The European Journal of Psychology Applied to Legal Context (2021) 13(2) 99-110

The European Journal of Psychology Applied to Legal Context

https://journals.copmadrid.org/ejpalc

Reality Monitoring: A Meta-analytical Review for Forensic Practice

Yurena Gancedo[a], Francisca Fariña[b], Dolores Seijo[a], Manuel Vilariño[a], and Ramón Arce[a]

[a]*Universidad de Santiago de Compostela, Spain;* [b]*Universidad de Vigo, Spain*

ARTICLE INFO

Article history:
Received 2 May 2021
Accepted 2 June 2021

Keywords:
Imagined memories
Perceived memories
Forensic assessment
Witness credibility
Reality monitoring

A B S T R A C T

Reality Monitoring (RM) criteria has been proposed as a forensic tool in order to discern between perceived and imagined memories. However, no systematic evidence has been provided on its validity for use in testimony evaluation. Thus, a meta-analytic review was designed to study its validity in forensic setting. A total of 40 primary studies were found, yielding 251 effect sizes. Random-effects meta-analyses correcting the effect size for sampling error and criterion unreliability were performed. The results showed that the total RM score discriminated, d = 0.542 (δ = 0.562), between imagined and perceived memories of events. In relation to individual criteria, the results showed support for the model's predictions (more external attributes in perceived memories) for clarity, d = 0.361 (δ = 0.399), sensory information, d = 0.359 (δ = 0.397), spatial information, d = 0.250 (δ = 0.277), time information, d = 0.509 (δ = 0.563), reconstructability of the story, d = 0.441 (δ = 0.488), and realism, d = 0.420 (δ = 0.464), but not for affective information, d = 0.024 [-0.081, 0.129]. Nevertheless, except for temporal information, the results are not generalized (negative effects may be found). For cognitive operations, the results corroborated, although the magnitude of the effect was lower than small, the hypothesis (more cognitive operations in imagined memories), d = -0.107 [-0.178, -0.036] (δ = -0.119). The moderating effects of age (more cognitive operations on imagined memories in adults, and on perceived memories in underage), evocation type (external attributes discern between imagined and perceived memories, in both self-experienced and non-experimented accounts), and criteria score (the results varied by score) moderators were studied. As conclusions, forensic implications for the validity of the RM technique in court proceedings are discussed.

Reality Monitoring: una revisión meta-analítica para la práctica forense

Palabras clave:
Memorias imaginadas
Memorias percibidas
Evaluación forense
Credibilidad del testimonio
Reality monitoring

R E S U M E N

Los criterios del *Reality Monitoring* (RM) han sido propuestos como una herramienta forense para discriminar entre memorias percibidas e imaginadas. Sin embargo, no se han facilitado pruebas sistemáticas de su validez para su uso en la evaluación del testimonio, motivo por el cual se planificó una revisión metaanalítica para estudiar su validez en el contexto forense. Se encontró un total de 40 estudios primarios, de los que se extrajeron 251 tamaños del efecto. Se llevaron a cabo meta-análisis de efectos aleatorios que corregían el tamaño del efecto por el error de muestreo y la falta de fiabilidad del criterio. Los resultados mostraron que la puntuación total en el RM discriminaba, d = 0.542 (δ = 0.562), entre memorias de eventos imaginados y percibidos. En relación con los criterios, los resultados avalaron las predicciones del modelo (más atributos externos en memorias percibidas) en los criterios claridad, d = 0.361 (δ = 0.399), información sensorial, d = 0.359 (δ = 0.397), información espacial, d = 0.250 (δ = 0.277), información temporal, d = 0.509 (δ = 0.563), reconstrucción de la historia, d = 0.441 (δ = 0.488), y realismo, d = 0.420 (δ = 0.464), pero no para el criterio información afectiva, d = 0.024 [-0.081, 0.129]. Sin embargo, excepto para el criterio información temporal, los resultados no son generalizables (se pueden hallar efectos negativos). Para el criterio operaciones cognitivas, los resultados corroboraron, aunque la magnitud del efecto era menor que pequeña, la hipótesis (más operaciones cognitivas en memorias imaginadas), d = -0.107 [-0.178, -0.036] (δ = -0.119). Se estudiaron como moderadores los efectos de la edad (más operaciones cognitivas en memorias imaginadas en adultos y en memorias percibidas en menores de edad), tipo de evocación (los atributos externos discernen entre memorias imaginadas y percibidas, tanto en relatos experimentados por uno mismo como no experimentados) y la puntuación del criterio (los resultados difirieron según la puntuación del criterio). Se comentan las implicaciones de los resultados de cara a la validez del RM como técnica forense en los procedimientos judiciales.

Cite this article as: Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality monitoring: A meta-analytical review for forensic practice. *The European Journal of Psychology Applied to Legal Context, 13*(2), 99-110. https://doi.org/10.5093/ejpalc2021a10

Funding: This research has been sponsored by a grant of the Spanish Ministry of Economy, Industry and Competitiveness (PSI2017-87278-R) and by a grant of the *Consellería de Cultura, Educación e Ordenación Universitaria, Xunta de Galicia* (ED431B 2020/46). Correspondence: ramon.arce@usc.es (Ramón Arce).

The verisimilitude attributed to witnesses has been, and keep being, the cornerstone of the vast majority of judicial cases, especially in crimes committed in the private sphere (e.g., sexual offences or family violence). This is so because prosecution's evidence is often reduced to the testimony of the complainant and the evaluation of the harm to the complainant. As the burden of proof corresponds to prosecution and although the testimony of the complainant may be sufficient evidence for conviction, it is usually not sufficient because there may be some benefit for the complainant by the conviction of the accused beyond the legitimate interest in conviction, such as revenge, enmity, resentment, an economic motive, or the existence of a previous relationship between complainant and accused (Arce, 2017). This contingency, which is very frequent in criminal cases (Novo & Seijo, 2010), implies that a complainant's testimony is endowed with probative ability with other means of proof. Evaluation of credibility of testimony is the main mean to provide a complainant's testimony with evidential aptitude validating his/her testimony (Novo & Seijo, 2010). A number of techniques (i.e., physiological indicators, nonverbal and paraverbal indicators, content analysis of statements) and with different objectives (i.e., correctly classify the truth or the lie) have been developed in this regard. The techniques aimed at classifying lies in the testimony have been judicially ruled out since they do not fulfill the task of providing plaintiff's testimony with evidentiary capacity (burden of proof) and in its application to an accused, because he/she has the right not to testify against himself/herself and not to confess guilt (e.g., Art. 24.2 of the Spanish Constitution) and, most importantly, a false testimony of the accused does not prove his/her guilt. In short, only knowledge and techniques based on scientific evidence classifying testimonies as true and referring to the testimony of the complainant have forensic validity. In light of this, the results of the investigation regarding classification of lies in a defendant's testimony have no forensic value, so that physiological evidence, as well as non-verbal and para-verbal indicators associated with lying, are not valid. Furthermore, they have not been really effective in classifying lies either (Sporer & Schwandt, 2006, 2007). Besides, the content analysis of testimonies has been effective in discriminating between memories of lived events (truth) and fabricated memories of events, as well as in classification of memories of lived events (Amado et al., 2015; Amado et al., 2016; Oberlader et al., 2016; Vrij et al., 2021). Two approaches have been formulated, tested, and used commonly in forensic practice for content analysis, one based on reality criteria (Criteria Based Content Analysis - CBCA; Steller & Köhnken, 1989) that are associated with a memory of actual experiences, and the other based on memory attributes or characteristics (Reality Monitoring - RM; Johnson & Raye, 1981) that allow discerning between memories of internal origin (memories derived from thoughts) and external (memories from perceptual experiences). Both approaches share the study of memory as primary register and that its objective is, based on memories content analysis, classification of "real" memories (i.e., resulting from outside perceptual experiences) of "past" (forensic task in contrast to reality testing centered in present perception) acts or events or discrimination between memories of real past acts or events and memories of fabricated or imagined past acts or events. CBCA, which is based on the Undeutsch hypothesis (the memory of truthful accounts of events differ significantly and noticeably in content and quality from false accounts) and is an update of the Statement Reality Analysis – SRA; Undeutsch, 1967, 1982), is part of a forensic technique, SVA, that defines the protocol to be followed for the application of the technique (case file analysis, semi-structured interview, statement content analysis with CBCA criteria, and validity checklist). In this way, the SVA/CBCA adjusts to the demand of justice: to provide supportive evidence of the complainant's testimony. Although the authors did not provide scientific evidence of the validity of CBCA categories of reality for the classification of true testimony, the subsequent literature did, as it is systematically deduced from meta-analytic reviews (Amado et al., 2015; Amado et al., 2016; Oberlader et al., 2016). Succinctly, albeit authors have not provided empirical support for the validity of reality criteria to classify true accounts and to discriminate between false and true accounts, these are equally valid for all types of memories of events (criminal types, events), populations (children, adults, women, men), and testimonies (victim/complainant, eyewitness, accused). In short, the Undeutsch hypothesis has passed from a hypothesis to a scientific truth. However, CBCA, as a measurement instrument, does not comply with psychometric characteristics of reliability and validity (Amado et al., 2015; Amado et al., 2016), nor with the judicial and law of precedent criteria required to a forensic evidence (i.e., error rate is unknown, it does not guarantee compliance with the principle of presumption of innocence, an objective decision rule is not provided, it does not evaluate persistence, it does not prescribe how the statement is obtained and, hence, does not include guarantees that the test was obtained lawfully; Arce, 2017; Daubert vs. Merrell Dow Pharmaceuticals, 1993).

Furthermore, the Reality Monitoring model aims to identify processes used by people to decide whether information (memory) has an internal (imagined) or external (perceived) origin. As for this, Johnson and Raye (1981) defined attributes that characterize a memory of external origin (external memory attributes: contextual information, sensory information, and semantic information) and internal origin (internal memory attributes: cognitive operations, i.e., thoughts, reasoning), and created an instrument for subjects to evaluate their imagined and perceived memories, the Memory Characteristics Questionnaire (MCQ), consisting of 39 items (Johnson et al., 1988). Originally, each item was taken as an attribute to discern between memories, but Suengas and Johnson (1988), after observing that the items could be grouped, factorialized (main components, N = 144) the instrument, identifying 5 composite factors (they refer to composite factors as they carried out two separate factor analyses for memories of perceived events –seven factors– and imagined events –six factors– creating composite factors with those that were more or less common to both memories): clarity, sensory information, contextual information, thoughts and feelings, and intensity of feelings. Schooler et al. (1986) applied a content analysis to differentiate between suggested memories and real memories of witnesses, taking two categories from Reality Monitoring (i.e., sensory information and cognitive processes). Alonso-Quecuty (1992) made the final leap into the field of testimony, applying a categorical content analysis system based on Johnson and Raye's (1981) model (i.e., sensory information, contextual information, idiosyncratic information) to which the declaration length was added to differentiate between true (external origin) and false (internal origin) statements. Sporer and Küpper (2004) published (the study was carried out in 1994 and presented as a paper at a congress) a new factorialization (N = 100) of the MCQ scale, finding 8 factors (i.e., clarity, sensory information, spatial information, time information, affective information, reconstructability, realism, and cognitive information), suggesting two applications of it: one, in line with the original proposal by Johnson and Raye (1981), for self-evaluations of the origin of memory (Self-ratings of Memory Characteristics Questionnaire - SMCQ), and another for the assessment of others memory (Judgment of Memory Characteristics Questionnaire - JMCQ). Actually, it was not exactly the result of a robust exploratory factor analysis (N = 100, with a ratio between subjects and items of 2.56), so the results were corrected to fit the factors to the theoretical model. In fact, the realism factor was comprised only for 1 item (a factor cannot be made up of less than 2 items, since a correlation of a single measure cannot be obtained, that is, internal consistency); thus, Sporer and Hamilton (1996) added 4 items to the questionnaire referring to this factor (the more items a factor has, the greater the reliability; Cronbach, 1951). However, Sporer and Sharman (2006) reduced the scale to 42, disappearing item 43 (believability) and

reassigning items (e.g., item 24, which was in the affect factor, becomes realism). In any case, the factors are maintained, but not the items that compose them. Finally, Vrij et al. (2004a, 2004b) proposed a model with 4 categories: visual, auditory, temporal, and spatial details. In sum, there are no standardized criteria that make up RM, but rather different classifications of RM criteria.

All these models and categories of analysis have been tested in investigation designs of judging other people's memory (forensic task), being the state of the question synthesized in narrative and meta-analytic reviews. Among the first, Sporer (2004) concluded that RM is as valid as CBCA, being clarity, temporal information, and realism criteria the most effective; Masip et al. (2005) stated that results are not conclusive, although contextual information and realism seem to be the criteria that best discriminate; and Vrij (2008) contended that results are not clear. In meta-analytic reviews, DePaulo et al. (2003) found a significant effect size for the criterion realism (d = -0.42, less in liar accounts, k = 1) and not significant for sensory information (d = -0.17, k = 4), idiosyncratic information (d = 0.01, k = 2), clarity (d = -0.01), reconstructability (d = -0.01, k = 1), and cognitive processes (d = 0.91, k = 1). In any case, results are not robust because there is insufficient k (< 3) or N (< 400). Finally, Oberlader et al. (2016) encountered that the total score in RM criteria discerned significantly and with an effect size greater than large (d/g = 1.26) between truthful and fabricated statements.

In this state of the art, a meta-analytic review with the aim of testing the validity of the RM model for discrimination between memories (global score in the RM) and of the different models (total score in the original criteria of Johnson & Raye, 1981; Sporer & Küpper, 2004; and Vrij et al., 2004a, 2004b) as well as of each one of the categories of analysis to know the adequacy of each category to the hypothesis of the origin of memories and to compare them between them, was set out. Additionally, and in the case of observing heterogeneity in the distribution of studies, the effects of moderators investigated in the literature of interest for forensic practice and evaluation will be studied: age group of participants (adults, younger children, and older children; Roberts & Lamb, 2010), type of evocation (self-experienced and non-experimented events; Monteiro et al., 2018), and scoring of criteria (scoring scales, categorical measure –presence vs. absence –, and frequency/density; Arce, 2017; Masip et al., 2005; Sporer, 2004; Vrij, 2008).

Method

Search for Studies

The bibliographic search focused on identifying those studies addressing the effectiveness of RM to differentiate between memories of perceived and imagined events. For this, at first the systematic and meta-analytic reviews already existing on this instrument were identified, as well as the primary studies that they include. Next, a search was carried out using the terms "reality monitoring approach", "reality monitoring", and "source monitoring", both independently (OR command) and in combination (AND command) with "testimony", "statement", "witness", "credibility", "perceived memory"," fabricated memory", "invented memory", and "imagined memory", in scientific reference databases (Web of Science, Scopus, PsycInfo and Dialnet), in the doctoral dissertation database Proquest Dissertations & Theses, as well as in the meta search-engine Google Scholar. To these initial descriptors, those identified in the sources (e.g., self-experienced accounts, invented accounts) were added until an exhaustive search was completed. The inclusion criteria were that: a) full text was available; b) analyzed protocols were testimonies; c) perceived and invented events be compared; d) criteria derived from the theory of RM be used to determine the internal or external origin of the account; e) the publication was in a medium subject to peer review or was a doctoral thesis (scientific evidence, Daubert standard); and f) the effect size be provided or, failing that, sufficient data to calculate it. Likewise, the following exclusion criteria were used: a) participants performed self-ratings of their memory; b) the study was part of a training plan (e.g., End-of-degree Project, Master degree thesis); and c) unpublished manuscripts. Applying this search strategy and the selection and exclusion criteria, we selected 40 primary studies (see flow diagram in Figure 1), from which 251 effect sizes were obtained.

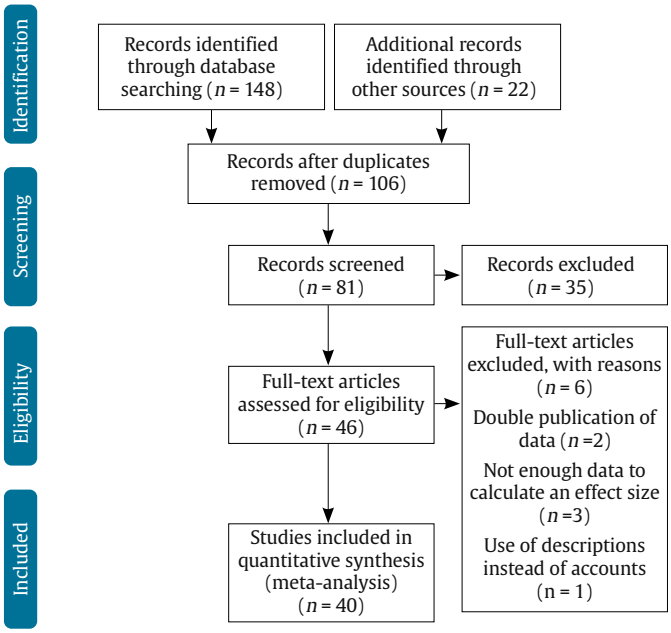

Figure 1. Flow Diagram of the Meta-analysis.

Coding of Primary Studies

The studies were coded according to the following categories: a) primary study reference; b) document type (article, doctoral thesis, proceeding paper, book, unpublished study); c) sample characteristics (i.e., age, gender, size); d) type of exposure to the reported event (real experience vs. video); e) evaluated criteria; f) scoring of criteria; and g) effect size or, where appropriate, the data necessary to calculate it. Two experienced and trained raters coded independently the studies in the precedent categories. After 10 days of the original coding, each rater repeated 50% of the coding of the studies (within-rater concordance). The between- and within-rater concordance was assessed in true kappa ($\bar{k}$; Fariña et al., 2002). Kappa corrects the concordance for the random agreement. Nevertheless, a systematic source of error is not controlled: the correspondence between coding (true kappa). Succinctly, if the exact correspondence is not verified, two errors may be encoded as an agreement. This correction is called true kappa. The results showed a total concordance ($\bar{k}$ = 1). Additionally, these raters were consistent in other studies, i.e., in other contexts (Fariña et al., 2017). Thus, verified between- and within-rater and inter-contexts concordance, the coding was reliable, i.e., another(s) trained rater(s) would find the same results (Wicker, 1975).

Data Analysis

The effect sizes were standardized in d, taking: a) from the primary study, when the d value was available (if data were provided for its

Table 1. Meta-analysis for the Total Reality Monitoring Score

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	$SD_δ$	% Var	95% CI_d	80% $CI_δ$
Total Score (any total score)										
16	1,696	0.546	0.3439	0.1985	0.2807	0.562	0.2884	33.44	0.449, 0.643	0.193, 0.931
Sporer & Küpper's (2004) Total Score (8 criteria)										
7	563	0.462	0.3132	0.2269	0.21604	0.475	0.2218	52.55	0.294, 0.630	0.191, 0.759
Original Total Score (4 criteria)										
3	224	0.617	0.0661	0.2379	0.0000	0.634	0.0000	100[1]	0.348, 0.886	0.634, 0.634
Vrij et al.'s (2004a, 2004b) Total Score (4 criteria)										
2	376	0.801	0.2536	0.1518	0.2031	0.813	0.2061	35.97	0.590, 1.012	0.594, 1.077

Note. *k* = number of effect sizes; *N* = total sample size; d_w = sample size weighted mean effect size; SD_d = standard deviation of *d*; SD_{pre} = standard deviation predicted for sampling error alone; SD_{res} = standard deviation of *d* after removing sampling error variance; δ = mean true effect size; $SD_δ$ = the standard deviation of δ; % Var = percent of observed variance accounted by artifactual errors; 95% CI_d = 95% confidence interval for *d*; 80% $CI_δ$ = 80% credibility interval for δ.
[1]The predicted variance overcomes the observed variance, rounding it to 100%.

calculation in the primary study, its accuracy was verified, as well as the application of the formula of Cohen, Hedges, or Glass when applicable for between designs and *d* average for within designs, and correcting bias effect size –Hedges's correction–); b) in those studies that did not provide this data, but did provide the mean and standard deviation (or alternatively, the standard error or variance) as well as the *N*s of the perceived memory group and imagined memory group, *d* was calculated with the formula Cohen's when $N1 = N2$, with Hedges's *g* when $N1 \neq N2$ and with Glass's Δ when the assumption of homogeneity of variances was violated for between designs and with *d* average for within designs, correcting bias effect size –Hedges's correction–; c) in studies in which the effect size was provided by another estimator (e.g., *r*, η^2), this was converted to *d*; d) when the value of *t* or *F* was well provided and the degrees of freedom were obtained *d* from these; and e) when there was more than one effect size in the same experiment (experimental manipulations with the same subjects) the combined means and variances were calculated and from there *d* was obtained. The authors created an excel spreadsheet for all the computations that were verified for the correctness of their operation by contrasting it with a manual execution.

Next step was to perform a meta-analysis of random effects correcting the effect size by the sample error, and the unreliability criterion (Schmidt & Hunter, 2015). Thus, two different mean effect sizes were computed in each meta-analysis: *d* (bare-bones procedure: correcting for sampling error alone) and δ (correcting *d* for criterion unreliability). As for this, the following statistics were calculated: effect size weighted for sampling error (d_w); standard deviation of *d* (SD_d); standard deviation of *d* predicted by artifactual errors (SD_{pre}); standard deviation of *d*, after removal of variance due to artifactual errors (SD_{res}); mean true effect size, corrected for criterion unreliability (δ); standard deviation of δ ($SD_δ$); variance accounted by artifactual errors (% Var); 95% confidence interval for *d* (95% CI_d); and 80% credibility interval for δ (80% $CI_δ$). If the confidence interval has no zero, it reported the effect size was significant. If the credibility interval has no zero, it confirmed it encompassed 80% of potential studies on the same population, meaning 90% of all the studies would be above the lower limit. If artifactual variance (% Var) explained the bulk of the variance, > 75% (75% rule; Hunter et al., 1982), then non-explained variance was not systematic (homogeneous data). Conversely, if it explained less than 75%, unexplained variance is due to moderators (heterogeneous data). Formulas were taken from Schmidt and Hunter (2015). Though *d* and δ mean effect sizes are valuable for deriving implications for forensic practice, additional analyses were performed to complement them: the study of cases and comparison of effects (raw effects were computed in the same measure). As for the study of cases, the probability of an inferiority score (PIS; Gallego et al., 2019; Redondo et al., 2019) was performed to know error probability in classifying an imagined memory as perceived memory (non-admissible error in forensic task as it infringes the principle of presumption of innocence). Overlapping the distributions of two populations, it consists of an estimation of the probability of obtaining in the interest population a score below the mean of the contrast population. The magnitude of the effects was interpreted in terms of Cohen's (1988) small, medium, and large, corresponding to a PS_{ES} of .556, .637, and .716) categories, adding a supplementary one, a more than large effect size ($d/δ$ > 1.20, corresponding to a PS_{ES} of .802, i.e., an effect size larger than 80.2% of all possible and than 60.4 of the positive or negative ones; Arce et al., 2015) and quantified in terms of the probability of superiority of the effect size (PS_{ES}; Arce et al., 2020; Arias et al., 2020). It consists of converting the effect size to a percentile. Comparisons of effect sizes were executed computing q (Cohen, 1988).

Criterion Reliability

Inter-rater reliability (*r*) was not reported in all studies and some primary papers reported agreement instead of reliability. On the basis of the lack of data about coding reliability in all studies, an average reliability was estimated for the criteria and for the total RM score due to the fact that reliability for the instrument (total score) and criteria is different. As for the total RM score, reliability was estimated with Spearman-Brown prophetic formula, obtaining an *r* of .947 (*SD* = .046), whereas reliability for the individual criteria was calculated using the reliability coefficients of each study obtaining an average *r* of .822 (*SD* = .148).

Results

Analysis of Atypical Values

Data were explored in search of extreme values (± 3 * IQR), outliers (± 1.5 * IQR) and abnormal with the application of Chauvenet's criterion (± 2 *SD*). For this, the sizes were segmented into: sizes for total RM score, sizes of external criteria, and sizes of internal criteria. In total RM score, no extreme, outlier, or abnormal values were found. An extreme value was found in internal memory criteria and was eliminated (Krackow, 2010). Finally, the exploration of the distribution of effect sizes in external criteria again identified Krackow's study as an extreme value, 5 outliers and 3 out of the range of the Chauvenet's criterion. The extreme value was eliminated because it was observed that they were not due to the effect of a moderator (Tukey, 1960); 4 of the outlier values were found to be inconvenient results (Arce et al., 2020) and explainable by moderators (explored segmented by moderators are no longer outliers) and 1 outlier (Santtila et al., 1998) in line with the hypothesis that it affected one of the 8 effect sizes of the study and it was not observed that it was the consequence of a moderator,

Table 2. Meta-analysis of the Reality Monitoring Categories

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
Clarity (and Vividness)										
7	700	0.361	0.8684	0.2024	0.8445	0.399	0.9335	5.53	0.058, 0.364	-0.796, 1.594
Sensory Information										
18	1,092	0.359	0.6564	0.2602	0.6026	0.397	0.6658	15.88	0.061, 0.239	-0.455, 1.249
Spatial Information										
25	2,290	0.250	0.6507	0.2106	0.6157	0.277	0.6806	10.55	0.168, 0.332	-0.594, 1.148
Time Information										
23	2,290	0.509	0.4326	0.2044	0.3813	0.563	0.4196	23.08	0.426, 0.592	0.026, 1.100
Affective Information										
15	1,397	0.024	0.6112	0.2081	0.5746	0.027	0.6355	11.59	-0.081, 0.129	-0.786, 0.840
Reconstructability of the Story										
5	463	0.441	0.4333	0.2110	0.3784	0.488	0.4170	24.28	0.256, 0.626	-0.046, 1.022
Realism										
9	908	0.420	0.6841	0.2020	0.6536	0.464	0.7221	8.92	0.288, 0.552	-0.460, 1.388
Cognitive Information (cognitive operations, idiosyncratic information)										
33	3,052	-0.107	0.6119	0.2089	0.5751	-0.119	0.6360	11.68	-0.178, -0.036	-0.933, 0.695

so it was removed. It was contrasted that the 3 outliers outside the range of Chauvenet's criterion were inconvenient results explained by moderators.

Study of the Total Reality Monitoring Score

The results of the meta-analysis for the total RM score (see Table 1) revealed a significant (when the confidence interval has no zero, indicating the effect size was significant), positive (higher scores in memories of perceived events in comparison with memories of imagined events), generalizable (the lower limit of credibility interval is 0.193, indicating the minimum expected effect size for 90% of any other study would be beyond 0.193), and medium magnitude (δ > 0.5; an effect size above 31.1.6%, PS_{ES} = .311) mean true effect size (δ). The margin of error (probability of a higher score in memories of imagined events than in memories of perceived events) of the total RM score would be of 28.7 (PIS = .287; classification of an imagined memory as a perceived one). Moreover, the percentage of explained variance for artifactual errors is less than 75%, indicating heterogeneity between primary studies. Thus, and given that the total score included different groupings of criteria (models), it was determined to study the models as a moderator.

Study of the Models of Total Reality Monitoring Score

The results of the meta-analyses of the total score showed a significant, positive, and generalizable mean true effect size for the three models (original, Sporer & Küpper, 2004, and Vrij et al., 2004a, 2004b), and of a magnitude between small and medium magnitude (0.20 > δ < 0.5; an effect size above 26.6% of all positives, PS_{ES} = .266) for the Sporer and Küpper's (2004) model, medium for the original model (δ > 0.5; above 34.7% of all positives, PS_{ES} = .347), and large (δ > 0.8; above 43.1% of all positives, PS_{ES} = .431) for Vrij et al.'s (2004a, 2004b) model. The probability of error was 31.7% (PIS = .317), 26.3% (PIS = .263), and 20.8% (PIS = .208) for Sporer and Küpper's, original and Vrij et al.'s model, respectively. Nevertheless, the results for Sporer and Küpper's and Vrij et al.'s models are explained by moderators (% Var < 75%). As for the original model, the variance explained by artifactual errors was 100%, properly of a second order sampling error, i.e., primary studies were not randomly distributed (insufficient *N* = 224).

Comparatively, the explanatory power of Sporer and Küpper's (2004) model is lower than Vrij et al.'s (2004a, 2004b), $q_s(N' = 451) = 0.161$, $z = 2.41$, $p < .05$. Comparisons for the original model were not performed as primary studies were not randomly distributed. In sum, the addition of 4 criteria from Sporer and Küpper above the Vrij et al.'s criteria (sensory-visual and auditory- spatial and time information) is not reflected in a greater explanatory power of the model.

Study of RM Criteria

For the cognitive operations criterion, the results (see Table 2) exhibited a negative (higher scores in imagined memories) and significant mean effect size. Nonetheless, the magnitude of the effect is lower than small (δ < 0.20; above 6.4% of all negatives, PS_{ES} = .064). In addition, there is heterogeneity between primary studies (% VAR = 13.86), while lower and upper limits of credibility intervals (-0.933 and 0.695) warn that results can be found even with large effect sizes that support the hypothesis (more cognitive information in imagined memories), but also of a close to large magnitude that refutes it (more cognitive information in perceived memories). In terms of practical utility, with the application of this criterion the probability of finding more cognitive information among perceived memories than in imagined memories (error) is of 45.3% (PIS = .453).

Among the criteria of memories of external origin (see Table 2), the results of the meta-analysis confirmed the prediction of the model, that is, a higher score in memories of external origin (positive mean effect size) and significant in clarity, sensory information, spatial information, time information, reconstructability of the story, and realism criteria. In terms of effect magnitude, the mean true effect size was between small and medium (0.20 > δ < 0.50) for clarity (above 22.1% of all positives, PS_{ES} = .221), sensory (above 22.1% of all positives, PS_{ES} = .221), spatial information (above 15.9% of all positives, PS_{ES} = .159), reconstructability of the story (above 27.4% of all positives, PS_{ES} = .274), and realism (above 25.9% of all positives, PS_{ES} = .259), and medium (δ > 0.50) for time information (above 36.2% of all positives, PS_{ES} = .362). Comparatively, the effect size for time information was significantly higher than for sensory, $q_s(N' = 1{,}478) = 0.080$, $z = 2.14$, $p < .05$, and spatial information, $q_s(N = 2{,}290) = 0.138$, $z = 4.67$, $p < .001$; the effect size for reconstructability of the story was significantly higher than for spatial information, $q_s(N' = 769) = 0.103$, $z = 2.01$, $p < .05$, and that the effect size for realism was significantly higher than for spatial information, $q_s(N' = 1300) = 0.092$, $z = 2.34$, $p < .05$. For other comparisons no differences were observed, $z < 1.82$, *ns*. Furthermore, for time information criterion, positive effect sizes are generalizable to the population of studies, i.e., the least expected effect size in studies is positive (the lower limit

Table 3. Meta-analysis of the Sensory Information Subcategories

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
Visual										
12	1,321	0.452	0.3942	0.1937	0.3433	0.500	0.3779	24.86	0.343, 0.561	0.016, 0.984
Auditory										
11	1,197	0.662	0.5057	0.1976	0.4656	0.732	0.5121	16.20	0.546, 0.778	0.077, 1.387

Table 4. Meta-analysis of the Reality Monitoring Memory Attributes in Adults

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
External Memory Attributes										
90	8,990	0.457	0.4779	0.2034	0.4324	0.505	0.4768	18.62	0.415, 0.499	-0.105, 1.115
Internal Memory Attributes										
23	2,041	-0.275	0.5882	0.2142	0.5478	-0.304	0.6055	13.38	-0.362, -0.188	-1.079, 0.471

of the credibility interval is positive). Nevertheless, the probability of classifying memories imagined as perceived (error) was 28.7% for time information (PIS = .287). Conversely, results for clarity, sensory information, spatial information, reconstructability of the story and realism criteria are not generalizable, i.e., negative effects may be found (the lower limit of the credibility interval is negative) with a probability of error of 34.5, 34.6, 39.1, 31.3, and 32.1% (PIS = .345, .346, .391, .313, and .321), correspondingly. Additionally, unexplained variance is due to moderators (% Var < 75).

However, the results do not confirm the prediction of the model in the affective information criterion. Thus, the average effect size, although positive, is not significant (the confidence interval for *d* has zero). In addition, both large positive and negative effect sizes (upper and lower limits of the credibility interval) can be found, variability that is explained by moderators (% Var = 10.82). In terms of practical utility, with the application of this criterion, the probability (error) of finding more affective information among perceived memories than in imagined memories is 48.9% (PIS = .4890).

In relation to the subcriteria of sensory information, meta-analytic results (see Table 3) showed a significant, positive (more sensory information in perceived memories), medium magnitude for visual sensations (an effect size above 27.4% of all positives, PS_{ES} = .274), close to large for auditory sensations (an effect size above 39.7% of all positives, PS_{ES} = .397), and generalizable mean true effect size. Nevertheless, as the percentage of variance explained by artifactual errors was < 75%, the results are influenced by moderators, while the probability of error was 30.9% (PIS = .309) and 23.2% (PIS = .232) for visual and auditory sensations, respectively.

For the subcategories of smell, taste, and physical sensations, meta-analyses could not be calculated due to insufficient primary studies (k = 1). However, the effect sizes observed in smell (d = 0.258 [-0.188, 0.704], δ = 0.285, n = 80, 1 – β = .847), taste (d = 0.040 [-0.404, 0.484], δ = 0.044, n = 80, 1 – β = .925), and physical (d = 0.273 [-0.154, 0.700], δ = 0.301, n = 87, 1 – β = .908) sensations were no significant (confidence interval for d has no zero), that is, data does not support the validity of these criteria, and these categories are practically unproductive (≤ .05, trivial presence).

Comparison of the meta-analytical results exposed a significantly larger effect size for auditory sensations than for visual sensations, $q_s(N' = 1256) = 0.111$, $z = 2.78$, $p < .01$. In short, more auditory than visual information in memories of perceived events is registered.

Moderators Study

Age. On the basis that the quality of an account is directly related to an individual's cognitive and language development (Davies, 1994; DePaulo et al., 2003), it has been hypothesized that, due to the limitations imposed by cognitive and language development, the accounts of lived events will contain lower number of criteria in children than in adults (Roberts & Lamb, 2010; Vrij et al., 2004a). Although, in primary studies there is concordance when referring to adults as people aged ≥ 18 years, there is no uniform criterion for grouping non-adults. As a minimum age, 3 years are taken, but they can reach up to 16; that is, they are classified as non-adults, children, and adolescents. Nevertheless, since the hypothesis that relates age to the quality of the account (productivity of the RM criteria) is based on the limitations imposed by cognitive and language development, it is not equally applicable to all underage. In this regard, a shared criterion in primary studies for classification of children with limitations in cognitive and language development was not found, being Roberts and Lamb's (2010) classification the only one reflected as such: younger (3-8 years) and older (9-16 years) children. As a consequence, meta-analyses for younger and older children were performed.

The results of the meta-analysis for attributes of external memories (it could not be calculated with the total RM score because k was insufficient) in adults (see Table 4) showed a positive, significant, and medium magnitude (an effect size above 28.1% of all positives PS_{ES} = .281) mean true effect size. However, the probability of error was 30.7% (PIS = .307). Moreover, the results are subject to the effect of moderators (% VAR < 75), and negative effects may be found (lower limit for 80% credibility interval was -0.105).

As for internal attributes, the results of the meta-analysis for adults (see Table 4) disclosed a significant, negative (more internal attributes in imagined memories) and small magnitude (an effect size above 22.1% of all negative effects, PS_{ES} = .221) mean true effect

Table 5. Meta-analysis of Reality Monitoring Memory Attributes in Underage

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
External Memory Attributes										
15[1]	403	1.096	1.9769	0.4184	1.9321	1.212	2.1350	4.65	0.886, 1.306	-1.521, 3.945
40[2]	2,261	0.384	0.7002	0.2703	0.6459	0.424	0.7137	15.07	0.301, 0.467	-0.489, 1.337
Internal Memory Attributes										
4[1]	129	0.632	1.1643	0.3643	1.1059	0.699	1.2220	9.95	0.276, 0.988	-0.865, 2.263
7[2]	374	0.222	0.7981	0.2764	0.7487	0.246	0.8279	12.04	0.018, 0.426	-0.814, 1.306

Note. [1]Younger children samples (3-8 years); [2]older children samples (9-16 years).

size. Nonetheless, the probability of error was 38.1% (PIS = .381), and studies are not homogeneous (% Var < 75%), suggesting the presence of moderators of the effect and, conversely, positive effect sizes may be found (the upper limit for the 80% credibility interval was 0.471).

The meta-analysis performed for external memory attributes (see Table 5) revealed for younger children (see Table 5) a significant, positive, and more than large magnitude ($\delta > 1.2$; an effect size above 61.0% of all positives, PS_{ES} = .610) mean true effect size. However, unexplained variance is due to moderators (% VAR < 75), with a high dispersion in its effects, oscillating the limits of 80% of all studies (credibility interval) between a more than large negative effect size, -1,521, and an extraordinary large positive effect size, 3.945, and with a probability of error of 11.3% (PIS = .113). Similarly, meta-analytic results for older children reported a significant, positive, and close to medium magnitude (δ = 0.424; an effect size above 23.6% of all positives, PS_{ES} = .236) mean true effect size. Nonetheless, results are explained by moderators (% VAR < 75), and with a high dispersion in their effects, oscillating the limits of 80% of all studies (credibility interval) between a medium negative effect size, -0.489, and a more than large positive effect size, 1.337, and a probability of error of 33.6% (PIS = .336). Comparatively, the observed effect for younger children (d = 1.212) was significantly higher, $q_s(N' = 683) = 0.364$, $z = 6.71$, $p < .001$, than for older children (d = 0.424).

As for the internal attributes, the results of the meta-analysis for younger children (see Table 5) displayed a significant, positive, and between medium and large ($0.5 < \delta < 0.8$; an effect size above 37.6% of all positives, PS_{ES} = .376) mean true effect size. However, the results are not generalizable, i.e., negative effects may be found (the lower limit of the credibility interval is negative and of a large magnitude, -0.865), whereas the probability of error rises to 24.2% (PIS = .242), and the unexplained variance is due to moderators (% Var < 75). Similarly, meta-analytic results for older children exhibited a significant, positive, and small magnitude (an effect size above 13.5% of all positives, PS_{ES} = .135) mean true effect size. Once again, results are intervened by moderators (% Var < 75), not generalized (the lower limit of the credibility interval is negative and of a large magnitude, -0.814), and with a probability of error of 40.3% (PIS = .403). Although results for younger and older children are insufficient to establish invariant conclusions ($N < 400$), and with this safeguard, the effect size for internal attributes was significantly higher for younger than for older children, $q_s(N' = 191) = 0.220$, $z = 2.13$, $p < .05$.

Comparatively, external memory attributes discriminate significantly more between memories of perceived events and fabricated memories of events in younger children (d = 1.212) than in older children (d = 0.424), $q_s(N' = 682) = 0.288$, $z = 6.71$, $p < .001$, and adults (d = 0.505), $q_s(N' = 769) = 0.324$, $z = 6.34$, $p < .001$. No differences were observed between adults and older children samples, $q_s(N' = 3{,}612) = 0.040$, $z = 1.70$, *ns*. On the other hand, internal memory attributes discriminate significantly between memories of perceived events and memories of fabricated events, contrary to the prediction of the model in underage (higher scores in perceived memories), while in adults scores significantly more internal attributes were registered in fabricated memories.

Type of Evocation

Two methods of evocation (bringing to memory) of perceived memories were used in research designs, self-experienced events and non-experienced events, watched on video, which has been proposed as a moderator of the effects (Masip et al., 2005).

The results of the meta-analysis run for memories of self-experienced events on external memory attributes (see Table 6) showed a significant, positive, and between small and medium magnitude ($0.2 < \delta < 0.5$; an effect size above 21.3% of all positives, PS_{ES} = .213) mean true effect size. Nevertheless, primary studies are not homogeneous (% VAR < 75), advertising that results are influenced by moderators; positive effects are not generalizable (the lower limit of the credibility interval is negative, -0.335) to all the population of studies, and the probability of error in the classification of perceived memories applying this criterion grows to 35.3% (PIS = .353). No significant effect (the confidence interval for d has zero) was observed for internal memory attributes in memories of self-experienced events.

By other hand, the results of the meta-analysis performed on external attributes for non-experimented events (see Table 7) displayed a significant, positive, and medium magnitude (an effect size above 32.6% of all positives, PS_{ES} = .326) mean true effect size. Nevertheless, primary studies are not homogeneous (% Var < 75), indicating that the effect size is conditioned by moderators; positive effects are not generalizable to all the studies population (the lower limit of the credibility interval is negative); and the probability of error was of 27.7% (PIS = .277). As for the internal attributes, the results revealed a significant, negative (more internal attributes in imagined memories), and small magnitude (an effect size above 12.7% of all negative effects, PS_{ES} = .127) mean true effect size. Nonetheless, the probability of error was 41.6% (PIS = .413); studies are not homogeneous (% Var < 75%), suggesting the presence of moderators of the effect; and positive effect sizes may be found (the upper limit for the 80% credibility interval was 0.501).

Comparatively, external memory attributes discriminate significantly better between perceived and imagined memories, $q_s(N' = 3723) = 0.104$, $z = 4.44$, $p < .001$, in memories of non-experienced events (d = 0.537 vs. d = 0.342 in memories of self-experienced events), while internal attributes do discern significantly between perceived and imagined memories of non-experienced events but not of self-experienced events.

Table 6. Meta-analysis of Reality Monitoring Memory Attributes in Self-Experienced Events

k	N	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
External Memory Attributes										
94	9,749	0.342	0.5423	0.1985	0.5047	0.378	0.5575	13.62	0.302, 0.382	-0.335, 1.091
Internal Memory Attributes										
21	1,942	-0.026	0.5564	0.2088	0.5158	-0.028	0.5704	14.08	-0.115, 0.063	-0.758, 0.702

Table 7. Meta-analysis of the Reality Monitoring Memory Attributes in Non-Experimented Events

k	N	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
External Memory Attributes										
25	2,302	0.537	0.5122	0.2130	0.4658	0.593	0.5133	17.90	0.454, 0.620	-0.064, 1.250
Internal Memory Attributes										
9	830	-0.199	0.5509	0.2097	0.5095	-0.220	0.5632	14.55	-0,336, -0.062	-0.941, 0.501

Table 8. Meta-analysis of the Reality Monitoring Memory External Attributes for the Moderator 'Scoring of Criteria'

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
Categorical: Presence vs. Absence										
2	64	0.809	0.0097	0.3706	0.0000	0.895	0.0000	100[1]	0.291, 1.327	0.895, 0.895
Rating Scales										
65	5,339	0.256	0.6659	0.2225	0.6277	0.283	0.6939	11.23	0.202, 0.310	-0.605, 1.171
Frequency/Density Counts										
69	7,345	0.466	0.4993	0.1972	0.5058	0.516	0.5059	16.07	0.420, 0.512	-0.131, 1.163

Criterion Scoring

Three units of measurement were employed in primary studies to evaluate the effects of content categories based on memory attributes: scoring scales, categorical measure (presence vs. absence), and frequency/density counts (i.e., standardization of the frequency by the duration of the account or by a number of words).

For a categorical scoring (see Table 8), adjustable to lawsuits, a positive, significant, and large magnitude (δ = 0.8) mean true effect size was obtained, but coming only from two effect sizes, not randomly distributed (% Var = 100), and an *N* of 64 that are insufficient to draw any certain conclusion. In the evaluation of the external attributes in rating scales, the meta-analysis exhibited a positive (higher scores in memories of perceived events in comparison with memories of imagined events), significant, and small magnitude (an effect size above 22.8% of all positives, PS_{ES} = .228) mean true effect size. Moreover, the results are no generalized to the population of studies measured the effect in rating scales (credibility intervals range from a negative medium effect size, -0.605, to positive large effect size, 1.171); primary studies are not homogeneous (% VAR < 75), advertising of the influence of moderators in the results, and the probability of error was of 38.8% (PIS = .388). Likewise, in frequency/density counts, the meta-analysis exhibited a positive, significant, and medium magnitude (δ = 0.5; an effect size above 28.1% of all positives, PS_{ES} = .281) mean true effect size. However, results are no generalized to the population of studies, the effect measured in frequency/density counts (credibility interval ranges from a negative effect size, -0.131, to a positive effect size, 1.163); primary studies are not homogeneous (%VAR < 75), i.e., the results are explained by moderators; and the probability of error is estimated in 30.3% (PIS = .303).

Meta-analytic results stated a significantly higher effect size when external attributes are registered in frequency/density counts (d = 0.516) than in rating scales (d = 0.283), $q_s(N' = 6183) = 0.114$, $z = 6.34$, $p < .001$.

For the evaluation on a scale of categorical measurement of internal attributes, only one study with an effect size of 0 was found. As for the rating scale measurement, the results of the meta-analysis (see Table 9) revealed a significant, positive (more internal attributes in perceived memories), generalizable, and between small and medium magnitude (0.20 > δ < 0.5; an effect size above 25.9% of all positives, PS_{ES} = .259) mean true effect size. Nonetheless, studies are not homogeneous (% Var < 75%), suggesting the presence of moderators of the effect; and the error in the classification of perceived memories as imagined memories (empirical model, contrary to the hypothesized model) with this criterion raises to 32.2% (PIS = .322). In the frequency/density count measure, the results of the meta-analysis (see Table 9) displayed a significant, negative (more internal attributes in imagined memories), and small magnitude (an effect size above 11.9% of all negatives, PS_{ES} = .119) mean true effect size. However, studies are not homogeneous (% Var < 75%), i.e., results are conditioned by moderators, negative results are not generalizable (the upper limit for the 80% credibility interval was positive, 0.603), and the error in the classification of imagined memories as perceived applying this criterion rises to 41.4% (PIS = .414).

The contrast of the results as measured in rating scales and frequency/density counts showed that significantly more internal attributes are registered in memories of perceived events when measuring in rating scales, while conversely significantly more internal attributes are registered in memories of imagined events when measuring in frequency/density counts.

Discussion

The results of meta-analyses are subject to limitations that must be borne in mind. First, the fidelity of the inter-context coding is not controlled, that is, between studies, so there is no verification that analysis categories have been coded in the same way in the different studies (Arce et al., 2000). Second, almost exclusively laboratory studies, although generally high-fidelity, were designed which have been shown to give qualitatively different results from field studies in the forensic research setting, so that findings are not directly generalizable to forensic practice (Konecny & Ebbesen, 1979). In this regard, it has been found that coders use different decision strategies (Fariña et al., 1994) in laboratory (more liberal in the coding of categories that associate a higher performance of the model as it does not have judicial implications, i.e., confirmation bias; Sporer et al., 2021) and in the field studies (more conservative, in this case, in the coding of external categories because these are linked to guilty verdicts), and that participants have less involvement and motivation, which is associated with a decrease in memory production, especially in the condition of imagined memories (Alonso-Quecuty & Hernández-Fernaud, 1997; Rogers, 2018). Third, stories have been evaluated, that are insufficient evidence (although many do not report the length, it was found that 62-word stories have been taken as enough) for a categorical content analysis that discriminates between memories of perceived and imagined events. In this way, productivity of content categories decreased (Arce, 2017; Köhnken, 2004). Fourth, the model was unexpectedly applied in a forensic setting to classify false memories (Masip et al., 2005; Vrij, 2008), when this classification has no forensic utility (the test of

Table 9. Meta-analysis of Reality Monitoring Memory Internal Attributes for the Moderator 'Scoring of Criteria'

k	*N*	d_w	SD_d	SD_{pre}	SD_{res}	δ	SD_δ	% Var	95% CI_d	80% CI_δ
Rating scales										
8	580	0.418	0.3111	0.2384	0.1999	0.462	0.21848	59.71	0.253, 0.583	0.182, 0.742
Frequency/Density Counts										
24	2,320	-0.196	0.6148	0.2047	0.5797	-0.217	0.6410	11.14	-0.278, -0.114	-1.037, 0.603

credibility of the testimony is aimed at providing value of evidence to the complainant's testimony, not to classify it as false) and the assumption that the lack of criteria is not correct (lack of evidence is not evidence –only one criterion, cognitive operations, is related to memories of internal origin by what the classification of memory as of internal origin is explained by the lack of criteria of external origin) has proved false (in forensic context other alternatives are possible as lack of cooperation or loss of memory) (Arce, 2017). Fifth, the type of interview to obtain the account, that has direct effects on the contents of the account, has not been exactly defined (Memon et al., 2010). Sixth, the effects of the interviewer on collected protocols (interviews), that may be biasing the results, are not controlled. Seventh, the method of specifying content categories, exploratory factorial analysis, does not guarantee a factorial invariance that a categorical content analysis system is required to be methodical, i.e., reliable and valid (Weick, 1985).

The results of the meta-analyses carried out confirm the usefulness of the total score in any of its Reality Monitoring measures to discriminate between memories of perceived and imagined events. Reversing this effect to a trivial effect (.10) would require 158 missing studies averaging null findings (FDA; Schmidt & Hunter, 2015). In addition, the results are generalizable between studies (results contrary to the model are not expected) and to all kinds of perceived memories (they are not limited to sexual abuse, as has been erroneously concluded occasionally in science and is frequently argued in forensic practice; Arce, 2017). However, there is no harmonization to this extent. In fact, three groupings were found with more than one study and four singular ones in which the total score is the result of different groupings of criteria. Contrary to the theory of the measure (the more criteria, the greater the reliability and, by extension, the validity of the measure; Cronbach, 1951), the model of Vrij et al. (2004a, 2004b) composed of 4 criteria, it discriminates between perceived and imagined memories better than Sporer and Küpper's (2004) of 8 criteria. This can happen for two reasons: that some criteria do not really measure what they are believed to measure and that the criteria of Vrij et al. conform to the core criteria, or that the studies are insufficient to guarantee a random distribution ($k < 3$ for Vrij et al.'s, 2004a, 2004b model). There is also no harmonization on how to score the criterion in the total RM score: some reversed the internal score and added to the total, while others subtracted the raw internal score to the sum of the external score. Anyway, these results are insufficient for the transfer to forensic practice since the margin of error (not admissible in forensic practice since it violates the principle of presumption of innocence, therefore it is not sufficient evidence to give evidence value to the testimony of the victim-complainant) in the classification of perceived memories (the classification of imagined memories is not a forensic task) oscillates, depending on whether one or another estimate of the total RM score is applied, between approximately 20 and 30%. In other words, in forensic evaluation it is not enough to ascertain that in the memory of the complainant-victims (the test of evaluation of the credibility of the testimony is executed as a prosecution test to provide the testimony of the complainant with evidential aptitude-victim) there is a higher score in the total RM score; it is necessary to classify the origin of the memory as external (memories of perceived events), along with the margin of error in such a classification (Daubert vs. Merrell Dow Pharmaceuticals, 1993). Thus, the resulting evidence is not judicial evidence (e.g., Sentencia del Tribunal Constitucional [Spanish Constitutional Court sentence] 16/2012, de 13 de febrero, 2012) valid and sufficient (it does not undermine the principle of presumption of innocence by not knowing 'strict decision criterion' that prevents any memory of fabricated events from being classified as memory of self-experienced events, that is, incriminating an innocent; Art 11.1 of the Universal Declaration of Human Rights; United Nations, 1948).

With regard to the study of the criteria, mixed results were found. Thus, results validate the model (a higher score in perceived memories) in clarity and vividness, sensory information, spatial information, time information, reconstructability of the story, and realism criteria (external attributes). Moreover, for the time information criterion they are generalizable between-studies. However, these results are not generalizable (adverse effects to the prediction of the model may be obtained) for clarity and vividness, sensory information, spatial information, reconstructability of the story, and realism criteria; and unexplained variance is due to moderators. This implies that future research has to identify potential explanatory moderators of adverse results. Furthermore, the probability of error in the classification of memories of perceived events with these criteria ranged from around 29 to 39%. Consequently, they are not strict in the classification of memories of perceived events, so for forensic practice they have to be taken as a whole (i.e., total RM score). Conversely, a non-significant effect was observed for affective information criterion. In short, this criterion does not discriminate between memories of perceived and imagined events. For this reason, it introduces noise into the total RM score, thus partially explaining the lower performance of the model with 8 criteria compared to that of 4. On the other hand, although the cognitive operations criterion (internal attribute) discerns significantly between imagined and perceived memories of events in line with the model prediction (higher scores in imagined memories), the magnitude of the effect is practically nil and, on the contrary, the margin of error (classification of memories imagined as perceived) rises to 45.3%. Furthermore, the direction of the effect is not generalizable, and it is possible to find effects contrary to the prediction of the model that the study of moderators of the future literature should identify. Hence, results do not support the introduction of this criterion in the computation of the total RM score. For forensic practice, this criterion would not be valid either, since it classifies imagined memories, when the forensic task is to classify perceived memories as such, not to classify memories as imagined or to rule out their being imagined (Arce, 2017).

With regard to the sub-criteria of the sensory criterion, the results showed that the visual information and auditory sensations discriminate (significantly more the auditory than visual sensations) significantly between perceived and imagined memories of events. These results are generalizable (no adverse results are expected) and with remarkable effect sizes. Future research has to establish whether segregation increases validity over the joint measure. If validity is increased, these two criteria should be taken as independent categories. On the other hand, smell, taste, and physical sensations subcategories are not productive, so they have to be dispensed with or added to a larger category for the correct preparation of a methodical categorical system, i.e., reliable and valid (Bardin, 1996).

Age has been shown to be a key moderator for forensic practice. Not surprisingly, the forensic application of this type of tool has been mainly limited to children and sexual abuse. In this regard, the results exhibited that internal and external memory attributes distinguish between imagined and perceived memories of events in adults and underage, both older and younger children. However, the direction of the effects varies in the attributes of internal origin from one type of population to another: negative (more internal attributes in imagined memories), confirming the model prediction in adults and positively (more internal attributes in perceived memories), refusing the model prediction in underage, both older and younger children. Nevertheless, these results are in terms of average; contrary results can be found in the three conditions (the results are not generalizable). For this reason, in addition to the fact that classification of imagined memories is not a forensic task, the use of this criterion by age groups, both in isolation (the probability of error ranges between 24 and 40%) and for the computation of the total RM score (adverse results may be found, age not being the moderator that explains them), is not validated by the results. In attributes related to memories of external origin, the predictions of the model are fulfilled in the three populations and, to a greater extent, among younger children. For forensic practice

these results do not validate the technique as it is observed that the results are not generalizable, estimating the probability of errors in 11.3, 33.6 and 30.7% (for younger children, older children and adults, respectively) in the classification of imagined memories as perceived, and it is not specified a strict decision criterion that corrects the error of classification of imagined memories as perceived.

The results of the evocation type moderator have reflected that the criteria that the model relates to memories of external origin significantly differentiate between imagined and perceived memories, both of self-experienced and non-experienced (watched on video) events. These results invalidate the technique as a whole for its forensic use, because in this setting the burden of proof requires the forensic evidence to discriminate memories of lived events from memories of non-lived events. In sum, external memory attributes discriminate between perceived and imagined memories, but not between perceived memories of a self-experienced and non-experienced event (both are perceived memories), the true object of forensic incriminating evidence. With regard to internally sourced memory attributes, the results do not support the model in memories of self-experienced events, while they do support it in memories of perceived but non-experienced events. Again, these results are not generalizable and extensible to forensic setting.

Unfortunately, from the last moderator studied, criterion scoring for the categorical measure (presence vs. absence) has no evidence (for internal attributes) or sufficient evidence (for external attributes; $N = 64$, $k = 2$). This is an adequate measure for forensic practice. From this it is possible to respond to legal demands to forensic evidence (the court requires the forensic evidence of charge to comply with the principle of presumption of innocence, full security, not high probability): a strict decision criterion controlling false positives may be drawn (Sentencia del Tribunal Supremo de 29 de octubre de 1981) and an estimation of the error must be provided (Daubert vs. Merrell Dow Pharmaceuticals, 1993). Surprisingly, the results varied according to the type of measure of RM criteria, indicating an imperfect construct validity (it does not mean invalidity). Thus, contradictory results were obtained in internal criteria: higher scores in imagined memories when measured in frequency/density counts, while higher scores were observed in perceived memories when measured in rating scales. Significantly higher external attributes were registered when measured in frequency/density counts than in rating scales. In sum, the type of measure affects the results, so future research must establish the causes.

Conflict of Interest

The authors of this paper declare no conflict of interest.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Akehurst, L., Easton, S., Fuller, E., Drane, G., Kuzmin, K., & Litchfield, S. (2017). An evaluation of a new tool to aid judgements of credibility in the medico-legal setting. *Legal and Criminological Psychology, 22*(1), 22-46. https://doi.org/10.1111/lcrp.12079

Alonso-Quecuty, M. (1992). Deception detection and reality monitoring: A new answer to an old question? In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (pp. 228-332). Walter de Gruyter.

*Alonso-Quecuty, M., & Hernández-Fernaud, E. (1997). Play it again Sam: Retelling a lie. *Estudios de Psicología, 18*(57), 29-37. https://doi.org/10.1174/021093997320972025

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and criteria based content analysis: A meta-analytic review. *European Journal of Psychology Applied to Legal Context, 7*(1), 3-12. https://doi.org/10.1016/j.ejpal.2014.11.002

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-based content analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology, 16*(2), 201-210. https://doi.org/10.1016/j.ijchp.2016.01.002

Arce, R. (2017). Análisis de contenido de las declaraciones de testigos: evaluación de la validez científica y judicial de la hipótesis y la prueba forense [Content analysis of the witness statements: Evaluation of the scientific and judicial validity of the hypothesis and the forensic proof]. *Acción Psicológica, 14*(2), 171-190. https://doi.org/10.5944/ap.14.2.21347

Arce, R., Arias, E., Novo, M., & Fariña, F. (2020). Are interventions with batterers effective? A meta-analytical review. *Psychosocial Intervention, 29*(3), 153-164. https://doi.org/10.5093/pi2020a11

Arce, R., Fariña, F., & Fraga, A. (2000). Género y formación de juicios en un caso de violación [Gender and juror judgment making in a case of rape]. *Psicothema, 12*(4), 623-628. http://www.psicothema.com/pdf/381.pdf

Arce, R., Fariña, F., Seijo, D., & Novo, M. (2015). Assessing impression management with the MMPI-2 in child custody litigation. *Assessment, 22*(6), 769-777. https://doi.org/10.1177/1073191114558111

Arias, E., Arce, R., Vázquez, M. J., & Marcos, V. (2020). Treatment efficacy on the cognitive competence of convicted intimate partner violence offenders. *Anales de Psicología/Annals of Psychology, 36*(3), 427-435. https://doi.org/10.6018/analesps.428771

Bardin, L. (1996). *El análisis de contenido* [Content analysis] (2nd ed.). Akal.

*Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005). Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. *Applied Cognitive Psychology, 19*(8), 985-1001. https://doi.org/10.1002/acp.1139

*Bembibre, J., & Higueras, L. (2012). Comparative analysis of true or false statements with the source monitoring model and the cognitive interview: Special features of the false accusation of innocent people. *Psychology, Crime & Law, 18*(10), 913-928. https://doi.org/10.1080/1068316X.2011.589387

*Bogaard, G., Colwell, K., & Crans, S. (2019). Using the reality interview improves the accuracy of the criteria-based content analysis and reality monitoring. *Applied Cognitive Psychology, 33*(6), 1018-1031. https://doi.org/10.1002/acp.3537

*Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*(3), 313-329. https://doi.org/10.1002/acp.1087

Cohen, J. (1988). *Statistical power analysis for behavioral sciences* (2nd ed.). LEA.

*Colwell, K., Hiscock-Anisman, C. K., Memon, A., Taylor, L., & Prewett, J. (2007). Assessment criteria indicative of deception (ACID): An integrated system of investigative interviewing and detecting deception. *Journal of Investigative Psychology and Offender Profiling, 4*(3), 167-180. https://doi.org/10.1002/jip.73

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. https://doi.org/10.1007/BF02310555

Daubert vs. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786 (1993).

Davies, G. M. (1994). Children's testimony: Research findings and police implications. *Psychology, Crime, & Law, 1*(2), 175-180. https://doi.org/10.1080/10683169408411951

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118. https://psycnet.apa.org/doi/10.1037/0033-2909.129.1.74

*Elntib, S., & Wagstaff, G. (2017). Are reality monitoring differences between truthful and deceptive autobiographical accounts affected by standardisation for word-count and the presence of others? *Psychology, Crime & Law, 23*(7), 699-716. https://doi.org/10.1080/1068316X.2017.1298762

*Elntib, S., Wagstaff, G. F., & Wheatcroft, J. M. (2015). The role of account length in detecting deception in written and orally produced autobiographical accounts using reality monitoring. *Journal of Investigative Psychology and Offender Profiling, 12*(2), 185-198. https://doi.org/10.1002/jip.1420

Fariña, F., Arce, R., & Novo, M. (2002). Heurístico de anclaje en las decisiones judiciales [Anchorage in judicial decision making]. *Psicothema, 14*(1), 39-46. http://www.psicothema.com/pdf/684.pdf

Fariña, F., Arce, R., & Real, S. (1994). Ruedas de identificación: de la simulación y la realidad [Lineups: A comparision of high fidelity research and research in a real context]. *Psicothema, 6*(3), 395-402. http://www.psicothema.com/pdf/935.pdf

Fariña, F., Redondo, L., Seijo, D., Novo, M., & Arce, R. (2017). A meta-analytic review of the MMPI validity scales and indexes to detect defensiveness in custody evaluations. *International Journal of Clinical and Health Psychology, 17*(2), 128-138. https://doi.org/10.1016/j.ijchp.2017.02.002

Gallego, R., Novo, M., Fariña, F., & Arce, R. (2019). Child-to-parent violence and parent-to-child-violence: A meta-analytic review. *European Journal of Psychology Applied to Legal Context, 11*(2), 51-59. https://doi.org/10.5093/ejpalc2019a4

*Gnisci, A., Caso, L., & Vrij, A. (2010). Have you made up your story? The effect of suspicion and liars' strategies on reality monitoring. *Applied Cognitive Psychology, 24*(6), 762-773. https://doi.org/10.1002/acp.1584

*Granhag, P. A., Strömwall, L. A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. *Legal and Criminological Psychology, 11*(1), 81-98. https://doi.org/10.1348/135532505X49620

*Hernández-Fernaud, E., & Alonso-Quecuty, M. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes? *Applied*

Cognitive Psychology, 11(1), 55-68. https://doi.org/10.1002/(SICI)1099-0720(199702)11:1%3C55::AID-ACP423%3E3.0.CO;2-G

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Sage.

*Izotovas, A., Vrij, A., Hope, L., Mann, S., Granhag, P. A., & Strömwall, L. A. (2018). Facilitating memory-based lie detection in immediate and delayed interviewing: The role of mnemonics. *Applied Cognitive Psychology, 32*(5), 561-574. https://doi.org/10.1002/acp.3435

*Jang, K. W., & Lee, J. H. (2010). The combination of P3-based GKT and reality monitoring in detecting deception. *International Journal of Psychophysiology, 77*(3), 330. https://doi.org/10.1016/J.IJPSYCHO.2010.06.261

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General, 117*(4), 371-376. https://doi.org/10.1037/0096-3445.117.4.371

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*(1), 67-85. https://doi.org/10.1037/0033-295X.88.1.67

*Jupe, L. M., Vrij, A., Leal, S., & Nahari, G. (2018). Are you for real? Exploring language use and unexpected process questions within the detection of identity deception. *Applied Cognitive Psychology, 32*(5), 622-634. https://doi.org/10.1002/acp.3446

*Kleinberg, B., Nahari, G., Arntz, A., & Verschuere, B. (2017). An investigation on the detectability of deceptive intent about flying through verbal deception detection. *Collabra: Psychology, 3*(1), 21. https://doi.org/10.1525/collabra.80

Köhnken, G. (2004). Statement validity analysis and the 'detection of the truth'. In A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 41-63). Cambridge University Press.

Konecny, V. J., & Ebbesen, E. B. (1979). External validity of research in legal psychology. *Law and Human Behavior, 3*(1-2), 39-70. https://psycnet.apa.org/doi/10.1007/BF01039148

*Krackow, E. (2010). Narratives distinguish experienced from imagined childhood events. *American Journal of Psychology, 123*(1), 71-80. https://doi.org/10.5406/amerjpsyc.123.1.0071

*Larsson, A. S., & Granhag, P. A. (2005). Interviewing children with the cognitive interview: Assessing the reliability of statements based on observed and imagined events. *Scandinavian Journal of Psychology, 46*(1), 49-57. https://doi.org/10.1111/j.1467-9450.2005.00434.x

*Logue, M., Book, A. S., Frosina, P., Huizinga, T., & Amos, S. (2015). Using reality monitoring to improve deception detection in the context of the cognitive interview for suspects. *Law and Human Behavior, 39*(4), 360-367. https://doi.org/10.1037/lhb0000127

*Mac Giolla, E., Ask, K., Granhag, P. A., & Karlsson, A. (2019). Can reality monitoring criteria distinguish between true and false intentions? *Journal of Applied Research in Memory and Cognition, 8*(1), 92-97. https://doi.org/10.1016/j.jarmac.2018.08.002

*Manzanero, A. L., Alemany, A., Recio, M., Vallet, R., & Aróztegui, J. (2015). Evaluating the credibility of statements given by persons with intellectual disability. *Anales de Psicología/Annals of Psychology, 31*(1), 338-344. https://doi.org/10.6018/analesps.31.1.166571

*Manzanero, A. L., López, B., & Aróztegui, J. (2016). Underlying processes behind false perspective production. *Anales de Psicología/Annals of Psychology, 32*(1), 256-265. https://doi.org/10.6018/analesps.32.1.194461

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law, 11*(1), 99-122. https://doi.org/10.1080/10683160410001726356

*McDougall, A. J., & Bull, R. (2015). Detecting truth in suspect interviews: The effect of use of evidence (early and gradual) and time delay on criteria-based content analysis, reality monitoring and inconsistency within suspect statements. *Psychology, Crime & Law, 21*(6), 514-530. https://doi.org/10.1080/1068316X.2014.994631

*Memon, A., Fraser, J., Colwell, K., Odinot, G., & Mastroberardino, S. (2010). Distinguishing truthful from invented accounts using reality monitoring criteria. *Legal and Criminological Psychology, 15*(2), 177-194. https://doi.org/10.1348/135532508X401382

Memon, A., Meissner, C. A., & Fraser, J. (2010). Cognitive interview. A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law, 16*(4), 340-372. https://psycnet.apa.org/doi/10.1037/a0020518

Monteiro, A., Vázquez, M. J., Seijo, D., & Arce, R. (2018). ¿Son los criterios de realidad válidos para clasificar y discernir entre memorias de hechos auto-experimentados y de eventos vistos en vídeo? [Are the reality criteria valid to classify and to discriminate between memories of self-experienced events and memories of video-observed events?]. *Revista Iberoamericana de Psicología y Salud, 9*(2), 149-160. https://doi.org/10.23923/j.rips.2018.02.020

*Nahari, G. (2018). Reality monitoring in the forensic context: Digging deeper into the speech of liars. *Journal of Applied Research in Memory and Cognition, 7*(3), 432-440. https://doi.org/10.1016/j.jarmac.2018.04.003

*Nahari, G., Vrij, A., & Fisher, R. P. (2014). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology, 19*(2), 227-239. https://doi.org/10.1111/j.2044-8333.2012.02069.x

Novo, M., & Seijo, D. (2010). Judicial judgement-making and legal criteria of testimonial credibility. *European Journal of Psychology Applied to Legal Context, 2*(2), 91-115. https://journals.copmadrid.org/ejpalc/art/0b7e926154c1274e8b602ff0d7c133d7

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior, 40*(4), 440-457. https://psycnet.apa.org/doi/10.1037/lhb0000193

Redondo, L., Fariña, F., Seijo, D., Novo, M., & Arce, R. (2019). A meta-analytical review of the responses in the MMPI-2/MMPI-2-RF clinical and restructured scales of parents in child custody dispute. *Anales de Psicología/Annals of Psychology, 35*(1), 156-165. https://doi.org/10.6018/analesps.35.1.338381

*Roberts, K. P., & Lamb, M. E. (2010). Reality-monitoring characteristics in confirmed and doubtful allegations of child sexual abuse. *Applied Cognitive Psychology, 24*(8), 1049-1079. https://doi.org/10.1002/acp.1600

Rogers, R. (2018). Researching response styles. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 592-614). Guilford Press.

*Santtila, P., Roppola, H., & Niemi, P. (1998). Assessing the truthfulness of witness statements made by children (aged 7-8, 10-11, and 13-14) employing scales derived from Johnson and Raye's model of Reality Monitoring. *Expert Evidence, 6*(4), 273-289. https://doi.org/10.1023/A:1008930821076

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting errors and bias in research findings* (3rd ed.). Sage.

Schooler, J. W., Gerhard, D., & Loftus, E. F. (1986). Qualities of the unreal. *Journal of Experimental Psychology: Learning, Memory & Cognition, 12*(2), 171-181. https://psycnet.apa.org/doi/10.1037/0278-7393.12.2.171

Sentencia del Tribunal Constitucional 16/2012, de 13 de febrero. (2012). *Boletín Oficial del Estado, 61*, 8-15. https://www.boe.es/boe/dias/2012/03/12/pdfs/BOE-A-2012-3531.pdf

Sentencia del Tribunal Supremo de 29 de octubre. (1981). https://vlex.es/vid/-76757546

*Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11*(5), 373-397. https://doi.org/10.1002/(SICI)1099-0720(199710)11:5%3C373::AID-ACP461%3E3.0.CO;2-0

Sporer, S. L. (2004). Reality monitoring and the detection of deception. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 64-102). Cambridge University Press.

Sporer, S. L., & Hamilton, S. C. (1996, June). *Should I believe it? Reality monitoring of invented and self-experienced events from early and late teenage years.* Poster presented at the NATO Advanced Study Institute in Port de Bourgenay, France.

Sporer, S. L., & Küpper, B. (2004). Fantasie und wirklichkeit – erinnerungsqualitäten von erlebten und erfundenen geschichten [Fantasy and reality – memory qualities of true and invented stories]. *Zeitschrift für Psychologie, 212*(3), 135-151. https://doi.org/10.1026/0044-3409.212.3.135

Sporer, S. L., Manzanero, A. L., & Masip, J. (2021). Optimizing CBCA and RM research: Recommendations for analyzing and reporting data on content cues to deception. *Psychology, Crime & Law, 27*(1), 1-39. https://doi.org/10.1080/1068316X.2020.1757097

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology, 20*(4), 421-446. https://doi.org/10.1002/acp.1190

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13*(1), 1-34. https://psycnet.apa.org/doi/10.1037/1076-8971.13.1.1

*Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. *Applied Cognitive Psychology, 20*(6), 837-854. https://doi.org/10.1002/acp.1234

Steller, M., & Köhnken, G. (1989). Criteria-based content analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217-245). Springer-Verlag.

*Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. *Applied Cognitive Psychology, 18*(6), 653-668. https://doi.org/10.1002/acp.1021

*Stromwall, L. A., & Granhag, P. A. (2005). Children's repeated lies and truths: effects on adults' judgments and reality monitoring scores. *Psychiatry, Psychology and Law, 12*(2), 345-356. https://doi.org/10.1375/pplt.12.2.345

Suengas, A. G., & Johnson, M. K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined complex events. *Journal of Experimental Psychology: General, 117*(4), 377-389. https://doi.org/10.1037/0096-3445.117.4.377

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, J. G. Ghurye, W. Hoeffding, W. G. Madoo, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 448-485). Stanford University Press.

Undeutsch, U. (1967). Beurteilung der glaubhaftigkeit von zeugenaussagenn [Assessing the credibility of witnesses]. In U. Undeutsch (Ed.), *Handbuch der psychologie, Vol. II: Forensische psychologie* (pp. 26-181). Verlag für Psychologie.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 27-56). Norsted.

United Nations. (1948). *Universal declaration of human rights.* https://undocs.org/en/A/RES/217(III)

*Valverde, M. J., Ruiz, J. A., & Llor, B. (2013). Valoración de la credibilidad del testimonio: Aplicación del modelo reality monitoring [Statement validity assessment: The reality monitoring tool]. *Revista Internacional de Psicología, 12*(2), 1-30. https://doi.org/10.33670/18181023.v12i02.68

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). John Wiley and Sons.

*Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004a). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research, 30*(1), 8-41. https://doi.org/10.1111/j.1468-2958.2004.tb00723.x

*Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 36*(2), 113-126. https://doi.org/10.1037/h0087222

*Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*(4), 239-263. https://doi.org/10.1023/A:1006610329284

*Vrij, A., Mann, S., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior, 32*(3), 253-265. https://doi.org/10.1007/s10979-007-9103-y

*Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior, 31*(5), 499-518. https://doi.org/10.1007/s10979-006-9066-4

Vrij, A., Palena, N., Leal, S., & Caso, L. (2021). The relationship between complications, common knowledge details and self-handicapping strategies and veracity: A meta-analysis. *European Journal of Psychology Applied to Legal Context, 13*(2). https://doi.org/10.5093/ejpalc2021a7

Weick, K. E. (1985). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 1, pp. 567-634). LEA.

Wicker, A. W. (1975). An application of a multitrait-multimethod logic to the reliability of observational records. *Personality and Social Psychology Bulletin, 1*(4), 575-579. https://doi.org/10.1177%2F014616727500100405

*Willén, R. M., & Strömwall, L. A. (2011). Offenders' uncoerced false confessions: A new application of statement analysis? *Legal and Criminological Psychology, 17*(2), 346-359. https://doi.org/10.1111/j.2044-8333.2011.02018.x

*Williams, S. M., Talwar, V., Lindsay, R. C. L., Bala, N., & Lee, K. (2014). Is the truth in your words? Distinguishing children's deceptive and truthful statements. *Journal of Criminology,* 547519. http://doi.org/10.1155/2014/54751