 = .44$, and a significant main effect of item type, $F(2, 182) = 49.71$, $p < .001$, $\eta_p^2 = .35$. These main effects were qualified by a significant interaction, however, $F(4, 182) = 3.37$, $p = .01$, $\eta_p^2 = .07$. To elucidate this interaction, two separate ANOVAs comparing reminiscent items to forgotten and consistent items, respectively, were conducted. The mixed ANOVA with item type (reminiscent, forgotten) as within-subjects factor, and group (observation, online estimation, external estimation) as between-subjects factor again revealed a significant main effect of item type, $F(1, 92) = 57.04$, $p < .001$, $\eta_p^2 = .38$, a significant main effect of group, $F(2, 94) = 27.46$, $p < .001$, $\eta_p^2 = .37$, as well as a significant interaction of both factors, $F(2, 92) = 6.29$, $p < .01$, $\eta_p^2 = .12$. As can be seen in Table 2, the estimation-observation gap was larger for reminiscent (46.28) than for forgotten (23.84) items. Moreover, estimation conditions, again, did not differ from one another, $p = .92$, whereas both differed significantly from the observation group, *p*'s < .001 (Bonferroni corrected *p*s). Hence, accuracy of reminiscent items was underestimated to a larger extent than accuracy of forgotten items. The pattern of results was different, however, when comparing reminiscent and consistent items. There was only a significant main effect of item type, $F(1, 93) = 79.96$, $p < .001$, $\eta_p^2 = .46$, as well as a significant main effect of group, $F(2, 93) = 51.38$, $p < .001$, $\eta_p^2 = .53$, but no significant interaction, $F(2, 93) = 0.59$, $p = .56$. Although the results descriptively match the expectation that the estimation-observation gap is larger for reminiscent items (46.28) than for consistently recollected items (37.15), their difference was not significant in the ANOVA conducted. Noticeably, in this case the nonparametric analysis yielded a much lower *p*-value, $\chi^2(2) = 3.92$, $p = .14$.

Taken together, there was a substantial observation-estimation gap regarding accuracy. Actual recollections were much more accurate than had been expected. The discrepancy was pronounced for reminiscent items, though only in comparison with forgotten items but not with consistently recalled items.

Additional Analyses

Accuracy of the various item types. In order to test for differences with regard to the accuracy of consistently recollected, forgotten, and reminiscent items, an ANOVA was conducted in the observation group only with item type (consistent, forgotten, reminiscent) as within-subjects factor on accuracy. It yielded a marginally significant main effect of item type, $F(2, 36) = 3.22$, $p = .06$, $\eta_p^2 = .15$. The corresponding non-parametric analysis was far from significance, $\chi^2(2) = 0.28$, $p = .87$, however. Pairwise comparisons revealed that reminiscent items proved to be significantly less accurate than consistent items, $t(20) = 2.74$, $p = .01$, $d = 0.84$, $Z = 2.22$, $p = .03$, but comparably accurate as forgotten items, $t(18) = 0.69$, $p = .50$, $d = 0.84$, $Z = 0.65$, $p = .52$. Comparing consistent and forgotten items yielded a significant difference in a paired *t*-test, $t(23) = 2.12$, $p < .05$, $d = 0.59$, but only a trend in the Wilcoxon-Test, $Z = 1.61$, $p = .11$.

Overall memory performance. Overall accuracy of the eyewitness accounts was analyzed in a 2 (point in time: t_1, t_2) x 3 (group: observation, online estimation, external estimation) mixed ANOVA. The analysis yielded a significant main effect of group, $F(2, 99) = 95.20$, $p < .001$, $\eta_p^2 = .66$, a significant main effect of point in time, $F(1, 99) = 146.38$, $p < .001$, $\eta_p^2 = .60$, as well as a significant interaction of both factors, $F(2, 99) = 36.08$, $p < .001$, $\eta_p^2 = .42$. As can be seen in Table 3, actual accuracy was significantly higher than expected by both estimation conditions, *t*'s > 7, *p*'s < .001. Moreover, participants remained accurate over one week, $p > .42$. Estimated accuracy, however, significantly decreased over time in both estimation groups *t*'s > 7, *p*'s < .001. Thus, participants expected accuracy (not amount recalled!) to significantly decrease from t_1 to t_2. In reality, however, it did not.

Table 3
Estimated and Observed Mean Overall Accuracy (SD) for both Points in Time

	t_1	t_2
Observation group	88.60 (9.24)	87.37 (10.80)
Online estimation group	56.61 (13.30)	32.59 (16.41)
External estimation group	61.38 (16.48)	38.10 (16.68)

Identification Task

Target present condition. One-sample *t*-tests indicated that the observed rate of correct identifications did not differ from the estimations provided by the external estimation group (see Table 4), $t(24) = 1.06$, $p = .30$, but was by trend overestimated by the online estimation group, $t(9) = 1.91$, $p = .09$. Estimation groups did not differ from one another, $t(33) = 0.16$, $p = .88$. The percentage of false identifications in contrast was overestimated by both, the online, $t(17) = 2.74$, $p = .01$, as well as the external estimation group, $t(24) = 2.64$, $p = .01$, which, again, did not differ from one another, $t(41) = 0.70$, $p = .49$.

Table 4
Percentage of Witnesses Observed or Estimated to Provide Correct or Incorrect Identifications.

		Observation	Online estimation	External estimation
Target present	Correct identifications	31.30	44.00 (21.06)	35.32 (18.92)
	False identifications	31.30	47.78 (25.57)	42.72 (21.64)
Target absent	False identifications	45.50	53.00 (17.03)	52.39 (28.48)

Note. Percentages do not sum up to 100% due to the fact that it was possible not to choose a person from the lineup.

Target absent condition. The rate of false identifications expected by the online estimation group as well as the external estimation condition did not significantly exceed the actual percentage of eyewitnesses who mistakenly identified an innocent person, $t(9) = 1.39$, $p = .19$ and $t(22) = 1.16$, $p = .26$, respectively.

Discussion

The primary aim of this study was to explore whether the estimation-observation gap extends to natural encoding conditions. By using a staged event, the estimation-observation gap was replicated. Actual memory performance was underestimated by future attorneys, lawyers, and judges, and the extent of the underestimation was sizeable. Moreover, this pattern was significantly more pronounced for reminiscent details than for forgotten items, which further documents the counterintuitive nature of reminiscence (Fisher et al., 2009; Gilbert & Fisher, 2006). The estimation-observation gap was also descriptively higher for reminiscence when compared to consistently recollected items; this difference did not reach statistical significance, however.

At first glance, this is inconsistent with Oeberst (2012) who found significant differences to both kinds of items in a study involving an overall smaller *N*. A closer inspection of the data points to smaller effect sizes in this study ($\eta p^2 = .01$; Oeberst, 2012: $\eta p^2 = .09$), which suggests that the study might have lacked the power to detect the effect. However, future research taking the power issue into account is needed to reliably answer this question.

Another slight difference to previous findings under non-naturalistic encoding conditions (Oeberst, 2012) regards the accuracy of reminiscent items. On average, they were less accurate (71%; Oeberst, 2012: 84%). Could this be due to the different encoding conditions administered in both studies? On the one hand, the mean might not represent the data ideally because the distribution was extremely left skewed (the median was 100%). On the other hand, other authors who had used videotapes instead of staged events reported comparable average rates to the ones found in the current study (e.g., Brock et al., 1999; Turtle & Yuille, 1994). However, there is no study directly comparing videos to actually witnessed events that analyzed reminiscence (see Roebers, Gelhaar, & Schneider, 2004; Thierry & Spence, 2004 for comparisons regarding overall memory performance). Thus, further research is needed.

The second objective of this study was to explore the generalizability of the findings to eyewitness identification tasks. In this case, the pattern of results was less straightforward. Although observations met expectations in some instances, deviations occurred in both directions—once overestimating the frequency of wrongful identifications and another time the frequency of correct identifications (or in other words: underestimating the percentage of witnesses, who refrained from an identification). Despite the partly small sample sizes, it is noteworthy that it was not so much that estimations differ from the estimations regarding the accuracy of overall recollections. Rather, actual memory performance was much worse than in the case of free recollections. Essentially, the identification task tests specific recognition memory, and previous research indicates some crucial differences. The presentation of other (but highly similar) items at the time of test may increase uncertainty among witnesses (e.g., Robinson & Johnson, 1996) and ultimately impair performance (e.g., Tulving & Thomson, 1971). It remains unclear, however, whether estimators are aware of these differences and could adjust their estimations accordingly. After all, estimations of others' performance have never been assessed in this paradigm before. Future research will need to clarify this.

Moreover, the potential influence of other relevant aspects that may come along with real-world settings posits relevant questions for future studies. On the one hand, this includes aspects of the encoding (e.g., direct involvement, trauma, representative forensic witness sample; Ihlebæk et al., 2003). On the other hand, however, it concerns the numerous other aspects that characterize the legal system and decision-making therein (Konečni & Ebbesen, 1992). It has been questioned, for instance, whether elicitation of the recollections in the lab (with mock witnesses) actually provides participants with the identical task that real witnesses face (Fariña et al., 1994). This argument might apply to various other aspects and thus, it has been convincingly argued that evidence from simulations may have "face validity" but actually lacks a consideration of the relevant factors that are present (Konečni & Ebbesen, 1992). Strictly spoken, evidence may thus only be generalized to real world settings if all factors operating in the legal system have been taken into account (as is the case in field studies). With regard to the present phenomenon, for instance, it remains still an open question of whether such inconsistencies have an effect upon the perceived credibility of a witness in a real trial, let alone on the verdicts (see also Konečni & Ebbesen, 1984).

The aim of the present study may thus be understood as the first step of a gradual approximation of real world circumstances: it was tested whether the observation-estimation gap, that was previously only demonstrated under artificial encoding condition, replicates under naturalistic encoding conditions. The present findings indeed showed that eyewitness performance was still underestimated by future actors of the legal system and this was particularly true for reminiscent items, which still seem to represent a stronger violation of laypersons' intuition. There is still a long way to go, however, to ensure general ecological validity.

Nevertheless, this leads us back to the fundamental concern of this paper that inspired this research: the actors in the legal system, who evaluate, question, defend, and judge the recollections of others almost every day, do so without any expertise regarding the issue at hand. This notion has been pointed out and empirically supported before (e.g., Benton, Ross, Bradshaw, Tomas, & Bradshaw, 2006; Magnussen et al., 2008) and 75% of the judges surveyed by Wise and Safer (2003) expressed a need for more training on eyewitness testimony. Lacking professional knowledge about human memory, however, likely compels decision makers to rely on their intuition instead (except for those cases, in which expert testimony is sought). This paper is not the first to suggest that such intuitions (or expectations) may be systematically wrong. The literature on wrongful convictions speaks to the same argument, although it is suggestive of deviations in the opposite direction—an *over*estimation of witnesses' capabilities (e.g., Huff, 1996; Levi, 1998). This makes clear that there is a fine line and I do not intend to advocate a more positive evaluation of eyewitness memory in general when pointing to the underestimation of actual eyewitness accounts. Rather, more research is needed with regard to discrepancies between scientifically approved knowledge about (eyewitness) memory on the one hand and the beliefs and expectations about memory that are held by decision makers within the legal system on the other. And crucially, it needs ecologically valid research about this (Konečni & Ebbesen, 1992) as well as potential countermeasures (e.g., Wise, Dauphinais, & Safer, 2007).

Conclusion

Because expectations guide the evaluation of eyewitnesses' accounts in decision-makers within the legal system, it is very important to learn more about these expectations. The present study suggests that these expectations might be wrong in case of reminiscent details. At the same time, differences were smaller when it came to an identification task. Further research is needed in order to shed more light on this important issue.

Conflict of Interest

The author of this article declares no conflict of interest.

Acknowledgements

I am indebted to Isabel Lindner for her valuable comments on earlier versions of this paper.

References

Ballard, P. B. (1913). Obliviscence and reminiscence. *British Journal of Psychology Monograph Supplements*, *1*, 1–82.

Baugerud, G. A., Magnussen, S., & Melinder, A. (2014). High accuracy but low consistency in children's long-term recall of a real-life stressful event. *Journal of Experimental Child Psychology*, *126*, 357–368. http://dx.doi.org/10.1016/jecp.2014.05.009

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior*, *25*, 475–491, doi: 10.1023/A: 1012840831846.

Benton, T. R., Ross, D. F., Bradshaw, E., Thomas, W. N., & Bradshaw, G. S. (2006). Eyewitness memory is still not common sense: Comparing jurors, judges and law enforcement to eyewitness experts. *Applied Cognitive Psychology*, *20*, 115–129. http://dx.doi.org/10.1002/acp.1171

Berman, G. L., & Cutler, B. L. (1996). Effects of inconsistencies in eyewitness testimony on mock-juror decision making, *Journal of. Applied Psychology*, *81*, 170–177. http://dx.doi.org/10.1037/0021-9010.81. 2.170

Bluck, S., Levine, L. J., & Laulhere, T. M. (1999). Autobiographical remembering and hypermnesia: A comparison of older and younger adults. *Psychology and Aging*, *14*, 671–682. http://dx.doi.org/10.1037/0882-7974.14.4.671

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instruction, foil similarity and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30. http://dx.doi.org/10.1037/1076-898X.12.1.11

Brock, P., Fisher, R. P., & Cutler, B. L. (1999). Examining the cognitive interview in a double-test paradigm. *Psychology, Crime & Law*, *5*, 29–45. http://dx.doi.org/10.1080/10683169908414992

Buschke, H. (1974). Spontaneous remembering after recall failure. *Science*, *184*, 579–581. http://dx.doi.org/10.1126/science.184.4136.579

Campbell, J., Nadel, L., Duke, D., & Ryan, L. (2011). Remembering all that and then some: Recollection of autobiographical memories after a one-year delay. *Memory*, *19*, 406–415. http://dx.doi.org/10.1080/09658211.2011.578073

Dunning, D., & Stern, L. B. (1992). Examining the generality of eyewitness hypermnesia: A close look at time delay and question type. *Applied Cognitive Psychology*, *6*, 643–657. http://dx.doi.org/10.1002/acp.2350060707

Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur Experimentellen Psychologie* [Memory: A contribution to experimental psychology]. Leipzig, Germany: Duncker & Humblot.

Ellison, L. (2001). The mosaic art? Cross-examination and the vulnerable witness. *Legal Studies*, *21*, 353–375, doi: 10.1111/j.1748-121X. 2001.tb00172.x.

Erdelyi, M. H. (2010). The ups and downs of memory. *American Psychologist*, *65*, 623–633. http://dx.doi.org/10.1037/a0020440

Fariña, F., Arce, R., & Real, S. (1994). Ruedas de identificación: De la simulación y la realidad [Lineups: A comparison of high fidelity research and research in a real context.]. *Psicotema*, *7*, 395–402.

Fisher, R. P., Brewer, N., & Mitchell, G. (2009). The relation between consistency and accuracy of eyewitness testimony: Legal versus cognitive explanations. In T. Williamson, R. Bull, & T. Valentine (Eds.), *Handbook of psychology of investigative interviewing: Current developments and future directions* (pp. 121–136). Chichester, UK: John Wiley.

Florida Supreme Court Standard Jury Instructions 3d. 2009. Retrieved from www.floridasupremecourt.org/[Februrary 1 2011].

Gilbert, J. A. E., & Fisher, R. P. (2006). The effects of varied retrieval cues on reminiscence in eyewitness memory. *Applied Cognitive Psychology*, *20*, 723–739. http://dx.doi.org/10.1002/acp1232

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior*, *33*, 70–82. http://dx.doi.org/10.1007/s10979-008-9136-x

Huff, R. C. (1996). *Convicted but Innocent: Wrongful Conviction and Public Policy*. Thousand Oaks, CA: Sage.

Ihlebæk, C., Løve, T., Eilertsen, D. E., & Magnussen, S. (2003). Memory for a staged criminal event witnessed live and on video. *Memory*, *11*, 319–327. http://dx.doi.org/10.1080/09658210244000018

Innocence Project (2012, April 11). Retrieved from http://www.innocenceproject.org/understand/Eyewitness-Misidentification.php

Konečni, V. J., & Ebbesen, E. B. (1984). The mythology of legal decision making. *International Journal of Law and Psychiatry*, *7*, 5–16.

Konečni, V. J., & Ebbesen, E. B. (1992). Methodological issues in research on legal decision-making, with special reference to experimental simulations. In F. Lösel, D. Bender, & T. Bliesner (Eds.), *Psychology and law: International perspectives* (pp. 413–423). Berlin: de Gruyter.

Leippe, M. R., & Romanczyk, A. (1989). Reactions to child (versus adult) eyewitnesses: The influence of juror's preconceptions and witness behavior. *Law and Human Behavior*, *13*, 103–132. http://dx.doi.org/10.1007/BF01055919

Levi, A. M. (1998). Are defendants guilty if they were chosen in a lineup? *Law and Human Behavior*, *22*, 389–407, doi: 10.1023/A: 1025718909499.

Magnussen, S., Wise, R. A., Raja, A. Q., Safer, M. A., Pawlenko, N., & Stridbeck, U. (2008). What judges know about eyewitness testimony: A comparison of Norwegian and US judges. *Psychology, Crime & Law*, *14*, 177–188. http://dx.doi.org/10.1080/10683160701580099

McCloskey, M., & Egeth, H. E. (1983). Eyewitness identification: What can a psychologist tell a jury? *American Psychologist*, *38*, 550–563. http://dx.doi.org/10.1037//0003-066X.38.5.550

Nadel, L., Campbell, J., & Ryan, L. (2007). Autobiographical memory retrieval and hippocampal activation as a function of repetition and the passage of time. *Neural Plasticity*,. Article 90472. doi: 10.1155/2007/90472.

Oeberst, A. (2012). If anything else comes to mind… better keep it to yourself? Delayed recall is discrediting – unjustifiably. *Law and Human Behavior*, *36*, 366–374. http://dx.doi.org/10.1007/s10979-011-9282-4

Pozzulo, J. D., Crescini, C., & Panton, T. (2008). Does methodology matter in eyewitness identification research?: The effect of live versus video exposure on eyewitness identification accuracy. *International Journal of Law and Psychiatry*, *31*, 430–437. http://dx.doi.org/10.1016/j.ijlp.2008.08.006

Prager, I. R., Moran, G., & Sanchez, J. (1996). Job analysis of felony assistant public defenders: The most important tasks and most useful knowledge, skills, and abilities. *Psychology, Crime & Law*, *3*, 37–49. http://dx.doi.org/10.1080/10683169608409793

Read, J. D., Tollestrup, P., Hammersley, R., McFadzen, E., & Christensen, A. (1990). The unconscious transference effect: Are innocent bystanders ever misidentified? *Applied Cognitive Psychology*, *4*, 3–31. http://dx.doi.org/10.1037/0021-9010.79.6.918

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence-accuracy correlation. *Journal of Applied Psychology*, *81*, 587–594. http://dx.doi.org/10.1037/0021-9010.81.5.587

Roebers, C. M., Gelhaar, T., & Schneider, W. (2004). It's magic!" The effects of presentation modality on children's event memory, suggestibility, and confidence judgments. *Journal of Experimental Child Psychology*, *87*, 320–335. http://dx.doi.org/10.1016/j.jecp.2004.01.004

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual Innocence: Five Days to Execution and Other Dispatches from the Wrongly Convicted*. New York, NY: Doubleday.

Schmitt, K. L., & Anderson, D. R. (2002). Television and reality: Toddlers' use of visual information from video to guide behavior. *Media Psychology*, *4*, 51–76. http://dx.doi.org/10.1027/S1532785XMEP0401_03

Scrivner, E., & Safer, M. A. (1988). Eyewitnesses show hypermnesia for details about a violent event. *Journal of. Applied Psychology*, *73*, 371–377. http://dx.doi.org/10.1037/0021-9010.73.3.371

Sixth Circuit Criminal Pattern Jury Instructions. (2005). Retrieved from www.ca6.uscourts.gov/internet/crim_jury_insts.htm/on April 12, 2011.

Thierry, K. L., & Spence, M. J. (2004). A real-life event enhances the accuracy of preschoolers' recall. *Applied Cognitive Psychology*, *18*, 297–309. http://dx.doi.org/10.1002/acp.965

Tollestrup, P., Turtle, J. W., & Yuille, J. C. (1994). Actual victims and witnesses to robbery and fraud: An archival analysis. In F. D. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult Eyewitness Testimony. Current Trends and Developments* (pp. 144–160). New York, NY: Cambridge University Press.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: effects of associative context. *Journal of Experimental Psychology*, *87*, 116–124. http://dx.doi.org/10.1037/h0030186

Turtle, J. W., & Yuille, J. C. (1994). Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hypermnesia. *Journal of Applied Psychology*, *79*, 260–271. http://dx.doi.org/10.1037/0021-9010.79.2.260

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science*, *7*, 45–75. http://dx.doi.org/10.1111/j.1529-1006.2006.00027.x

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*, 155–167, doi: 101037/1076-898X.8.3.155.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*, 1–39, doi: 10.1023/A: 1025750605807.

Wise, R. A., Dauphinais, K. A., & Safer, M. A. (2007). A tripartite solution to eyewitness error. *The Journal of Criminal Law & Criminology*, *97*, 807–871.

Wise, R. A., & Safer, M. A. (2003). A survey of judges' knowledge and beliefs about eyewitness testimony. *Court Review*, *40*, 6–16.

Yuille, J. C., & Cutshall, J. L. (1986). A case study of eyewitness memory of a crime. *Journal of. Applied Psychology*, *71*, 291–301. http://dx.doi.org/10.1037/0021-9010.71.2.291

Yuille, J. C., & Wells, G. L. (1991). Concerns about the application of research findings: The issue of ecological validity. In J. Doris (Ed.), *The Suggestibility of Children's Recollections: Implications for Eyewitness Testimony (p* (pp. 118–128). Washington, DC: American Psychological Association.