Reality Monitoring: A Meta-analytical Review for Forensic Practice

Yurena Gancedo, Francisca Fariña, Dolores Seijo, Manuel Vilariño, and Ramón Arce

*Universidad de Santiago de Compostela, Spain; †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

ABSTRACT

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, $d = 0.542$ ($\delta = 0.562$), 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, $d = 0.361$ ($\delta = 0.399$), sensory information, $d = 0.359$ ($\delta = 0.397$), spatial information, $d = 0.250$ ($\delta = 0.277$), time information, $d = 0.509$ ($\delta = 0.563$), reconstructability of the story, $d = 0.441$ ($\delta = 0.488$), and realism, $d = 0.420$ ($\delta = 0.464$), 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] ($\delta = -0.119$). The moderating effects of age (more cognitive operations on imagined memories in adults, and on perceived memories in underage), evocation type (more 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.

RESUMEN

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

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 metanálítica 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-analisis 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, $d = 0.542$ ($\delta = 0.562$), entre memorias de eventos imaginados y percibidas. En relación con los criterios, los resultados avalaron las predicciones del modelo (más atributos externos en memorias percibidas) en los criterios claridad, $d = 0.361$ ($\delta = 0.399$), información sensorial, $d = 0.359$ ($\delta = 0.397$), información espacial, $d = 0.250$ ($\delta = 0.277$), información temporal, $d = 0.509$ ($\delta = 0.563$), reconstrucción de la historia, $d = 0.441$ ($\delta = 0.488$), y realismo, $d = 0.420$ ($\delta = 0.464$), 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] ($\delta = -0.119$). 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 discriernen 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.
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. Schooeler 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 factorization (N = 100) of the MCQ scale, finding 8 factors (i.e., clarity, sensory information, spatial information, time information, affective information, reconstructability, realizm, 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 larger than large (df/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 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”, “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. Additionally, 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.](image)

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 (κ; 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). Sufficiently, 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 (κ = 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
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 Ns 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 ≠ 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, η²), 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 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 sampling error variance; δ = mean true effect size; SD_d = the standard deviation of δ; % Var = percent of observed variance accounted by artifactual errors; 95% CI_d = 95% confidence interval for d; 80% CI_d = 80% credibility interval for δ.

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

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>d_w</th>
<th>SD_w</th>
<th>SD_pre</th>
<th>SD_d</th>
<th>δ</th>
<th>SD_δ</th>
<th>% Var</th>
<th>95% CI_δ</th>
<th>80% CI_δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1,696</td>
<td>0.546</td>
<td>0.3439</td>
<td>0.1985</td>
<td>0.2807</td>
<td>0.562</td>
<td>0.2884</td>
<td>33.44</td>
<td>0.449, 0.643</td>
<td>0.193, 0.931</td>
</tr>
<tr>
<td>7</td>
<td>563</td>
<td>0.462</td>
<td>0.3132</td>
<td>0.2269</td>
<td>0.21604</td>
<td>0.475</td>
<td>0.2218</td>
<td>52.55</td>
<td>0.294, 0.630</td>
<td>0.191, 0.759</td>
</tr>
<tr>
<td>3</td>
<td>224</td>
<td>0.617</td>
<td>0.0661</td>
<td>0.2379</td>
<td>0.0000</td>
<td>0.634</td>
<td>0.0000</td>
<td>100%</td>
<td>0.348, 0.886</td>
<td>0.634, 0.634</td>
</tr>
<tr>
<td>2</td>
<td>376</td>
<td>0.801</td>
<td>0.2536</td>
<td>0.1518</td>
<td>0.2031</td>
<td>0.813</td>
<td>0.2061</td>
<td>35.97</td>
<td>0.590, 1.012</td>
<td>0.594, 1.077</td>
</tr>
</tbody>
</table>

Note. k = number of effect sizes; N = total sample size; d_w = sample size weighted mean effect size; SD_w = standard deviation of d; SD_pre = standard deviation predicted for sampling error alone; SD_d = 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_δ = 95% confidence interval for d; 80% CI_δ = 80% credibility interval for δ.

1The predicted variance overcomes the observed variance, rounding it to 100%.

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, 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_E of .556, .637, and .716 categories, adding a supplementary one, a more than large effect size (d/δ > 1.20, corresponding to a PS_E > .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$_E$; 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).
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 34.7% of all positives, PS ES = .347), and large (δ > 0.8; above 6.4% of all negatives, PSES = .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 imagined memories than in perceived memories, but also of a close to large magnitude that refutes it (more cognitive information in perceived memories).

**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 < 0.50; an effect size above 52.6% of all positives, PS ES = .526) for the Sporer & Küpper’s (2004) model, medium for the original model (δ > 0.5; above 347.4% 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% (PS ES = .317), 26.3% (PS ES = .263), and 20.8% (PS ES = .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(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 43.1% of all negatives, PSES = .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% (PS ES = .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(N = 1,478) = 0.80, z = 2.14, p < .05, and spatial information, q(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(N = 769) = 0.013, z = 2.01, p < .05, and that the effect size for realism was significantly higher than for spatial information, q(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 2. Meta-analysis of the Reality Monitoring Categories**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>k</th>
<th>N</th>
<th>d_m</th>
<th>SD_m</th>
<th>SD_sd</th>
<th>δ</th>
<th>SD</th>
<th>% Var</th>
<th>95% CI</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity (and Vividness)</td>
<td>7</td>
<td>700</td>
<td>0.361</td>
<td>0.8684</td>
<td>0.2024</td>
<td>0.8445</td>
<td>0.399</td>
<td>0.9335</td>
<td>5.53</td>
<td>0.058, 0.364</td>
</tr>
<tr>
<td>Sensory Information</td>
<td>18</td>
<td>1,092</td>
<td>0.359</td>
<td>0.6564</td>
<td>0.2602</td>
<td>0.6026</td>
<td>0.397</td>
<td>0.6658</td>
<td>15.88</td>
<td>0.061, 0.239</td>
</tr>
<tr>
<td>Spatial Information</td>
<td>25</td>
<td>2,290</td>
<td>0.250</td>
<td>0.6507</td>
<td>0.2106</td>
<td>0.6157</td>
<td>0.277</td>
<td>0.6806</td>
<td>10.55</td>
<td>0.168, 0.332</td>
</tr>
<tr>
<td>Time Information</td>
<td>23</td>
<td>2,290</td>
<td>0.509</td>
<td>0.4326</td>
<td>0.2044</td>
<td>0.3813</td>
<td>0.563</td>
<td>0.4196</td>
<td>23.08</td>
<td>0.426, 0.592</td>
</tr>
<tr>
<td>Affective Information</td>
<td>15</td>
<td>1,397</td>
<td>0.024</td>
<td>0.6112</td>
<td>0.2081</td>
<td>0.5746</td>
<td>0.027</td>
<td>0.6355</td>
<td>11.59</td>
<td>-0.081, 0.129</td>
</tr>
<tr>
<td>Reconstructability of the Story</td>
<td>5</td>
<td>463</td>
<td>0.441</td>
<td>0.4333</td>
<td>0.2110</td>
<td>0.3784</td>
<td>0.488</td>
<td>0.4170</td>
<td>24.28</td>
<td>0.256, 0.626</td>
</tr>
<tr>
<td>Realism</td>
<td>9</td>
<td>908</td>
<td>0.420</td>
<td>0.6841</td>
<td>0.2020</td>
<td>0.6536</td>
<td>0.464</td>
<td>0.7221</td>
<td>8.92</td>
<td>0.288, 0.552</td>
</tr>
<tr>
<td>Cognitive Information (cognitive operations, idiosyncratic information)</td>
<td>33</td>
<td>3,052</td>
<td>-0.107</td>
<td>0.6199</td>
<td>0.2089</td>
<td>0.5751</td>
<td>-0.119</td>
<td>0.6360</td>
<td>11.68</td>
<td>-0.178, -0.036</td>
</tr>
</tbody>
</table>
Table 3. Meta-analysis of the Sensory Information Subcategories

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>d_w</th>
<th>SD_w</th>
<th>SD_mm</th>
<th>SD_res</th>
<th>δ</th>
<th>SD</th>
<th>% Var</th>
<th>95% CI_d</th>
<th>80% CI_d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>12</td>
<td>1,321</td>
<td>0.452</td>
<td>0.3942</td>
<td>0.1937</td>
<td>0.3433</td>
<td>0.500</td>
<td>0.3779</td>
<td>24.86</td>
<td>0.343, 0.561</td>
</tr>
<tr>
<td>Auditory</td>
<td>11</td>
<td>1,197</td>
<td>0.662</td>
<td>0.5057</td>
<td>0.1976</td>
<td>0.4656</td>
<td>0.732</td>
<td>0.5121</td>
<td>16.20</td>
<td>0.546, 0.778</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>d_w</th>
<th>SD_w</th>
<th>SD_mm</th>
<th>SD_res</th>
<th>δ</th>
<th>SD</th>
<th>% Var</th>
<th>95% CI_d</th>
<th>80% CI_d</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Memory Attributes</td>
<td>90</td>
<td>8,990</td>
<td>0.457</td>
<td>0.4779</td>
<td>0.2034</td>
<td>0.4324</td>
<td>0.505</td>
<td>0.4768</td>
<td>18.62</td>
<td>0.415, 0.499</td>
</tr>
<tr>
<td>Internal Memory Attributes</td>
<td>23</td>
<td>2,041</td>
<td>-0.275</td>
<td>0.5882</td>
<td>0.2142</td>
<td>0.5478</td>
<td>-0.304</td>
<td>0.6055</td>
<td>13.38</td>
<td>-0.362, -0.188</td>
</tr>
</tbody>
</table>

The results 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), respectively. 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 = .489).

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 [-0.404, 0.484], 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 not 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., 2004).

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 size.

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

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>d_w</th>
<th>SD_w</th>
<th>SD_mm</th>
<th>SD_res</th>
<th>δ</th>
<th>SD</th>
<th>% Var</th>
<th>95% CI_d</th>
<th>80% CI_d</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Memory Attributes</td>
<td>15</td>
<td>403</td>
<td>1.096</td>
<td>1.9769</td>
<td>0.4184</td>
<td>1.9321</td>
<td>1.212</td>
<td>2.1350</td>
<td>4.65</td>
<td>0.886, 1.306</td>
</tr>
<tr>
<td>40</td>
<td>2,261</td>
<td>0.384</td>
<td>0.7002</td>
<td>0.2703</td>
<td>0.6459</td>
<td>0.424</td>
<td>0.7137</td>
<td>15.07</td>
<td>0.301, 0.467</td>
<td>-0.489, 1.337</td>
</tr>
<tr>
<td>Internal Memory Attributes</td>
<td>4</td>
<td>129</td>
<td>0.632</td>
<td>1.1643</td>
<td>0.3643</td>
<td>1.1059</td>
<td>0.699</td>
<td>1.2220</td>
<td>9.95</td>
<td>0.276, 0.988</td>
</tr>
<tr>
<td>7</td>
<td>374</td>
<td>0.222</td>
<td>0.7981</td>
<td>0.2764</td>
<td>0.7487</td>
<td>0.246</td>
<td>0.8279</td>
<td>12.04</td>
<td>0.018, 0.426</td>
<td>-0.814, 1.306</td>
</tr>
</tbody>
</table>

Note. Younger children samples (3-8 years); older children samples (9-16 years).
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 (δ > 1.2; an effect size above 61.0% of all positives, PSES = .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, PSES = .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 (δ = 1.212) was significantly higher, q(N = 683) = 0.364, z = 6.71, p < .001, than for older children (δ = 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 < δ < 0.8; an effect size above 37.6% of all positives, PSES = .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, PSES = .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(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 (δ = 1.212) than in older children (δ = 0.424), q(N = 682) = 0.288, z = 6.71, p < .001, and adults (δ = 0.505), q(N = 769) = 0.324, z = 6.34, p < .001. No differences were observed between adults and older children samples, q(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 < δ < 0.5; an effect size above 21.3% of all positives, PSES = .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 δ 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, PSES = .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, PSES = .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(N = 3723) = 0.104, z = 4.44, p < .001, in memories of non-experienced events (δ = 0.537 vs. δ = 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

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>δ</th>
<th>SD</th>
<th>SD_{se}</th>
<th>SD_{ep}</th>
<th>b</th>
<th>SD</th>
<th>% Var</th>
<th>95% CI</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>9,749</td>
<td>0.342</td>
<td>0.5423</td>
<td>0.1985</td>
<td>0.5047</td>
<td>0.378</td>
<td>0.5575</td>
<td>13.62</td>
<td>0.302, 0.382</td>
<td>-0.335, 1.091</td>
</tr>
<tr>
<td>21</td>
<td>1,942</td>
<td>-0.026</td>
<td>0.5564</td>
<td>0.2088</td>
<td>0.5158</td>
<td>-0.028</td>
<td>0.5704</td>
<td>14.08</td>
<td>-0.115, 0.063</td>
<td>-0.758, 0.702</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>δ</th>
<th>SD</th>
<th>SD_{se}</th>
<th>SD_{ep}</th>
<th>b</th>
<th>SD</th>
<th>% Var</th>
<th>95% CI</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>2,302</td>
<td>0.537</td>
<td>0.5122</td>
<td>0.2130</td>
<td>0.4658</td>
<td>0.593</td>
<td>0.5133</td>
<td>17.90</td>
<td>0.454, 0.620</td>
<td>-0.064, 1.250</td>
</tr>
<tr>
<td>9</td>
<td>830</td>
<td>-0.199</td>
<td>0.5509</td>
<td>0.2097</td>
<td>0.5095</td>
<td>-0.220</td>
<td>0.5632</td>
<td>14.55</td>
<td>-0.336, -0.062</td>
<td>-0.941, 0.501</td>
</tr>
</tbody>
</table>
Table 8. Meta-analysis of the Reality Monitoring Memory External Attributes for the Moderator ‘Scoring of Criteria’

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>$d_m$</th>
<th>$SD_{m}$</th>
<th>$SD_{res}$</th>
<th>$\delta$</th>
<th>% Var</th>
<th>95% CI</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical: Presence vs. Absence</td>
<td>2</td>
<td>64</td>
<td>0.809</td>
<td>0.0097</td>
<td>0.3706</td>
<td>0.0000</td>
<td>0.895</td>
<td>0.0000</td>
</tr>
<tr>
<td>Rating Scales</td>
<td>65</td>
<td>5,339</td>
<td>0.256</td>
<td>0.6659</td>
<td>0.2225</td>
<td>0.6277</td>
<td>0.283</td>
<td>0.6939</td>
</tr>
<tr>
<td>Frequency/Density Counts</td>
<td>69</td>
<td>7,345</td>
<td>0.466</td>
<td>0.4993</td>
<td>0.1972</td>
<td>0.5058</td>
<td>0.516</td>
<td>0.5059</td>
</tr>
</tbody>
</table>

### 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 ($\delta = 0.8$) 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, $PS_{es} = \ldots$) 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% ($PIS = \ldots$). Likewise, in frequency/density counts, the meta-analysis exhibited a positive, significant, and medium magnitude ($\delta = 0.5$; an effect size above 28.1% of all positives, $PS_{es} = \ldots$) 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% ($PIS = \ldots$).

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 ($d = 0.283$), $q(N' = 6183) = 0.114$, $z = 6.34$, $p < 0.001$.

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), generalized, and between small and medium magnitude ($0.20 > \delta > 0.5$; an effect size above 25.9% of all positives, $PS_{es} = \ldots$) 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% ($PIS = \ldots$). 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, $PS_{es} = \ldots$) 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% ($PIS = \ldots$).

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 (Farina 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’

<table>
<thead>
<tr>
<th>k</th>
<th>N</th>
<th>$d_m$</th>
<th>$SD_{m}$</th>
<th>$SD_{res}$</th>
<th>$\delta$</th>
<th>% Var</th>
<th>95% CI</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating scales</td>
<td>8</td>
<td>580</td>
<td>0.418</td>
<td>0.3111</td>
<td>0.2384</td>
<td>0.1999</td>
<td>0.462</td>
<td>0.21848</td>
</tr>
<tr>
<td>Frequency/Density Counts</td>
<td>24</td>
<td>2,320</td>
<td>-0.196</td>
<td>0.6148</td>
<td>0.2047</td>
<td>0.5797</td>
<td>-0.217</td>
<td>0.6410</td>
</tr>
</tbody>
</table>
The studies are insufficient to guarantee a random distribution (K < 3) of criteria do not really measure what they are believed to measure. This can happen for two reasons: that is the result of different groupings of criteria. Contrary to the theory of the measure (the more criteria, the greater the reliability and, by harmonization to this extent. In fact, three groupings were found with been erroneously concluded occasionally in science and is frequently argued in forensic practice; Arce, 2017). However, there is no harmonicization 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 Köpper’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 harmonicization 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) discriminates 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 segmentation 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 (Rardin, 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, Inc., 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.


