The European Journal of Psychology Applied to Legal Context

https://journals.copmadrid.org/ejpalc

# The Relationship between Complications, Common Knowledge Details and Self-handicapping Strategies and Veracity: A Meta-analysis

Aldert Vrij[a], Nicola Palena[b], Sharon Leal[a], and Letizia Caso[c]

[a]University of Portsmouth, United Kingdom; [b]University of Bergamo, Italy; [c]University of Lumsa, Rome, Italy

ARTICLE INFO

*Article history:*
Received 25 January 2021
Accepted 12 April 2021
Available online 17 May 2021

*Keywords:*
Meta-analysis
Complications
Common knowledge details
Self-handicapping strategies
Bayes factors

## ABSTRACT

Practitioners frequently inform us that variable ‘total details’ is not suitable for lie detection purposes in real life interviews. Practitioners cannot count the number of details in real time and the threshold of details required to classify someone as a truth teller or a lie teller is unknown. The authors started to address these issues by examining three new verbal veracity cues: complications, common knowledge details, and self-handicapping strategies. We present a meta-analysis regarding these three variables and compared the results with ‘total details’. Truth tellers reported more details ($d$ = 0.28 to $d$ = 0.45) and more complications ($d$ = 0.51 to $d$ = 0.62) and fewer common knowledge details ($d$ = -0.40 to $d$ = -0.46) and self-handicapping strategies ($d$ = -0.37 to $d$ = -0.50) than lie tellers. Complications was the best diagnostic veracity cue. The findings were similar for the initial free recall and the second recall in which only new information was examined. Four moderators (scenario, motivation, modality, and interview technique) did not affect the results. As a conclusion, complications in particular appear to be a good veracity indicator but more research is required. We included suggestions for such research.

## La relación de las complicaciones, los detalles de observación común y las estrategias de autojustificación con la veracidad: un meta-análisis

*Palabras clave:*
Meta-análisis
Complicaciones
Detalles de dominio público
Estrategias de falta de capacidad
Factores de Bayes

## RESUMEN

Los profesionales dicen con frecuencia que la variable “detalles totales” es adecuada para la detección de mentiras en las entrevistas de la vida real. No pueden contar el número de detalles en tiempo real y se desconoce el umbral de detalles necesario para clasificar a alguien como sincero o mentiroso. Los autores comenzaron a abordar estos temas analizando tres nuevos indicadores verbales de veracidad: las complicaciones, los detalles de conocimiento público y las estrategias de falta de capacidad. Se presenta un meta-análisis de estas tres variables y se comparan los resultados con los “detalles totales”. Los sujetos que dicen la verdad dan más detalles ($d$ = 0.28 hasta $d$ = 0.45) y complicaciones ($d$ = 0.51 hasta $d$ = 0.62) y menos detalles de conocimiento público ($d$ = -0.40 hasta $d$ = -0.46) y estrategias de justificación ($d$ = -0.37 hasta $d$ = -0.50) que los que mienten. Las complicaciones resultaron ser el mejor indicador diagnóstico de veracidad. Los resultados fueron iguales en la primera entrevista en recuerdo libre y en la segunda entrevista, en la que solo se analizaba información nueva. No influyeron en los resultados cuatro moderadores: el escenario, la motivación, la modalidad y la técnica de entrevista. Como conclusión, las complicaciones parecen ser un buen indicador de veracidad aunque es necesaria más investigación. Se discuten las futuras líneas de investigación.

From verbal cues to deception that have been frequently examined, total details emerged as the best diagnostic cue (Amado et al., 2016). Truth tellers typically report more details than lie tellers ($d$ = 0.55), representing a medium effect size. However, practitioners frequently tell us that variable total details is not useful to them, mainly for two reasons: 1) it is not possible to count the number of details in an interview in real time and 2) it can never be established in an individual interview how many details are required to judge somebody as a truth teller or lie teller. Vrij et al. (2017) started to address these two problems by introducing three new verbal veracity cues: complications, common knowledge details, and self-handicapping strategies. To date they have published 18 samples measuring complications, 12 samples measuring common knowledge details, and 13 samples examining self-handicapping strategies. In this article we present a meta-analysis of that research. This meta-analysis has three aims: to determine the diagnostic value of these three cues to differentiate truth tellers from lie tellers; to assess the gaps in knowledge

Cite this article as: 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. *The European Journal of Psychology Applied to Legal Context.* Ahead of print. https://doi.org/10.5093/ejpalc2021a7

Funding: The time the first author spent working on this article was funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1).
Correspondence: aldert.vrij@port.ac.ukT (A. Vrij).

concerning these three cues; and to encourage other researchers to examine these and other cues.

## Complications, Common Knowledge Details, and Self-Handicapping Strategies

Complications are occurrences that affecs the story-teller and make a situation more complex ("Initially we did not see our friend, as he was waiting at a different entrance") (Vrij, Deeb et al., 2021). Complications are thought to occur more often in truthful statements than in deceptive statements. Making up complications requires imagination but lie tellers may not have adequate imagination to fabricate these (Köhnken, 2004; Vrij, 2008). In addition, lie tellers prefer to keep their stories simple (Hartwig et al., 2007), but adding complications makes the story more complex. Complications as introduced by Vrij and colleagues differs from the unexpected complications criterion that is part of the Criteria-Based Content Analysis (CBCA) tool (Amado et al., 2016). The main difference is that in CBCA complications are necessarily unexpected occurrences, whereas this is not the case in Vrij and colleagues' definition. Thus, when someone says that when driving from their hometown A to their holiday destination C they briefly visited town B, this visit to town B is considered a complication by Vrij and colleagues but not according to the CBCA definition. In other words, Vrij and colleagues' definition solely focuses on 'complexity' without taking any unexpected elements into account. A second, less important, difference is how the coding takes place. Vrij and colleagues always apply frequency coding whereas CBCA-coders often use scales (3-point or 5-point scales) (Vrij, 2008). Frequency coding is more detailed and probably more reliable than scale coding.

Common knowledge details refer to strongly invoked stereotypical information about events ("The event had an Oscars theme so everybody was dressed up"). Lie tellers are thought to report more common knowledge details than truth tellers. Truth tellers have personal experiences of an event and are likely to report these (DePaulo et al., 1996). If lie tellers do not have personal experiences of the event they report or do not have experiences of events related to the to-be-discussed event, they will draw upon general knowledge to construe the event (Sporer, 2016). Common knowledge details have never been examined before in deception research. Deception researchers have discussed scripts, without actually examining them (Köhnken, 2004; Sporer, 2016; Volbert & Steller, 2014). Scripts are different from common knowledge details. A script is a stereotyped sequence of actions that defines a well-known situation (Schank & Abelson, 1977, p. 41) (e.g., 'John went to a restaurant. He ordered lobster. He paid the check and left'). Common knowledge details do not necessarily involve a sequence of events. Thus, the sentence 'We spent a couple of hours at the National Museum' classifies as a common knowledge detail but not as a script.

Self-handicapping strategies refer to justifications as to why someone chooses not to provide information ("There isn't much to say about the actual bungee jump as it took only a few moments"). A real life example occurred with Dominic Cummings, a former advisor to the British Prime Minister Boris Johnson. Cummings travelled from London to Durham during lockdown. This is what he said when he was asked whether he had discussed this trip with Johnson: "At some point during the first week, when we were both sick and in bed, I mentioned to him what I had done. Unsurprisingly, given the condition we were in, neither of us remember the conversation in any detail." (https://inews.co.uk/news/dominic-cummings-lockdown-statement-pm-adviser-said-meant-barnard-castle-431111). For lie tellers, not having to provide information is an attractive strategy. However, they are also concerned about their credibility and believe that admitting lack of knowledge and/or memory appears suspicious (Ruby & Brigham, 1998). A potential solution is to provide a justification for the inability to provide information. Self-handicapping strategies have never been examined in deception research before.

## Real Time Coding

A detail is a unit of information and each new piece of information counts as a detail. This means that many details can occur in a statement, far too many to count in real time. Complications and common knowledge details are clusters of details. Therefore, fewer of them occur in a statement which makes them easier to count in real time. To return to the example mentioned earlier, the sentence "Initially we did not see our friend, as he was waiting at a different entrance" contains seven details but only one complication and the sentence "The event had an Oscars theme so everybody was dressed up" contains four details but only one common knowledge detail. Someone who just listens to those two sentences will have difficulty in counting the details but should be able to spot the complication and common knowledge detail. Self-handicapping strategies are also relatively easy to spot in real time as it typically contains a statement why some information cannot be provided followed by a justification for it.

Note that the coding of complications, common knowledge details, and self-handicapping strategies occurs in addition to frequency of details coding and not instead of such coding. Frequency of details coding can be further specified, for example in contextual details. Again, this would occur in addition to the coding of complications, common knowledge details, and self-handicapping strategies. For example, the sentence "We spent a couple of hours at the National Museum" includes two contextual details (a time detail – couple of hours – and a location detail – National Museum) but the entire sentence constitutes one common knowledge detail. Contextual details are typically reported more often by truth tellers than by lie tellers (Amado et al., 2016), but the way they are reported in this sentence makes it a common knowledge detail and such details are more likely to be reported by lie tellers than by truth tellers.

## Decision-Making in Individual Cases

Most verbal deception research examines verbal cues to truthfulness, that is, verbal cues that truth tellers report more frequently than lie tellers. For example, all 19 CBCA criteria, including the variable 'total details', are cues to truthfulness. Only examining cues to truthfulness poses a problem for practitioners: how many details should someone report to classify as a truth teller or a lie teller? This question is impossible to answer. The number of details reported is not only dependent on veracity but also on the interviewee and situation. That is, some individuals are more talkative than other individuals and some events can be described in much more detail than other events (Nahari & Pazuelo, 2015; Nahari & Vrij, 2014; Vrij, 2016). A practitioner would be in a much stronger position to determine that an interviewee is lying if s/he records not only cues that truth tellers are more likely to report (cues to truthfulness) but also cues that lie tellers are more likely to report (cues to deceit). Examining complications, common knowledge details, and self-handicapping strategies does exactly that: it measures a mixture of cues to truthfulness and cues to deceit. Someone could be more confident that somebody is telling the truth if complications are present and common knowledge details and self-handicapping strategies are largely absent (rather than just complications present) and, vice versa, someone could be more confident that someone is lying when complications are largely absent and common knowledge details and self-handicapping strategies occur.

## Hypotheses

Truth tellers will report more complications (Hypothesis 1) and fewer common knowledge details (Hypothesis 2) and self-handicapping strategies (Hypothesis 3) than lie tellers.

## Method

The procedure applied for conducting this meta-analysis followed the APA Meta-Analysis Reporting Standard (Cooper, 2011) and a recently published meta-analysis on the verifiability approach (Palena et al., 2020).

### Inclusion Criteria

We searched the literature for empirical studies examining complications, common knowledge details and/or self-handicapping strategies and used the following inclusion criteria: (1) the study involved one interviewee rather than pairs or groups (see for studying groups, Leal et al., 2018; Vernham et al., 2020); (2) the study measured the total frequency of complications, common knowledge details, and/or self-handicapping strategies; (3) the study used the coding procedure first outlined by Vrij et al. (2017); (4) veracity was manipulated (either within-subjects or between-subjects); (5) the study could either be based on a single interview or follow the within-subjects approach as outlined in Vrij, Leal, and Fisher (2018), where the interviewee is first asked a free recall, then s/he is exposed to any manipulation (e.g., the model statement), and then asked a second free recall; (6) statements were coded manually; and (7) articles were written in English. If in the selected articles 'total details' was reported, we included this variable in the analyses.

### Moderator Analyses

The first moderator was the deception scenario. This is an important moderator for applied reasons as it examines whether effects can be generalised across scenarios. We distinguished between two categories: i) trip/memorable event, when the participants discussed a trip they had made or a memorable event which was out of the ordinary, and ii) spy mission, where the participants performed a spy mission (Moderator 1). A second moderator was the level of motivation, because research has found that cues to deception are more evident in motivated rather than unmotivated senders (DePaulo et al., 2003). Building on DePaulo et al. (2003) and Suchotzki et al. (2017), we introduced two categories: (i) participants who received an incentive for their participation, such as money, university credits, a present etc., or (ii) participants who did not receive an incentive (Moderator 2).

Following Palena et al. (2020), we also coded the modality of the interview on two levels: i) whether the participants were interviewed, or ii) provided a written statement (Moderator 3). Last, since several interviewing techniques have been introduced in lie detection research (e.g., model statement, Vrij, Leal, & Fisher, 2018, sketching while narrating, Vrij, Leal, Fisher, et al., 2018), we also coded whether or not any manipulation was used in the experiment (Moderator 4). Analyses showed no variability for the moderators incentive (Moderator 2) and modality (Moderator 3) since all studies included an incentive and no study involved a written statement (see Appendix C). Hence, the effect of these moderators was not explored.

### Search Strategies and Studies Selection

We conducted the literature search in October 2020, and used the following databases: PsycINFO, PsycARTICLES, Web of Science, and Scopus. We looked for any type of work (article, review, book chapters, etc.) that included terms ("complication*" OR "common knowledge" OR "self-handicapping", in title/abstract/keywords) and ("decept*" OR "deceit" OR "lie" OR "lying" OR "truth*", in title/abstract/keywords) but did not include the terms ("collect*" OR "pair*", in title/abstract/keywords). We limited our research to the psychological, social sciences, and art and humanities areas, and to sources produced from 2017 onwards. We also: i) contacted scholars who previously published articles in which complications, common knowledge details, or self-handicapping strategies were reported; ii) visited the webpages of scholars working on these types of detail; iii) searched the reference list of the selected papers; iv) searched the 2019-2020 conference programs of the European Association of Psychology and Law (EAPL), and the 2017-2020 conference programs of the American Psychology and Law Society (AP-LS), and the International Investigative Interviewing Research Group (iIIRG); v) and conducted a search on ResearchGate.

The selection process was conducted by two researchers with experience in the field. Inter-rater agreement was calculated via Cohen's k and was 1.00 (100% of agreement). Appendix A shows the illustration of the selection process via the Prisma diagram (Moher et al., 2010) and Appendix B reports a description of the included studies and reasons for the exclusion of studies.

### Coding

A coding protocol was applied to extract all the relevant information needed for the meta-analysis from the selected papers.

The following variables were coded: (a) truth tellers' and lie tellers sample sizes; (b) mean and standard deviation of complications, common knowledge details, self-handicapping strategies, and total details; (c) the deception scenario; (d) whether or not an incentive was provided; (e) whether the participants provided an oral or a written statement; and (f) whether or not any interviewing technique (e.g., model statement, sketching while narrating, etc.) was used. Regarding (b), the descriptives required to compute the effect sizes were obtained from three different outcome variables: i) "initial recall" – this refers to the first statement provided by interviewees who were asked to provide more than one statement, or the only statement provided by interviewees who provided only one recall; ii) "second recall" – this refers to "new" information (complications, common knowledge details etc) not already mentioned in the initial recall provided by the interviewee in a second recall; and iii) "total recall" – this refers to the sum of initial and second recall.

Regarding Moderator 4 (presence of manipulation), for "initial recall" and "second recall" data, we coded each study as "manipulation present" if the interviewees were exposed to any manipulation at any of these recall stages and as "manipulation absent" if they were not. For "total recall" data, we coded manipulation as "present" if participants were exposed to any manipulation at their "initial recall", at "second recall", or at both (see Appendix B for more information concerning the coding of Moderator 4). Inter-rater agreement for non-categorical moderators was analysed via percentage of agreement and was 100%. Inter-rater agreement for categorical moderator was calculated via Cohen's k and was 1.00.

### Data Analysis

Effect sizes were computed as *d* with Cohen's formula when the study sample sizes were equal, and Hedges' formula when the sample sizes were unequal. We obtained Cohen's *d* from the selected studies wherever possible. In all other cases, we contacted the authors to request any missing data (data requested to authors are reported in Appendix B). Cohen's *d* was computed so that positive values indicate a higher frequency amongst truth tellers than amongst lie tellers. For each effect size, 95% CIs, standard error, variance, and the level of significance were computed. However, Cohen's *d* is not always straightforward to interpret. For this reason, we also reported two additional statistics for the estimation of the magnitude of the effect: the probability of superiority of the effect size ($PS_{ES}$) and the probability of inferiority score (PIS) (Arias et al., 2020; Monteiro et al., 2018). The former is a transformation of the observed effect size as a percentile. For example, if a positive effect size represents a higher frequency of details amongst truth tellers than lie tellers, a $PS_{ES}$ = .30 indicates that the observed effect size is greater than 30% of all possible effect sizes. PIS is the probability that truth tellers obtain a score that is lower than the mean score of lie tellers. For example, if PIS = .30, 30% of truth tellers would report an amount of detail lower than the mean score of lie tellers.

The effect sizes obtained from the selected sources were analysed via standard meta-analytic procedures (Borenstein et al., 2011) via the meta-analytic software ProMeta3. All effect sizes were pooled via the inverse-variance method. All outcomes concerning the initial recall, and total detail and complications for second and total recall were pooled via a random-effects model as it allows to account both for within-studies and between-studies variances. In contrast, common knowledge details and self-handicapping strategies for second recall and total recall were pooled using the fixed-effect method.

Heterogeneity was explored using the Q statistic (indicating a lack of homogeneity if significant), and the $I^2$ statistic, which estimates what proportion of the observed variance is related to real differences in the analysed effect sizes. An $I^2$ of 70% or more is deemed as a high difference, 50% as moderate, and 25% as low (Borenstein et al., 2011; Cooper, 2015). Moreover, we explored the standardised residuals of each study to detect possible outliers and sensitivity analyses (where one study at a time is removed) were conducted to explore the (in) stability of the results.

Publication bias was explored via the trim and fill method.

We also carried out meta-analyses via the Bayesian model averaging method (Gronau et al., 2017, 2020). See Appendix J.

## Results

### Included Studied Description

One-hundred and twenty-one records were found through the database search, and three records through other sources. After removing the duplicates, 105 records were screened in their titles and abstracts. Eighty-three records did not relate to verbal credibility assessment and were thus excluded. One was excluded because, although focusing on verbal credibility assessment, it did not analyse complications, common knowledge details, or self-handicapping strategies (Leal et al., 2018), and one was excluded because it focused on pairs of interviewees (Vernham et al., 2020). The remaining 20 records were read in full to evaluate if they matched the eligibility criteria. Three of them were excluded because they did not contain new data (Vrij & Leal, 2020; Vrij, Leal, Mann, et al., 2019; Vrij & Vrij, 2020); one was excluded because it was theoretical rather than empirical (Vrij, Leal, & Fisher, 2018); one was excluded because the only cue that they examined – complications – did not occur frequently enough to be analysed leaving us with no data to report (Verigin, Meijer, Vrij et al., 2020); and one was excluded because we did not manage to obtain the required data from the authors (Verigin, Meijer, & Vrij, 2020).

In the end, 14 studies were included in the meta-analysis, and all of them employed a between-subjects design for veracity (participants were either asked to tell the truth or to lie). Three articles included two independent subgroups, where one was exposed to a manipulation (e.g., the model statement) and the other was not (Vrij, Leal, Fisher, et al., 2018; Vrij et al., 2017; Vrij, Mann et al., 2020). Two studies included three recalls and four subgroups, where one subgroup was exposed to a manipulation only the first time they recalled the event, one was exposed to a manipulation only the second time they recalled the event (with a one-week delay), one was exposed to a manipulation at both times, and one at neither (see Appendix B, for an explanation concerning how we treated these groups in the analyses) (Deeb et al., 2020; Deeb et al., 2021). One article included two studies with independent samples (Vrij, Leal, et al., 2020), of which the second study included three subgroups. Also, Leal et al. (2019) and Vrij, Leal, Fisher, Mann, Jo, et al. (2019) included three subgroups. In Vrij, Leal, Deeb, et al. (2020), Leal et al. (2019), and Vrij, Leal, Fisher, Mann, Jo, et al. (2019) the three subgroups differed in their exposure (or not) to a specific manipulation technique. All subgroups of these studies were therefore independent. Appendix B reports the characteristics of all studies and specific notes for each of them, including how the independent subgroups were treated for the analyses. Complications were measured in every study, but the other variables were not (see Appendix C).

In the end, we analysed the 14 articles that fit eligibility criteria. Appendix C reports main characteristics of included studies, which are marked in the reference list with an asterisk.

### Overall Effect Sizes Estimation

**Initial recall.** A random-effects meta-analysis of the 17 samples related to the "frequency of total details" (*N* = 2,083) showed that truth tellers reported more total details than lie tellers, with a moderate effect size (Cohen, 1988), *d* = 0.45, 95% CI [0.32, 0.59], *p* < .001, $PS_{ES}$ = .251, PIS = .326 (see forest plot in Appendix D). The meta-analysis also showed moderate heterogeneity, Q(16) = 36.07, *p* < .01, $I^2$ = 55.65. When exploring the residuals of each study and their significance, one study appeared to be an outlier (Vrij, Mann, et al., 2020). A sensitivity analysis showed that the effect size ranged from *d* = 0.41, 95% CI [0.29, 0.53], *p* < .001, when the outlier study was excluded, to *d* = 0.48, 95% CI [0.35, 0.61], *p* < .001. Moderator's effects for manipulation (*k* = 13 for "manipulation absent", *d* = 0.44 and *k* = 4 for "manipulation present", *d* = 0.53) and scenario (*k* = 13 for "past trip event", *d* = 0.40 and *k* = 4 for "spy mission", *d* = 0.63) were not significant (Table 1), indicating that neither of the two moderators had a significant effect on the frequency of total details provided by truth tellers vs. lie tellers. The funnel plot showed to be asymmetric to some degree (Appendix E), and the trim and fill method trimmed two studies. Yet, both observed (*d* = 0.45) and estimated (*d* = 0.40, $PS_{ES}$ = .221 and PIS = .345) effect sizes were in the moderate effect size region.

**Table 1.** Initial Recall for Total details - Moderator Analysis

| Variable | $Q_{B(df)}$ | *k* | *d* | 95% CI | $Q_{W(df)}$ |
|---|---|---|---|---|---|
| Manipulation | $0.47_{(1)}$ | | | | |
| No | | 13 | $0.44^{***}$ | [0.27, 0.60] | $34.78_{(12)}^{**}$ |
| Yes | | 4 | $0.53^{***}$ | [0.33, 0.72] | $0.15_{(3)}$ |
| Scenario | $1.12_{(1)}$ | | | | |
| Past trip event | | 13 | $0.40^{***}$ | [0.28, 0.51] | $17.95_{(12)}$ |
| Spy mission | | 4 | $0.63^{**}$ | [0.21, 1.05] | $12.55_{(3)}^{**}$ |

*Note.* A positive *d* indicates that truth tellers reported more details than lie tellers; CI = confidence interval; $Q_B$ = heterogeneity between factors levels; $Q_W$ = heterogeneity within factors levels.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

A random-effects meta-analysis of the 18 samples for "frequency of complications" (*N* = 2,163) showed that truth tellers reported more complications than lie tellers, with a moderate effect size of *d* = 0.58, 95% CI [0.48, 0.68], *p* < .001, $PS_{ES}$ = .318, PIS = .281 (Appendix D). This supports Hypothesis 1. Q statistic was not significant, Q(17) = 21.32,

$p$ = .21, and $I^2$ = 20.25 showed low heterogeneity. Due to the lack of heterogeneity there is no need to conduct moderator analyses neither for manipulation ($k$ = 14 absent, $k$ = 4 present) nor for scenario ($k$ = 14 past trip event, $k$ = 4 spy mission). Sensitivity analyses showed that effect sizes ranged from $d$ = 0.56, 95% CI [0.47, 0.66], $p$ < .001, to $d$ = 0.62, 95% CI [0.53, 0.71], $p$ < .001, when the only outlier study (Vrij, Leal, Fisher, Mann, Deeb, et al., 2019) was excluded. The funnel plot was asymmetric (Appendix E) and three studies were trimmed. Yet, observed ($d$ = 0.58) and estimated ($d$ = 0.54) effect sizes were comparable.

A random-effects meta-analysis on the 12 samples for "frequency of common knowledge details" ($N$ = 1,498) showed that truth tellers reported fewer common knowledge details than lie tellers, with a moderate effect size of $d$ = -0.40, 95% CI [-0.52, -0.27], $p$ < .001, $PS_{ES}$ = .221, PIS = .655 (Appendix D). This supports Hypothesis 2. There was low heterogeneity between studies, $I^2$ = 26.91, Q(11) = 15.05, $p$ = .18. Due to the lack of heterogeneity, there is no need to conduct moderator analyses neither for manipulation ($k$ = 8 absent, $k$ = 4 present) nor for scenario (all studies belonged to the past trip event category). Sensitivity analyses showed that the effect ranged from $d$ = -0.37, 95% CI [-0.49, -0.25], $p$ < .001 to $d$ = -0.44, 95% CI [-0.55, -0.33], $p$ < .001, when the only outlier study was excluded (Vrij, Leal, Fisher, Mann, Deeb, et al., 2019). The funnel plot did not show a clear asymmetry, and two studies were trimmed (Appendix E). Again, observed ($d$ = -0.40) and estimated ($d$ = -0.36) effect sizes were comparable.

Finally, a random effects meta-analysis on the 13 samples for the "frequency of self-handicapping strategy" ($N$ = 1,620) showed that truth tellers reported fewer self-handicapping strategies than lie tellers with a small effect size of $d$ = -0.37, 95% CI [-0.53, -0.20], $p$ < .001 (Appendix D). This supports Hypothesis 3. There was moderate heterogeneity between studies, $I^2$ = 63.63, Q(12) = 32.99, $p$ < .01. Moderator effects for manipulation ("manipulation absent" $k$ = 9, "manipulation present" $k$ = 4, $Q_{between(1)}$ = 0.49, $p$ = .48) and scenario (past trip event $k$ = 12, spy mission $k$ = 1, $Q_{between(1)}$ = 0.68, $p$ = .41) were not significant. Sensitivity analyses showed that the effect size ranged from $d$ = -0.31, 95% CI [-0.44, -0.17], $p$ < .001, when the only outlier record was excluded (Vrij et al., 2017, subgroup 2, manipulation present), to $d$ = -0.41, 95% CI [-0.57, -0.24], $p$ < .001. The funnel plot did not show a clear asymmetry and no study was trimmed (Appendix E).

**Second recall.** A random-effects meta-analysis on the 17 samples focusing on "new total detail" ($N$ = 1,658) showed that truth tellers reported more new details than lie tellers with a small effect size of $d$ = 0.28, 95% CI [0.17, 0.39], $p$ < .001 (Appendix F). There was low heterogeneity, $I^2$ = 27.22, Q(16) = 21.98, $p$ = .14. Hence, this means that there is no need to conduct any moderator analysis for manipulation ($k$ = 6 absent, $k$ = 11 present) nor for scenario ($k$ = 13 past trip event, $k$ = 4 spy mission). Further, sensitivity analyses showed that the effect size ranged from $d$ = 0.25, 95% CI [0.16, 0.35], $p$ < .001, when the only outlier study was excluded (Vrij, Mann, et al., 2020, subgroup 2, manipulation present), to $d$ = 0.30, 95% CI [0.19, 0.41], $p$ < .001. The funnel plot appeared symmetric and no study was trimmed (Appendix G).

A random-effects meta-analysis on the 17 samples focusing on "new complications" ($N$ = 1,658) indicated that truth tellers reported more new complications than lie tellers with a moderate effect size, $d$ = 0.51, 95% CI [0.42, 0.61], $p$ < .001, $PS_{ES}$ = .281, PIS = .305 (Appendix F). This supports Hypothesis 1. There was no heterogeneity, $I^2$ = 0.00, Q(16) = 13.01, $p$ = .67; hence, no moderator analysis was conducted for manipulation ($k$ = 6 absent, $k$ = 11 present) nor for scenario ($k$ = 13 past trip event, $k$ = 4 spy mission). There were no outlier studies and the effect size ranged from $d$ = 0.50, 95% CI [0.40, 0.60], $p$ < .001 to $d$ = 0.53, 95% CI [0.43, 0.63], $p$ < .001. The trim and fill method trimmed no study and the funnel plot appeared symmetrical (Appendix G).

A fixed effect meta-analysis on the 10 samples focusing on "new common knowledge details" ($N$ = 1,149) showed that truth tellers report fewer new common knowledge details than lie tellers with a moderate effect size, $d$ = -0.46, 95% CI [-0.58, -0.35], $p$ < .001, $PS_{ES}$ = .259, PIS = .677 (Appendix F). This supports Hypothesis 2. There was no heterogeneity, $I^2$ = 0.00, Q(9) = 5.85, $p$ = .75; hence no moderator analysis was conducted for manipulation ($k$ = 4 absent, $k$ = 6 present) nor for scenario (all studies belonged to the past event trip category). No study was an outlier and a sensitivity analysis showed that the effect size ranged from $d$ = -0.45, 95% CI [-0.57, -0.32], $p$ < .001, to $d$ = -0.50, 95% CI [-0.62, -0.38], $p$ < .001. One study was trimmed via the trim and fill method (Appendix G), and the observed ($d$ = -0.46) and the estimated ($d$ = -0.45) were almost identical.

A fixed-effect meta-analysis on the 10 samples focusing on "new self-handicapping strategies" ($N$ = 1,067) showed that truth tellers report fewer new self-handicapping strategies than lie tellers with a moderate effect size, $d$ = -0.50, 95% CI [-0.56, -0.44], $p$ < .001, $PS_{ES}$ = .274, PIS = .691 (Appendix F). This supports Hypothesis 3. There was moderate heterogeneity, $I^2$ = 55.66, Q(9) = 20.30, $p$ < .05. There was a significant effect for manipulation, $Q_{between(1)}$ = 7.41, $p$ < .01. Studies in the "manipulation absent" category ($k$ = 4) obtained a smaller effect size ($d$ = -.23, 95% CI [-.43, -.04], $p$ < .05, $Q_{within(3)}$ = 1.58, $p_{Qwithin}$ = .66) than those in the "manipulation present" category ($k$ = 6, $d$ = -.52, 95% CI [-.59, -.46], $p$ < .001, $Q_{within(5)}$ = 11.31, $p$ = .05). The effect of the moderator "scenario" was not significant (past trip event $k$ = 8, spy mission $k$ = 2, $Q_{between(1)}$ = 1.81, $p_{Qwihin}$ = .18). There was one outlier study (Vrij, Leal, Jupe, et al., 2018). Sensitivity analyses showed that the effect ranged from $d$ = -0.27, 95% CI [-0.39, -0.15], $p$ < .001, when such study was excluded, to $d$ = -0.51, 95% CI [-0.57, -0.45], $p$ < .001. The trim and fill method showed that no study was trimmed (Appendix G).

**Total recall.** A series of meta-analyses for "total recall" showed a pattern that was similar to that of second recall (see Appendices H and I). Truth tellers reported more unique total details and unique complications but fewer unique common knowledge details and unique self-handicapping strategies than lie tellers. Heterogeneity analyses were not significant (Table 2).

## Discussion

The meta-analysis showed support for all three hypotheses and revealed that truth tellers reported more complications and fewer common knowledge details and self-handicapping strategies than lie tellers. Findings were very similar for first initial recall, the ond recall (where only new information after first recall was examined), and for total recall (initial and second recalls combined). The finding that the

**Table 2.** Meta-analysis for Total Recall Data

| | | | | | | Sensitivity analysis | | Trim and fill | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | $k$ | $N$ | $d$ [95% CI] | $Q_{(df)}$ | $I^2$ | Min $d$ [95% CI] | Max $d$ [95% CI] | Trimmed studies | Estimated effect size |
| Total details[1] | 15 | 1,454 | 0.45 [0.35, 0.55] | $13.93_{(14)}$ | 0.00 | 0.42 [0.32, 0.53] | 0.47 [0.36, 0.58] | 5 | $d$ = 0.37 |
| Complications[1] | 15 | 1,454 | 0.62 [0.49, 0.75] | $21.10_{(14)}$ | 33.64 | 0.59 [0.46, 0.71] | 0.65 [0.51, 0.78] | 4 | $d$ = 0.53 |
| Common knowledge details[2] | 8 | 945 | -0.43 [-0.56, -0.30] | $5.56_{(7)}$ | 0.00 | -0.39 [-0.53, -0.25] | -0.47 [-0.61, -0.33] | 0 | $d$ = -0.43 |
| Self-handicapping strategies[2] | 10 | 1,067 | -0.37 [-0.49, -0.25] | $7.21_{(9)}$ | 0.00 | -0.35 [-0.47, -0.22] | -0.40 [-0.52, -0.27] | 0 | $d$ = -0.37 |

*Note.* [1]random-effects model; [2]fixed effect model.

pattern of results obtained in a first recall tends to repeat itself in a second recall should increase confidence amongst practitioners when they use these variables in a two recall interview when attempting to detect deceit. It may also indicate robustness of the findings.

All effect sizes were moderate but they were somewhat larger for complications (*d* ranged from 0.51 to 0.62) than for common knowledge details (*d* ranged from -0.40 to -0.46) and self-handicapping strategies (*d* ranged from -0.37 to -.50). Further, the PIS indicated that a greater proportion of truth tellers would obtain the expected scores when focusing on complications (higher scores than lie tellers) than when focusing on common knowledge or self-handicapping strategies (lower scores than lie tellers). The relatively weaker results for common knowledge details and self-handicapping strategies (both cues to deceit) than for complications (cue to truthfulness) is unfortunate. The verbal deception field is in much stronger need of cues to deceit than cues to truthfulness because so few verbal cues to deceit do exist (Nahari et al., 2019).

Although complications can be coded in probably most statements, common knowledge details and self-handicapping strategies do not always occur (Vrij, Granhag, et al., 2021). They are typically examined in a 'travel' scenario where participants report a trip they allegedly have made in the last twelve months. Making a trip is arguably a somewhat scripted activity, which makes common knowledge details more likely to occur ("We visited the famous market, after which we went to the beach. We had dinner in a Mexican restaurant"). And when the trip was not recent, it gives lie tellers a good opportunity to include self-handicapping strategies ("I cannot remember which restaurants we went to in the evenings, because we went there three months ago"). The situation is different when someone describes a unique event that just happened. It seems reasonable to suggest that lie tellers are more likely to report common knowledge details when describing a somewhat scripted event than a unique event and that they are more likely to report self-handicapping strategies when describing an event that occurred in the past as opposed to an event that occurred recently.

Truth tellers may also include common knowledge details in their statements and perhaps particularly so when they do not see the relevance of describing the experience in more detail. These common knowledge details will be impossible to distinguish from those reported by lie tellers. A possible solution is to stress to interviewees that they should report every detail they can remember even the insignificant ones. An alternative solution is to expose interviewees to a model statement, an example of a detailed account (Leal et al., 2015), so that interviewees are made aware of the amount of detail they are expected to provide. A model statement has shown to be an effective method to generate information (Vrij, Leal, & Fisher, 2018).

Truth tellers may also admit lack of memory when an event happened some time ago and, as a result, may include self-handicapping strategies in their statement ("I cannot remember which restaurants we went to in the evenings, because we went there three months ago"). Admitting lack of memory is a CBCA criterion and truth tellers report those more frequently than lie tellers (Amado et al., 2016). There are two differences between self-handicapping strategies and admitting lack of memory. First, admitting lack of memory becomes a self-handicapping strategy only when it is followed by a justification: In the sentence above: "… because we were there three months ago." Thus, the sentence "I cannot remember which restaurants we went to in the evenings" would classify as admitting lack of memory in CBCA but does not constitute a self-handicapping strategy. Second, self-handicapping strategies do not just include admitting lack of memory; it also includes admitting lack of perceptual experiences. The example "There isn't much to say about the actual bungee jump as it took only a few moments" constitutes a self-handicapping strategy of the lack of perceptual experiences type. Perhaps a distinction between these two types of self-handicapping strategies, an admitting lack of memory type and an admitting lack of experiences type, may make the difference between truth tellers and lie tellers more pronounced. That is, perhaps truth tellers are more likely to include in their statements the memory-type of self-handicapping strategies than the experience-type. This suggests that the experience-type will be the strongest veracity indicator.

Complications was not only a stronger veracity indicator than common knowledge details and self-handicapping strategies, it was also a more diagnostic veracity indicator than total details. This was particularly the case in the second recall (*d* = 0.51 for complications and *d* = 0.28 for total details). The relatively low *d* score for total details in a second recall suggests that the finding obtained for complications (pattern of results obtained in a first recall repeats itself in a second recall) does not apply to total details to the same extent. This makes total details a more problematic cue to use for lie detection purposes than complications.

A possible benefit of total details is that the cue always can be examined, because even brief statements (perhaps with the exception of 'no comment') always include details. Very short statements may not include complications. Whether total details emerges as a strong veracity cue in short statements is an empirical question. This may not be the case, because verbal cues to deception are more likely to occur in longer statements because words are the carriers of verbal cues to veracity (Vrij et al. 2007). It is also an empirical question whether complications emerge as a stronger veracity indicator than total details in deception scenarios other than the ones examined in this meta-analysis. All we can conclude at this stage is that complications emerged as a stronger veracity indicator than total details in the scenarios examined in this meta-analysis.

The moderators did affect only the results for new self-handicapping strategies in the second recall. It means that effect sizes were mostly homogeneous across studies. Yet, this should be taken with some caution because of the (i) low number of studies and (ii) unbalanced groups (Borenstein et al, 2011). The absence of a moderator effect cannot be interpreted as evidence that interview techniques have no effect on the dependent variables (Moderator 4), because moderator analyses often have low power (Borenstein et al., 2011). To examine whether an interview technique has an effect on dependent variables someone should either analyse more studies or compare "within each experiment" an experimental ('technique present') condition with a control ('technique absent') condition. The latter is different from what happened in moderator analyses in the present meta-analysis. Here, across all samples included in the meta-analysis, 'technique absent' conditions were compared with 'technique present' conditions, but these absent and present conditions did not always belong to the same experiment.

We note four limitations. First, all the available research comes from Vrij's lab. This is not uncommon in deception research. For example, in a meta-analysis of Strategic Use of Evidence (SUE) research, Granhag was an author on every publication (Hartwig et al., 2014). It is even not uncommon in interviewing research. For example, also in a meta-analysis of the Scharff technique, Granhag was an author on every publication (Luke, 2021). Despite this, research carried out by other researchers seems essential. At present we cannot rule out that the way Vrij and colleagues operationalise and code the three variables is idiosyncratic and that this is driving the effects. The general lack of heterogeneity in the effects presented in this meta-analysis is unusual for deception research (DePaulo et al., 2003) and psychology research in general (Stanley et al., 2018). This could be due to small sample sizes (Borenstein et al., 2011, 2016). Another factor could be a lack of variance in deception scenarios in which cues have been examined. A wider spread of deception scenarios is thus welcome and the contribution of other researchers to this domain would facilitate this.

A second limitation is that all studies are lab-based studies but field studies testing hypotheses seem relevant. This could be a challenge due to the difficulty in obtaining ground truth in field studies. Third, the number of studies on which this meta-analysis was based was limited.

Although this is not uncommon in this field (Hartwig et al., 2014; Luke, 2021) more research is required, particularly regarding common knowledge details and self-handicapping strategies for second and total recalls as the number of included studies is insufficient.

Fourth, we stated that examining complications, common knowledge details, and self-handicapping details is advantageous compared to coding total details because the former three variables can be coded in real time whereas the latter variable cannot. Note that there is yet no empirical evidence that the former three variables can be coded in real time.

Several issues merit further research, such as in which types of setting complications, common knowledge details and self-handicapping strategies (1) can be examined and (2) yield the strongest effects. We already know that common knowledge details and self-handicapping strategies cannot be examined in certain situations and the search for alternative cues to deceit that occur in such settings seems urgent (Nahari et al., 2019). Another area of research is whether the three variables become more diagnostic if they are considered in relation to the type of detail they refer to. For example, truth tellers compared to lie tellers tend to include (1) more verifiable details in their statements (Palena et al., 2020) and (2) focus more on core aspects of an event (Sakrisvold et al., 2017). Are therefore complications reported in verifiable details and core events more diagnostic than complications reported in unverifiable details and peripheral events? In addition, as we already suggested above, it is worthwhile to compare how diagnostic these three variables are as veracity indicators compared to the total details variable.

Researchers should examine whether complications, common knowledge details, and self-handicapping strategies indeed can be counted in real time. In our training of practitioners, we focus on complications and self-handicapping strategies and our experience is that they can be counted in real time. However, we have never formally examined this. In addition, we have never examined whether practitioners can also count common knowledge details in real time. Finally, it seems likely that new verbal veracity indicators other than complications, common knowledge details, and self-handicapping strategies do exist. The field is in particular need of cues that lie tellers report more frequently than truth tellers (cues to deceit). In that respect, although it does not constitute a new cue to deceit, separating self-handicapping strategies into two types, one that does include admitting lack of memory and another type that includes admitting lack of perceptual experiences, may be a first step. We hope this meta-analysis stimulates researchers to address these and other issues.

## Conflict of Interest

The authors of this article declare no conflict of interest.

## References

*References marked with an asterisk indicate studies included in the metaanalysis.

Amado, B. G., Arce, R., Fariña, F., & Vilarino, 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

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis.* John Wiley & Sons.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2016). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5-18. https://doi.org/10.1002/jrsm.1230

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

Cooper, H. (2011). *Reporting research in psychology: How to meet journal article reporting standards.* American Psychological Association.

Cooper, H. (2015). *Research synthesis and meta-analysis: A step-by-step approach* (Vol. 2). Sage Publications.

*Deeb, H., Vrij, A., & Leal, S. (2020). The effects of a model statement on information elicitation and deception detection in multiple interviews. *Acta Psychologica, 207*, 103080. https://doi.org/10.1016/j.actpsy.2020.103080

*Deeb, H., Vrij, A., Leal, S., & Burkhardt, J. (2021). The Effects of sketching while narrating on information elicitation and deception detection in multiple interviews. *Acta Psychologica, 213*, 103236. https://doi.org/10.1016/j.actpsy.2020.103236

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*(5), 979-995. https://doi.org/10.1037/0022-3514.70.5.979

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://doi.org/10.1037/0033-2909.129.1.74

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2020). *A primer on bayesian model-averaged meta-analysis.* https://doi.org/10.31234/osf.io/97qup

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology, 2*(1), 123-138. https://doi.org/10.1080/23743603.2017.1326760

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic use of evidence during investigative interviews: The state of the science. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 1-36). Academic Press.

Hartwig, M., Granhag, P. A., & Strömwall, L. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime, & Law, 13*(2), 213-227. https://doi.org/10.1080/10683160600750264

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

*Leal, S., Vrij, A., Deeb, H., Hudson, C., Capuozzo, P., & Fisher, R. P. (2020). Verbal cues to deceit when lying through omitting information. *Legal and Criminological Psychology, 25*(2), 278-294. https://doi.org/10.1111/lcrp.12180

*Leal, S., Vrij, A., Deeb, H., & Kamermans, K. (2019). Encouraging interviewees to say more and deception: The ghostwriter method. *Legal and Criminological Psychology, 24*(2), 273-287. https://doi.org/10.1111/lcrp.12152

Leal, S., Vrij, A., Vernham, Z., Dalton, G., Jupe, L., Harvey, A., & Nahari, G. (2018). Cross-cultural verbal deception. *Legal and Criminological Psychology, 23*(2), 192-213. https://doi.org/10.1111/lcrp.12131

Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. (2015). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology, 20*(1), 129-146. https://doi.org/10.1111/lcrp.12017

Luke, T. J. (2021). A meta-analytic review of experimental tests of the interrogation technique of Hanns Joachim Scharff. *Applied Cognitive Psychology, 35*(2), 360-373. https://doi.org/10.1002/acp.3771

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International Journal of Surgery, 8*, 336-341. https://doi.org/10.1371/journal.pmed.1000097

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., Ashkenazi, T., Fisher, R. P., Granhag, P. A., Hershkovitz, I., Masip, J., Meijer, E., Nisin, Z., Sarid, N., Taylor, P. J., Verschuere, B., & Vrij, A. (2019). Language of lies: Urgent issues and prospects in verbal lie detection research. *Legal and Criminological Psychology, 24*(1), 1-23. https://doi.org/10.1111/lcrp.12148

Nahari, G., & Pazuelo, M. (2015). Telling a convincing story: Richness in detail as a function of gender and priming. *Journal of Applied Research in Memory and Cognition, 4*(4), 363-367. https://doi.org/10.1016/j.jarmac.2015.08.005

Nahari, G., & Vrij, A. (2014). Are you as good as me at telling a story? Individual differences in interpersonal-reality monitoring. *Psychology, Crime, & Law, 20*(6), 573-583. https://doi.org/10.1080/1068316X.2013.793771

Palena, N., Caso, L., Vrij, A., & Nahari, G. (2020). The verifiability approach: A meta-analysis. *Journal of Applied Research in Memory and Cognition.* Advance online publication. https://doi.org/10.1016/j.jarmac.2020.09.001

Ruby, C. L., & Brigham, J. C. (1998). Can criteria-based content analysis distinguish between true and false statements of African-American speakers? *Law and Human Behavior, 22*(4), 369-388. https://doi.org/10.1023/A:1025766825429

Sakrisvold, M. L., Granhag, P. A., & Mac Giolla, E. (2017). Partners under pressure: Examining the consistency of true and false alibi statements.

*Behavioral Sciences & the Law, 35*(1), 75-90. https://doi.org/10.1002/bsl.2275

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Erlbaum.

Sporer, S. L. (2016). Deception and cognitive load: Expanding our horizon with a working memory model. *Frontiers in Psychology: Hypothesis and Theory, 7,* 420. https://doi.org/10.3389/fpsyg.2016.00420

Stanley, T., Carter, E., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325-1346. https://doi.org/10.1037/bul0000169

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428-453. https://doi.org/10.1037/bul0000087

Verigin, B. L., Meijer, E. H., & Vrij, A. (2020). Embedding lies into truthful stories does not affect their quality. *Applied Cognitive Psychology, 34*(2), 516-525. https://doi.org/10.1002/acp.3642

Verigin, B. L., Meijer, E. H., Vrij, A., & Zauzig, L. (2020). The interaction of truthful and deceptive information. *Psychology, Crime & Law, 26*(4), 367-383. https://doi.org/10.1080/1068316X.2019.1669596

Vernham, Z., Vrij, A., Nahari, G., Leal, S., Mann, S., Satchell, L., & Orthey, R. (2020). Applying the verifiability approach to deception detection in alibi witness situations. *Acta Psychologica, 204,* 103020. https://doi.org/10.1016/j.actpsy.2020.103020

Volbert, R., & Steller, M. (2014). Is this testimony truthful, fabricated, or based on false memory? Credibility assessment 25 years after Steller and Köhnken (1989). *European Psychologist, 19*(3), 207-220. https://doi.org/10.1027/1016-9040/a000200

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities, second edition.* John Wiley and Sons.

Vrij, A. (2016). Baselining as a lie detection method. *Applied Cognitive Psychology, 30*(6), 1112-1119. https://doi.org/10.1002/acp.3288

Vrij, A., Deeb, H., Leal, S., Granhag, P. A., & Fisher, R. P. (2020). Plausibility: A verbal cue to veracity worth examining? *The European Journal of Psychology Applied to Legal Context.* Advance online publication. https://doi.org/10.5093/ejpalc2021a4

Vrij, A., Granhag, P. A., Leal, S., Fisher, R. P., Kleinman, S. M., & Ashkenazi, T. (2021). The present and future of verbal lie detection. In D. Matteo & K. C. Scherr (Eds.), *The Oxford handbook of psychology and law.* Oxford Press.

Vrij, A., & Leal, S. (2020). Proportion of complications in interpreter-absent and interpreter-present interviews. *Psychiatry, Psychology and Law, 27*(1), 155-164. https://doi.org/10.1080/13218719.2019.1705197

*Vrij, A., Leal, S., Deeb, H., Chan, S., Khader, M., Chai, W., & Chin, J. (2020). Lying about flying: The efficacy of the information protocol and model statement for detecting deceit. *Applied Cognitive Psychology, 34*(1), 241-255. https://doi.org/10.1002/acp.3614

Vrij, A., Leal, S., & Fisher, R. P. (2018). Verbal deception and the model statement as a lie detection tool. *Frontiers in Psychiatry, 9*(492). https://doi.org/10.3389/fpsyt.2018.00492

*Vrij, A., Leal, S., Fisher, R. P., Mann, S., Dalton, G., Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2018). Sketching as a technique to eliciting information and cues to deceit in interpreter-based interviews. *Journal of Applied Research in Memory and Cognition, 7*(2), 303-313. https://doi.org/10.1016/j.jarmac.2017.11.001

*Vrij, A., Leal, S., Fisher, R. P., Mann, S., Deeb, H., Jo, E., Castro Campos, C., & Hamzeh, S. (2019). The efficacy of using countermeasures in a model statement interview. *The European Journal of Psychology Applied to Legal Context, 12*(1), 23-34. https://doi.org/10.5093/ejpalc2020a3

*Vrij, A., Leal, S., Fisher, R. P., Mann, S., Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2019). Eliciting information and cues to deceit through sketching in interpreter-based interviews. *Applied Cognitive Psychology, 33*(6), 1197-1211. https://doi.org/10.1002/acp.3566

*Vrij, A., Leal, S., Jupe, L., & Harvey, A. (2018). Within-subjects verbal lie detection measures: A comparison between total detail and proportion of complications. *Legal and Criminological Psychology, 23*(2), 265-279. https://doi.org/10.1111/lcrp.12126

*Vrij, A., Leal, S., Mann, S., Dalton, G., Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2017). Using the model statement to elicit information and cues to deceit in interpreter-based interviews. *Acta Psychologica, 177,* 44-53. https://doi.org/10.1016/j.actpsy.2017.04.011

*Vrij, A., Leal, S., Mann, S., Fisher, R. P., Dalton, G., Jo, E., Shaboltas, A., Khaleeva, M., Granskaya, J., & Houston, K. (2018). Using unexpected questions to elicit information and cues to deceit in interpreter-based interviews. *Applied Cognitive Psychology, 32*(1), 94-104. https://doi.org/10.1002/acp.3382

Vrij, A., Leal, S., Mann, S., Shaboltas, A., Khaleeva, M., Granskaya, J., & Jo, E. (2019). Using the model statement technique as a lie detection tool: A cross-cultural comparison. *Psychology in Russia: State of the Art, 12*(2), 19-33. https://doi.org/10.11621/pir.2019.0202

*Vrij, A., Leal, S., Mann, S., Vernham, Z., Dalton, G., Serok-Jeppa, O., Rozmann, N., Nahari, G., & Fisher, R. P. (2020). "Please tell me all you remember": A comparison between British' and Arab' interviewees' free narrative performance and its implications for lie detection. *Psychiatry, Psychology and Law.* Advance online publication. https://doi.org/10.1080/13218719.2020.1805812

Vrij, A., Mann, S., Kristen, S., & Fisher, R. (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., Mann, S., Leal, S., & Fisher, R. P. (2021). Combining verbal veracity assessment techniques to distinguish truth tellers from lie tellers. *European Journal of Psychology Applied to Legal Context, 13*(1), 9-19. https://doi.org/10.5093/ejpalc2021a2

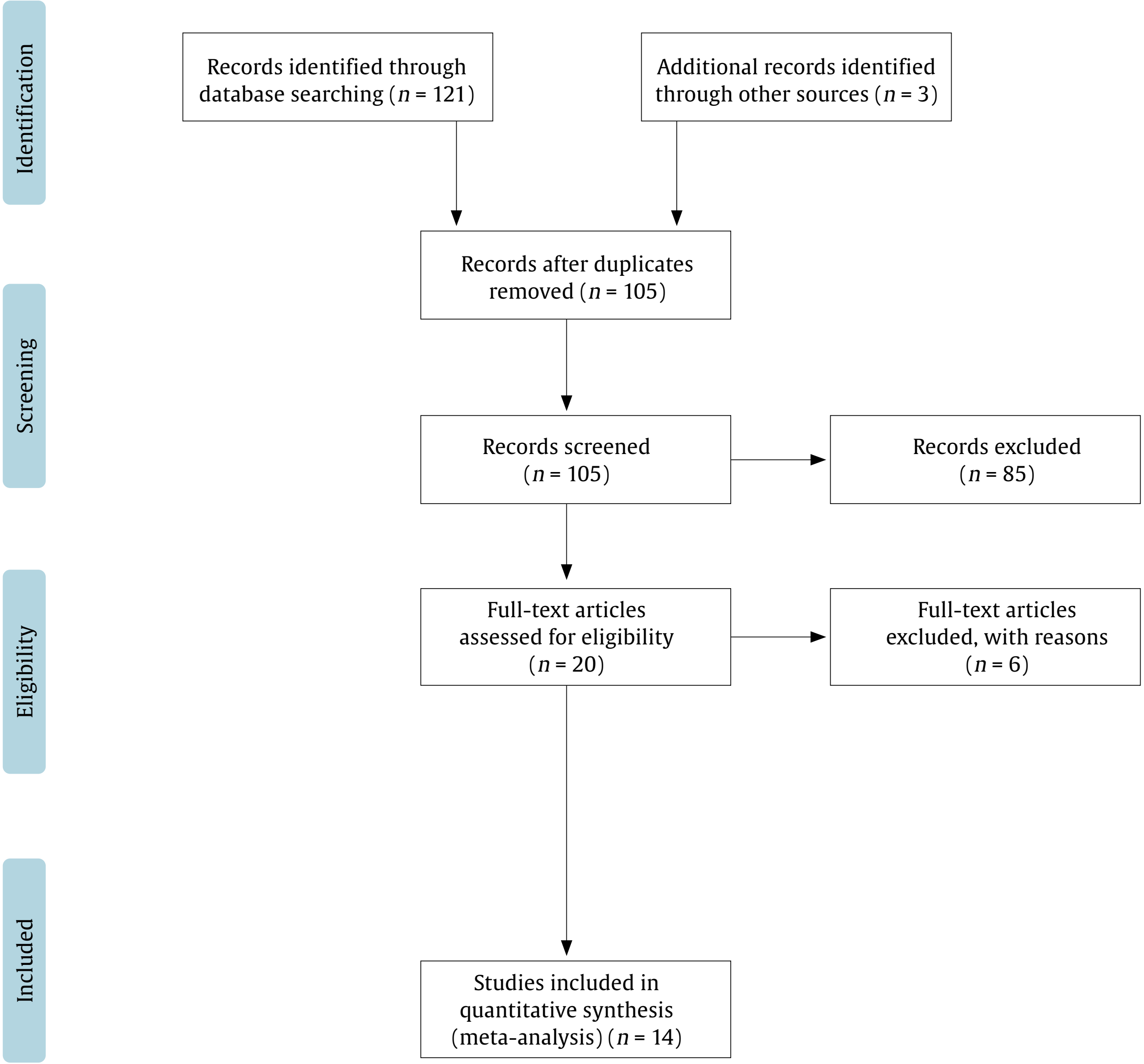
*Vrij, A., Mann, S., Leal, S., Fisher, R. P., & Deeb, H. (2020). Sketching while narrating as a tool to detect deceit. *Applied Cognitive Psychology, 34*(3), 628-642. https://doi.org/10.1002/acp.3646

Vrij, A., & Vrij, S. (2020). Complications travel: A cross-cultural comparison of the proportion of complications as a verbal cue to deceit. *Journal of Investigative Psychology and Offender Profiling, 17*(1), 3-16. https://doi.org/10.1002/jip.1538

## Appendix A

Prisma Flowchart

PRISMA 2009 Flow Diagram


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 121)
Additional records identified through other sources (n = 3)
Records after duplicates removed (n = 105)
Records screened (n = 105)
Records excluded (n = 85)
Full-text articles assessed for eligibility (n = 20)
Full-text articles excluded, with reasons (n = 6)
Studies included in quantitative synthesis (meta-analysis) (n = 14)

## Appendix B

Article Notes and Reasons for Excluded Sources (continued)

| Study | Notes | Data requested to authors |
|---|---|---|
| Deeb et al. (2020) | In this study participants were interviewed three times, each one week apart. T1 is the immediate interview, T2 is the second interview (after one week) and T3 is the final interview (after another week). Participants could be exposed to a Model Statement at T1, T2, T1 and T2 or not at all. We considered T1 as "Initial recall" and T2 as "Second recall". "Total recall" are the total of unique complications etc. reported during the entire interview (all repetitions disregarded).<br>For T1 ("Initial recall") we report an effect size for the two subgroups that received the Model Statement at T1 ("Manipulation present") and an effect size for the two subgroups who did not receive the Model Statement at T1 ("Manipulation absent").<br>For T2 ("Second recall") we report an effect size for participants that received the Model Statement at T2 ("Manipulation present") and an effect size for those that did not receive the MS at T2 ("Manipulation absent").<br>For T2 "Total recall", we report an effect size for participants that received the Model Statement at T1, T2 or both at T1 and T2 ("Manipulation present") and an effect size for those who did not receive the Model Statement at either T1 nor T2 ("Manipulation absent"). | Initial Recall (T1 data)<br>Effect sizes and confidence intervals for all variables for:<br>Subgroups 1 (MS at T1) and 3 (MS at T1 and T2) combined<br>Subgroups 2 (MS at T2) and 4 (control, no MS) combined<br>Second Recall (T2 data)<br>Effect sizes and confidence intervals for all variables (new details) for:<br>Subgroups 2 (MS at T2) and 3 (MS at T1 and T2) combined<br>Subgroup 1 (MS at T1) and 4 (control, no MS)<br>Total Recall (T1 + T2 data)<br>Effect sizes and confidence intervals for all variables (unique details) for:<br>Subgroups 1 (MS at T1) 2 (MS at T2) and 3 (MS at T1 and T2) combined.<br>Subgroup 4 (control, no MS) |
| Deeb et al. (2021) | This study used the same protocol as Deeb et al. (2020) except that the manipulation involved the Sketching technique rather than the Model Statement technique. Hence, we reported the effect sizes in the same way as we did for Deeb et al. (2020). | Initial Recall (T1 data)<br>Effect sizes and confidence intervals for all variables for:<br>Subgroups 1 (Sketching at T1) and 3 (Sketching at T1 and T2) combined.<br>Subgroups 2 (Sketching at T2) and 4 (control, no Sketching) combined.<br>Second Recall (T2 data)<br>Effect sizes and confidence intervals for all variables (new details) for:<br>Subgroups 2 (Sketching at T2) and 3 (Sketching at T1 and T2) combined.<br>Subgroups 1 (Sketching at T1) and 4 (control, no Sketching) combined.<br>Total Recall (T1 + T2 data)<br>Effect sizes and confidence intervals for all variables (unique details) for:<br>Subgroups 1 (Sketching at T1) 2 (Sketching at T2) and 3 (Sketching at T1 and T2) combined.<br>Subgroup 4 (control, no Sketching) |
| Leal et al. (2020) | Participants were interviewed twice. In Interview 2 they recalled what they had discussed in Interview 1 (hence Interview 1 was the stimulus material). In Interview 2 they firstly answered a free recall question ("Initial recall") followed by a Model Statement followed by a second free recall question ("Second recall"). Answers to the two phases of Interview 2 were used in the present study. For Initial recall, we coded Manipulation as "absent". For Second and Total recalls we coded Manipulation as "present". The authors did not report common knowledge details and self-handicapping strategies as these details did not occur frequently enough. | Total Recall<br>Confidence intervals for Cohen's d for both unique total detail and unique complications. |
| Leal et al. (2019) | The participants provided one statement, hence, for this experiment only the data for "Initial recall" are used. Manipulation was coded as "present" for the "ghost writer" group and as "absent" for the other two groups (no instruction and 'be detailed' instruction) combined. | *n* for truth tellers and n for lie tellers for control condition and be detailed condition combined.<br>Effect sizes and confidence intervals for all variables for control and be detailed conditions combined.<br>*n* for truth tellers and *n* for lie tellers for "ghost writer" condition. |
| Verigin, Meijer, & Vrij (2020) | Excluded. It was not possible to obtain the necessary information from the authors | Effect sizes and confidence intervals accounting for the entire statement for truth tellers and for the two lie tellers conditions combined. |
| Verigin, Meijer, Vrij, et al. (2020) | Excluded: The author coded complications but did not report them as they did not occur frequently enough. | |
| Vrij and Leal (2020) | Excluded: This paper does not contain new data. It aggregates the data from Vrij et al. (2017) and from Vrij, Leal, Fisher, et al. (2018). | |
| Vrij, Leal, Deeb, et al. (2020, Study 1) | Participants (real airport travellers) were interviewed about their trip by an immigration officer. They provided only one statement. Hence, for this experiment only the data for "Initial recall" are used. Complications was the only variable examined. No manipulation took place, hence for this study we coded Manipulation as "absent". | |

## Appendix B

Article Notes and Reasons for Excluded Sources

| Study | Notes | Data requested to authors |
|---|---|---|
| Vrij, Leal, Deeb, et al. (2020, Study 2) | Phase 1 of Study 2 was largely identical to Study 1. Study 2 also included a Phase 2, where participants were interviewed by a second interviewer after the manipulation (e.g., Model Statement) took place. The authors did not examine common knowledge details and self-handicapping strategies. For the Initial recall we report an effect size for the whole sample (101 truth tellers and 107 lie tellers) because there was no manipulation at Phase 1. Hence, for Initial recall Manipulation was coded as "absent". For Second and Total recalls, the interviewee could be exposed to no Manipulation (coded as "Manipulation absent"), a Model Statement (coded as "Manipulation present"), or a Model Statement plus an Information Protocol (coded as "Manipulation present"). | Second Recall<br>Effect sizes and confidence intervals for new total detail for each subgroup.<br>*n* for truth tellers and *n* for lie tellers for control condition<br>*n* for truth tellers and *n* for lie tellers for IP condition<br>*n* for truth tellers and *n* for lie tellers for IP + MS condition<br>Total Recall<br>Effect sizes and confidence intervals for unique total detail for each subgroup.<br>Effect sizes and confidence intervals for unique complications for each subgroup. |
| Vrij, Leal, and Fisher (2018) | Excluded: The paper is theoretical rather than empirical | |
| Vrij, Leal, Fisher, et al. (2018)[1] | This paper focuses on the sixth question asked of the interview protocol; the first five questions are reported in Vrij, Leal, Mann, et al. (2018) (see below). Half of the participants were asked to sketch while answering this question, whereas the other half were not asked to sketch. We coded this data as Second recall for the two subgroups (Manipulation present vs. Manipulation absent) independently. The authors did not examine self-handicapping strategies. | Second Recall<br>*n* for truth tellers and *n* for lie tellers for control condition<br>*n* for truth tellers and *n* for lie tellers for the Sketching condition.<br>Effect sizes and confidence intervals for all variables for the control condition.<br>Effect sizes and confidence intervals for all variables for the Sketching condition |
| Vrij, Leal, Fisher, Mann, Debb, et al. (2019) | In this study coaching was manipulated. Participants were either coached or not about the Model Statement and the Type of Details the interview would focus on. The whole sample was included in this meta-analysis (both the coached group and the control group) but the effect of coaching was not considered. First, when considering the whole interview (before and after the Model Statement was employed), there was no significant difference between the two coaching conditions on the various dependent variables. Second, there is no other study on the effect of coaching on the model statement and the type of details (e.g., complications). Hence, it would not be possible to examine the moderating role of this type of coaching. The Model Statement was applied before the Second recall. Hence, we coded Initial recall as "Manipulation absent" and Second and Total recalls as "Manipulation present". | Initial Recall<br>Effect sizes and confidence intervals for all variables considering the whole sample.<br>Second Recall<br>Effect size and confidence interval for new low and medium/high complication combined considering the whole sample.<br>Total Recall<br>Effect size and confidence interval for unique low and medium/high complication combined considering the whole sample. |
| Vrij, Leal, Fisher, Mann, Jo, et al. (2019) | In this study the participants were given the opportunity to discuss up to four events. To maintain the independency of the data, we only focused on the first event they discussed. Initial recall (phase 1 of the experiment) was provided before the sketching manipulation took place and focused on the entire event. As for Vrij, Leal, Deeb, et al. (2020, Study 2), we report an effect size for the whole sample. Second recall focused on a particularly memorable event of the entire story. For Second and Total recalls, participants could be exposed to no manipulation (coded as "Manipulation absent"), to a standard Sketching manipulation or to a Model Sketching manipulation. We combined the two Sketching condition (there was no difference between the two) and coded them as "Manipulation present". | Second Recall<br>*n* for truth tellers, *n* for lie tellers, effect sizes and confidence intervals for all variables for control condition.<br>*n* for truth tellers, *n* for lie tellers, effect sizes and confidence intervals for all variables for the two Sketching condition (Standard Sketching and Model Sketching) combined.<br>Total Recall<br>Effect sizes and confidence intervals for all variables for control condition.<br>Effect sizes and confidence intervals for all variables for the two Sketching condition (Standard Sketching and Model Sketching) combined. |
| Vrij, Leal, Jupe, et al. (2018) | Participants were interviewed about a trip they had made in the last 12 months. They reported the same event twice: Once before and once after the Model Statement. The first time the participants recalled the event is Initial recall (coded as "Manipulation absent"). The second time the participants recalled the event, after being exposed to the Model Statement, is Second recall. Hence, Second and Total recalls were coded as "Manipulation present". | Effect size and confidence intervals for Total Recall |
| Vrij et al. (2017)[2] | Participants of different ethnicities were interviewed about a trip they had made, either in their first language or through an interpreter. Since the interpreter manipulation was not significant, we considered the two interpreter conditions as a whole (interpreter present and interpreter absent combined). The authors also manipulated the presence vs. absence of a Model Statement, so that one subgroup was exposed whereas the other was not exposed to such a technique. We considered the two Model Statements subgroups independently. The group that was not exposed to the Model Statement was coded as "Manipulation absent", whereas the group that was exposed to the Model Statement was coded as "Manipulation present" | *n* for truth tellers and *n* for lie tellers for control condition<br>*n* for truth tellers and *n* for lie tellers for MS condition<br>Effect sizes and confidence intervals for total detail and self-Handicapping for:<br>Subgroups 1 (control)<br>Subgroup 2 (MS condition) |

## Appendix B

Article Notes and Reasons for Excluded Sources

| Study | Notes | Data requested to authors |
|---|---|---|
| Vrij, Leal, Mann, et al. (2018)[1] | In this paper, interviewees initially answered five questions about the planning and the execution of a trip they allegedly made. They then they answered a sixth (final) question focusing on the best thing that happened to them during the trip. The article only reported the results for the initial five questions and the answer to that sixth question was reported in Vrij, Leal, Fisher, et al. (2018). Here, we focused on the whole sample and coded Manipulation as "absent", as the sketching was introduced only after the fifth question. | Effect sizes and confidence intervals for all variables considering the whole sample (unexpected and expected questions combined). |
| Vrij, Leal, Mann, et al. (2019)[2] | Excluded: This paper does not contain new data as it is based on the dataset in Vrij et al. (2017). | |
| Vrij, Leal, Mann, et al. (2020)[3] | The British participants of this study were taken from the dataset used in Vrij, Mann, et al. (2020), whereas the Arab participants were new participants. Thus, for this study we only report data from the Arab participants. The authors did not examine common knowledge details and self-handicapping strategies. No interview technique was introduced so we coded Manipulation as "absent". | |
| Vrij, Mann, et al. (2021) | In this study, the authors did not examine common knowledge details and self-handicapping strategies. Further, this experiment explored the effect of applying multiple techniques in one interview. Hence, it was based on several recalls, one after each specific technique was introduced. Consequently, the effect size for Total recall is based on several recalls and several manipulation tactics. We coded Initial recall as "Manipulation absent" as no technique was introduced at this time. Interviewees were exposed to a Model Statement before Second recall. Hence, we coded Second recall as "Manipulation present". Interviewees were then exposed to three other techniques (reverse order, sketching, indication to provide checkable sources). Hence, we coded Total recall (that was the last interview that took place, after all techniques were employed) as "Manipulation present". | |
| Vrij, Mann, et al. (2020)[3] | In this experiment the participants completed a mock mission which included receiving a package from an agent. The Initial recall focused on the whole mission, whereas the Second recall focused only on the time of the package exchange. There was no manipulation at phase 1. Hence, as we did for Vrij, Leal, Deeb, et al. (2020, Study 2) and Vrij, Leal, Fisher, Mann, Jo, et al. (2019) for Initial recall, we report an effect size for the whole sample. After Initial recall, participants were either exposed (coded as "Manipulation present") or not (coded as "Manipulation absent") to the Sketching manipulation. Hence, for Second and Total recalls, we obtained an effect size for the Sketching-present condition and an effect size for the Sketching-absent condition. Further, the authors did not examine common knowledge details. Last, the author coded the details for different themes of the statements separately (e.g., information about the route of the trip and information about the location). We report effect size for all themes combined. | Initial Recall<br>Effect size and confidence intervals for details and complications, considering the whole sample.<br>Effect size and confidence interval for self-handicapping for the whole sample.<br>Second Recall<br>Effect sizes and confidence intervals for all variables (new total detail, new complications, new self-handicapping strategies) for each subgroup (control, Sketching) for all themes concerning the time of the exchange combined<br>*n* for truth tellers and *n* for lie tellers for control condition.<br>*n* for truth tellers and *n* for lie tellers for Sketching condition<br>Total Recall<br>Effect sizes and confidence intervals for all variables (unique total detail, unique complications, unique self-handicapping strategies) for each subgroup (control, Sketching) for all themes concerning the time of the exchange combined. |
| Vrij and Vrij (2020) | Excluded: This paper does not contain new data. It aggregates the data from Vrij, Leal, Fisher, Mann, Jo, et al. (2019), Vrij et al. (2017), and Vrij, Leal, Fisher, et al. (2018). | |

*Note.* Superscripts indicate that the papers were based on the same dataset ([1], [2]), or on a partial overlap of the datasets ([3]).

## Appendix C1

### Study Characteristics – Initial Recall

| Study | Truth tellers *n* | Lie tellers *n* | *N* | Initial Recall *d* [95% CI] | | | | Scenario | Manipulation[1] | Incentive | Modality |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Total details | Complications | Common knowledge details | Self-handicapping strategies | | | | |
| Deeb et al. (2020, subgroups 1 and 3 - MS at T1 and MS at T1 and T2) | 59 | 62 | 121 | 0.57 [0.21, 0.93] | 0.53 [0.17, 0.89] | -0.18 [-0.53, 0.17] | -0.15 [-0.50, 0.21] | Past trip event | MS (P) | yes | Oral |
| Deeb et al. (2020, subgroups 2 and 4 –MS at T2 and control) | 61 | 61 | 122 | 0.48 [0.12, 0.83] | 0.46 [0.10, 0.81] | -0.40 [-0.75, -0.04] | -0.30 [-0.65, 0.06] | Past trip event | None (A) | yes | Oral |
| Deeb et al. (2021, subgroups 1 and 3 - Sketching at T1 and Sketching at T1 and T2) | 60 | 60 | 120 | 0.50 [0.14, 0.85] | 0.69 [0.32, 1.05] | -0.53 [-0.89, -0.17] | -0.26 [-0.61, 0.10] | Past trip event | Sketching (P) | yes | Oral |
| Deeb et al. (2021, subgroups 2 and 4 - Sketching at T2 and control) | 62 | 61 | 123 | 0.39 [0.04, 0.74] | 0.63 [0.27, 0.98] | -0.51 [-0.86, -0.15] | -0.19 [-0.54, 0.16] | Past trip event | None (A) | yes | Oral |
| Leal et al. (2020) | 44 | 41 | 85 | 0.04 [-0.38, 0.46] | 0.34 [-0.08, 0.77] | did not occur frequently enough | did not occur frequently enough | Spy mission | None (A) | yes | Oral |
| Leal et al. (2019, subgroup 1 and 2 combined – no manipulation and "Be detailed") | 48 | 51 | 99 | 0.56 [0.16, 0.95] | 0.87 [0.47, 1.28] | -0.67 [-1.34, -0.01] | 0.11 [-0.27, 0.49] | Past trip event | None (A) | yes | Oral |
| Leal et al. (2019, subgroup 3 – Ghost Writer) | 26 | 25 | 51 | 0.58 [0.04, 1.12] | 0.87 [0.31, 1.42] | -0.64 [-1.18, -0.10] | -0.45 [-0.99, 0.08] | Past trip event | Ghost-writer (P) | yes | Oral |
| Vrij, Leal, Deeb, et al. (2020, Study 1) | 41 | 39 | 80 | not examined | 0.71 [0.27, 1.16] | not examined | not examined | Past trip event | None (A) | yes | Oral |
| Vrij, Leal, Deeb, et al. (2020, Study 2, whole sample – no manipulation at Phase 1) | 101 | 107 | 208 | 0.65 [0.37, 0.93] | 0.53 [0.26, 0.81] | not examined | not examined | Past trip event | None (A) | yes | Oral |
| Vrij, Leal, Fisher, Mann, Debb, et al. (2019, whole sample) | 97 | 104 | 201 | 0.11 [-0.16, 0.38] | 0.13 [-0.14, 0.40] | -0.09 [-0.36, 0.18] | -0.14 [-0.42, 0.14] | Past trip event | None (A) | yes | Oral |
| Vrij, Leal, Fisher, Mann, Jo, et al. (2019, whole sample – No manipulation at Phase 1) | 102 | 103 | 205 | 0.23 [-0.05, 0.51] | 0.68 [0.40, 0.95] | -0.34 [-0.62, -0.07] | -0.33 [-0.60, -0.06] | Past trip event | None (A) | yes | Oral |
| Vrij, Leal, Jupe, et al. (2018) | 27 | 26 | 53 | 0.19 [-0.34, 0.72] | 0.44 [-0.10, 0.97] | -0.14 [-0.67, 0.39] | -0.93 [-1.47, -0.38] | Past trip event | None (A) | yes | Oral |
| Vrij et al. (2017, subgroup 1 – no manipulation) | 50 | 50 | 100 | 0.59 [0.19, 0.98] | 0.64 [0.24, 1.04] | -0.23 [-0.62, 0.16] | -0.49 [-0.88, -0.09] | Past trip event | None (A) | yes | Oral |
| Vrij et al. (2017, subgroup 2 – Model Statement) | 49 | 50 | 99 | 0.49 [0.09, 0.88] | 0.80 [0.39, 1.21] | -0.74 [-1.15, -0.34] | -1.13 [-1.54, -0.72] | Past trip event | MS (P) | yes | Oral |
| Vrij, Leal, Mann et al. (2018, whole sample – unexpected and expected questions combined) | 102 | 102 | 204 | 0.09 [-0.18, 0.36] | 0.55 [0.27, 0.83] | -0.59 [-0.87, -0.31] | -0.62 [-0.89, -0.34] | Past trip event | None (A) | yes | Oral |
| Vrij, Leal, Mann, et al. (2020, subgroup 1 – Arab sample) | 38 | 38 | 76 | 0.72 [0.27, 1.18] | 0.36 [-0.09, 0.80] | not examined | not examined | Spy mission | None (A) | yes | Oral |
| Vrij, Mann, et al. (2021) | 47 | 47 | 94 | 0.70 [0.29, 1.11] | 0.81 [0.40, 1.23] | not examined | not examined | Spy mission | None (A) | yes | Oral |
| Vrij, Mann, et al. (2020, whole sample – no manipulation at phase 1) | 60 | 62 | 122 | 1.04 [0.67, 1.41] | 0.77 [0.40, 1.13] | not examined | -0.22 [-0.57, 0.13] | Spy mission | None (A) | yes | Oral |

*Note.* [1](P) indicates that manipulation was coded as "present", (A) indicates that manipulation was coded as absent.

## Appendix C2

Study Characteristics – Second Recall

| | | | | Second Recall *d* [95% CI] | | | | |
|---|---|---|---|---|---|---|---|---|
| Study | Truth tellers *n* | Lie tellers *n* | *N* | New total details | New complications | New common knowledge details | New self-handicapping strategies | Manipulation |
| Deeb et al. (2020, subgroups 2, and 3 - MS at T1, MS at T2, and MS at T1 and T2) | 61 | 62 | 123 | 0.12 [-0.23, 0.48] | 0.57 [0.21, 0.92] | -0.34 [-0.69, 0.01] | -0.27 [-0.62, 0.08] | MS (P) |
| Deeb et al. (2020, subgroups 1 and 4 – MS at T1 and control) | 59 | 61 | 120 | -0.04 [-0.40, 0.31] | 0.54 [0.18, 0.89] | -0.18 [-0.53, 0.18] | -0.22 [-0.57, 0.14] | None (A) |
| Deeb et al. (2021, subgroups 2 and 3 - Sketching at T2 and Sketching at T1 and T2) | 61 | 60 | 121 | 0.52 [0.16, 0.88] | 0.70 [0.34, 1.05] | -0.35 [-0.70, 0.00] | -0.18 [-0.53, 0.18] | Sketching (P) |
| Deeb et al. (2021, subgroups 1 and 4 – Sketching at T1 and control) | 61 | 61 | 122 | 0.34 [-0.01, 0.69] | 0.49 [0.13, 0.84] | -0.54 [-0.89, -0.18] | -0.37 [-0.72, -0.02] | None (A) |
| Leal et al. (2020) | 44 | 41 | 85 | 0.26 [-0.16, 0.68] | 0.76 [0.32, 1.19] | did not occur frequently enough | did not occur frequently enough | MS (P) |
| Vrij et al. (2020, Study 2, subgroups 1 – no manipulation) | 35 | 35 | 70 | 0.45 [-0.02, 0.91] | 0.31 [-0.15, 0.76] | not examined | not examined | None (A) |
| Vrij, Leal, Deeb, et al. (2020, Study 2, subgroups 2 – IP) | 35 | 38 | 73 | 0.30 [-0.15, 0.74] | 0.30 [-0.15, 0.74] | not examined | not examined | IP (P) |
| Vrij, Leal, Deeb, et al. (2020, Study 2, subgroups 3 – IP+MS) | 31 | 34 | 65 | 0.13 [-0.35, 0.60] | 0.45 [-0.03, 0.93] | not examined | not examined | IP + MS (P) |
| Vrij, Leal, Fisher, et al. (2018, subgroup 1 – no manipulation) | 49 | 51 | 100 | -0.09 [-0.48, 0.30] | 0.28 [-0.11, 0.67] | -0.69 [-1.14, -0.25] | not examined | None (A) |
| Vrij, Leal, Fisher, et al. (2018, subgroup 1 – sketching) | 53 | 51 | 104 | 0.21 [-0.17, 0.59] | 0.44 [0.05, 0.82] | -0.54 [-0.92, -0.15] | not examined | Sketching (P) |
| Vrij, Leal, Fisher, Mann, Debb, et al. (2019, whole sample) | 97 | 104 | 201 | 0.41 [0.13, 0.69] | 0.48 [0.20, 0.76] | -0.46 [-0.74, -0.18] | -0.37 [-0.65, -0.09] | MS (P) |
| Vrij, Leal, Fisher, Mann, Jo, et al. (2019, subgroup 1 – no manipulation) | 34 | 35 | 69 | 0.52 [0.06, 0.99] | 0.81 [0.34, 1.29] | -0.61 [-1.08, -0.14] | 0.00 [-0.46, 0.46] | None (A) |
| Vrij, Leal, Fisher, Mann, Jo, et al. (2019, subgroups 2 and 3 combined – standard sketching and model sketching) | 68 | 68 | 136 | 0.41 [0.07, 0.74] | 0.38 [0.04, 0.71] | -0.61 [-0.95, -0.27] | -0.32 [-0.65, 0.01] | Sketching (P) |
| Vrij, Leal, Jupe et al. (2018) | 27 | 26 | 53 | 0.37 [-0.15, 0.90] | 0.84 [0.29, 1.38] | -0.52 [-1.05, 0.00] | -0.57 [-0.64, -0.50] | MS (P) |
| Vrij, Mann et al. (2021) | 47 | 47 | 94 | 0.04 [-0.36, 0.44] | 0.80 [0.39, 1.22] | not examined | not examined | MS (P) |
| Vrij, Mann, et al. (2020, subgroup 1 – no manipulation) | 30 | 32 | 62 | 0.03 [-0.45, 0.51] | 0.61 [0.12, 1.11] | not examined | -0.27 [-0.75, 0.22] | None (A) |
| Vrij, Mann, et al. (2020, subgroup 2 - sketching) | 30 | 30 | 60 | 0.97 [0.45, 1.48] | 0.18 [-0.31, 0.67] | not examined | -0.27 [-0.75, 0.22] | Sketching (P) |

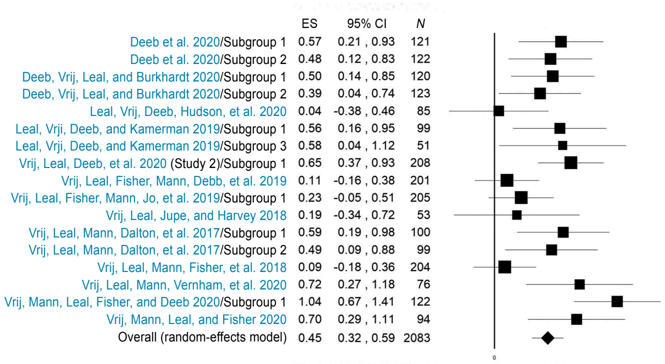
## Appendix C3

Study Characteristics – Total Recall

| | | | | Total Recall *d* [95% CI] | | | | |
|---|---|---|---|---|---|---|---|---|
| Study | Truth tellers *n* | Lie tellers *n* | *N* | Unique total details | Unique complications | Unique common knowledge details | Unique self-handicapping strategies | Manipulation |
| Deeb et al. (2020, subgroups 1, 2, and 3 - MS at T1, MS at T2, and MS at T1 and T2) | 91 | 93 | 184 | 0.28 [-0.01, 0.57] | 0.52 [0.23, 0.81] | -0.38 [-0.67, -0.09] | -0.41 [-0.70, -0.12] | MS (P) |
| Deeb et al. (2020, subgroup 4 – control) | 29 | 30 | 59 | 0.78 [0.27, 1.29] | 0.85 [0.33, 1.37] | -0.54 [-1.05, -0.04] | -0.45 [-0.95, 0.04] | None (A) |
| Deeb et al. (2021, subgroups 1, 2 and 3 - Sketching at T1, Sketching at T2 and Sketching at T1 and T2) | 91 | 90 | 181 | 0.49 [0.20, 0.78] | 0.67 [0.37, 0.96] | -0.56 [-0.85, -0.27] | -0.45 [-0.74, -0.16] | Sketching (P) |
| Deeb et al. (2021, subgroups 4 - control) | 31 | 31 | 62 | 0.54 [0.05, 1.04] | 0.62 [0.12, 1.10] | -0.45 [-0.94, 0.03] | 0.00 [0.48, -0.48] | None (A) |
| Leal et al. (2020) | 44 | 41 | 85 | 0.23 [-0.19, 0.64] | 0.80 [0.37, 1.24] | did not occur frequently enough | did not occur frequently enough | MS (P) |
| Vrij, Leal, Deeb, et al. (2020, Study 2, subgroups 1 – no manipulation) | 35 | 35 | 70 | 0.62 [0.16, 1.09] | 0.80 [0.33, 1.28] | not examined | not examined | None (A) |
| Vrij, Leal, Deeb, et al. (2020, Study 2, subgroups 2 – IP) | 35 | 38 | 73 | 0.50 [0.05, 0.96] | 0.38 [-0.08, 0.83] | not examined | not examined | IP (P) |
| Vrij, Leal, Deeb, et al. (2020, Study 2, subgroups 3 – IP+MS) | 31 | 34 | 65 | 0.67 [0.19, 1.16] | -0.10 [-0.57, 0.38] | not examined | not examined | IP + MS (P) |
| Vrij, Leal, Fisher, Mann, Debb, et al. (2019, whole sample) | 97 | 104 | 201 | 0.32 [0.04, 0.60] | 0.42 [0.15, 0.69] | -0.28 [-0.55, -0.01] | -0.33 [-0.66, 0.00] | MS (P) |
| Vrij, Leal, Fisher, Mann, Jo, et al. (2019, subgroup 1 – no manipulation) | 34 | 35 | 69 | 0.52 [0.06, 0.99] | 0.99 [0.50, 1.47] | -0.15 [-0.61, 0.31] | -0.12 [-0.58, 0.34] | None (A) |
| Vrij, Leal, Fisher, Mann, Jo, et al. (2019, subgroups 2 and 3 combined – standard sketching and model sketching) | 68 | 68 | 136 | 0.26 [-0.08, 0.59] | 0.60 [0.26, 0.94] | -0.67 [-1.01, -0.33] | -0.48 [-0.81, -0.14] | Sketching (P) |
| Vrij, Leal, Jupe et al. (2018) | 27 | 26 | 53 | 0.31 [-0.22, 0.83] | 0.69 [0.16, 1.22] | -0.38 [-0.91, 0.14] | -0.81 [-1.35, -0.27] | MS (P) |
| Vrij, Mann, et al. (2021) | 47 | 47 | 94 | 0.39 [-0.01, 0.78] | 1.04 [0.62, 1.46] | not examined | not examined | Several (P) |
| Vrij, Mann, et al. (2020, subgroup 1 – no manipulation) | 30 | 32 | 62 | 0.68 [0.18, 1.18] | 0.75 [0.25, 1.25] | not examined | -0.27 [-0.75, 0.22] | None (A) |
| Vrij, Mann et al. (2020, subgroup 2 - sketching) | 30 | 30 | 60 | 1.04 [0.51, 1.56] | 0.52 [0.03, 1.02] | not examined | -0.27 [-0.75, 0.22] | Sketching (P) |

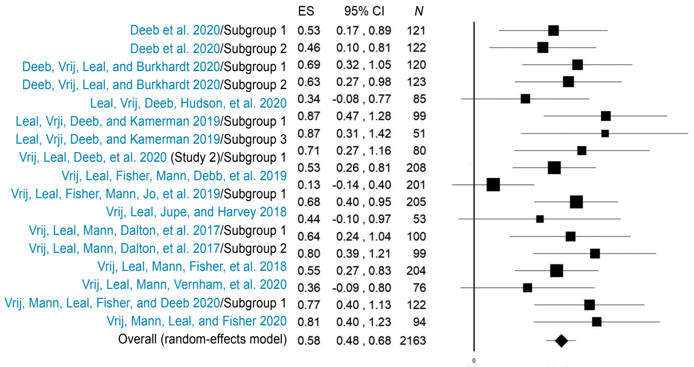
## Appendix D

### Initial Recall Data - Forest Plots



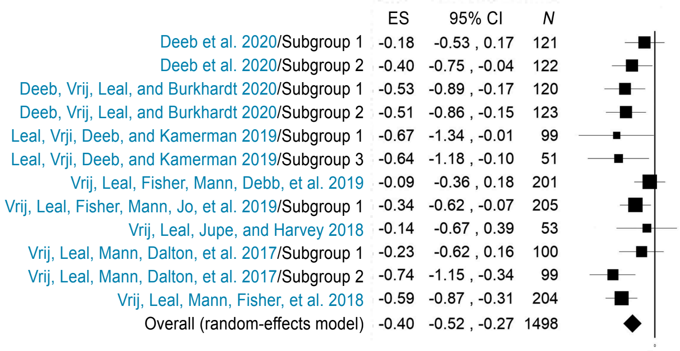
| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | 0.57 | 0.21 , 0.93 | 121 |
| Deeb et al. 2020/Subgroup 2 | 0.48 | 0.12 , 0.83 | 122 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | 0.50 | 0.14 , 0.85 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.39 | 0.04 , 0.74 | 123 |
| Leal, Vrij, Deeb, Hudson, et al. 2020 | 0.04 | -0.38 , 0.46 | 85 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 1 | 0.56 | 0.16 , 0.95 | 99 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 3 | 0.58 | 0.04 , 1.12 | 51 |
| Vrij, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 1 | 0.65 | 0.37 , 0.93 | 208 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | 0.11 | -0.16 , 0.38 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.23 | -0.05 , 0.51 | 205 |
| Vrij, Leal, Jupe, and Harvey 2018 | 0.19 | -0.34 , 0.72 | 53 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 1 | 0.59 | 0.19 , 0.98 | 100 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 2 | 0.49 | 0.09 , 0.88 | 99 |
| Vrij, Leal, Mann, Fisher, et al. 2018 | 0.09 | -0.18 , 0.36 | 204 |
| Vrij, Leal, Mann, Vernham, et al. 2020 | 0.72 | 0.27 , 1.18 | 76 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | 1.04 | 0.67 , 1.41 | 122 |
| Vrij, Mann, Leal, and Fisher 2020 | 0.70 | 0.29 , 1.11 | 94 |
| Overall (random-effects model) | 0.45 | 0.32 , 0.59 | 2083 |

Total details



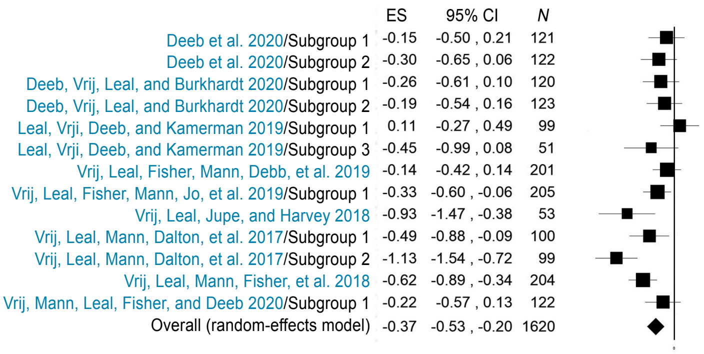
| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | 0.53 | 0.17 , 0.89 | 121 |
| Deeb et al. 2020/Subgroup 2 | 0.46 | 0.10 , 0.81 | 122 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | 0.69 | 0.32 , 1.05 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.63 | 0.27 , 0.98 | 123 |
| Leal, Vrij, Deeb, Hudson, et al. 2020 | 0.34 | -0.08 , 0.77 | 85 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 1 | 0.87 | 0.47 , 1.28 | 99 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 3 | 0.87 | 0.31 , 1.42 | 51 |
| Vrij, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 1 | 0.71 | 0.27 , 1.16 | 80 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | 0.53 | 0.26 , 0.81 | 208 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.13 | -0.14 , 0.40 | 201 |
| Vrij, Leal, Jupe, and Harvey 2018 | 0.68 | 0.40 , 0.95 | 205 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 1 | 0.44 | -0.10 , 0.97 | 53 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 2 | 0.64 | 0.24 , 1.04 | 100 |
| Vrij, Leal, Mann, Fisher, et al. 2018 | 0.80 | 0.39 , 1.21 | 99 |
| Vrij, Leal, Mann, Vernham, et al. 2020 | 0.55 | 0.27 , 0.83 | 204 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | 0.36 | -0.09 , 0.80 | 76 |
| Vrij, Mann, Leal, and Fisher 2020 | 0.77 | 0.40 , 1.13 | 122 |
| | 0.81 | 0.40 , 1.23 | 94 |
| Overall (random-effects model) | 0.58 | 0.48 , 0.68 | 2163 |

Complications



| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | -0.18 | -0.53 , 0.17 | 121 |
| Deeb et al. 2020/Subgroup 2 | -0.40 | -0.75 , -0.04 | 122 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | -0.53 | -0.89 , -0.17 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | -0.51 | -0.86 , -0.15 | 123 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 1 | -0.67 | -1.34 , -0.01 | 99 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 3 | -0.64 | -1.18 , -0.10 | 51 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | -0.09 | -0.36 , 0.18 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | -0.34 | -0.62 , -0.07 | 205 |
| Vrij, Leal, Jupe, and Harvey 2018 | -0.14 | -0.67 , 0.39 | 53 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 1 | -0.23 | -0.62 , 0.16 | 100 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 2 | -0.74 | -1.15 , -0.34 | 99 |
| Vrij, Leal, Mann, Fisher, et al. 2018 | -0.59 | -0.87 , -0.31 | 204 |
| Overall (random-effects model) | -0.40 | -0.52 , -0.27 | 1498 |

Common knowledge details

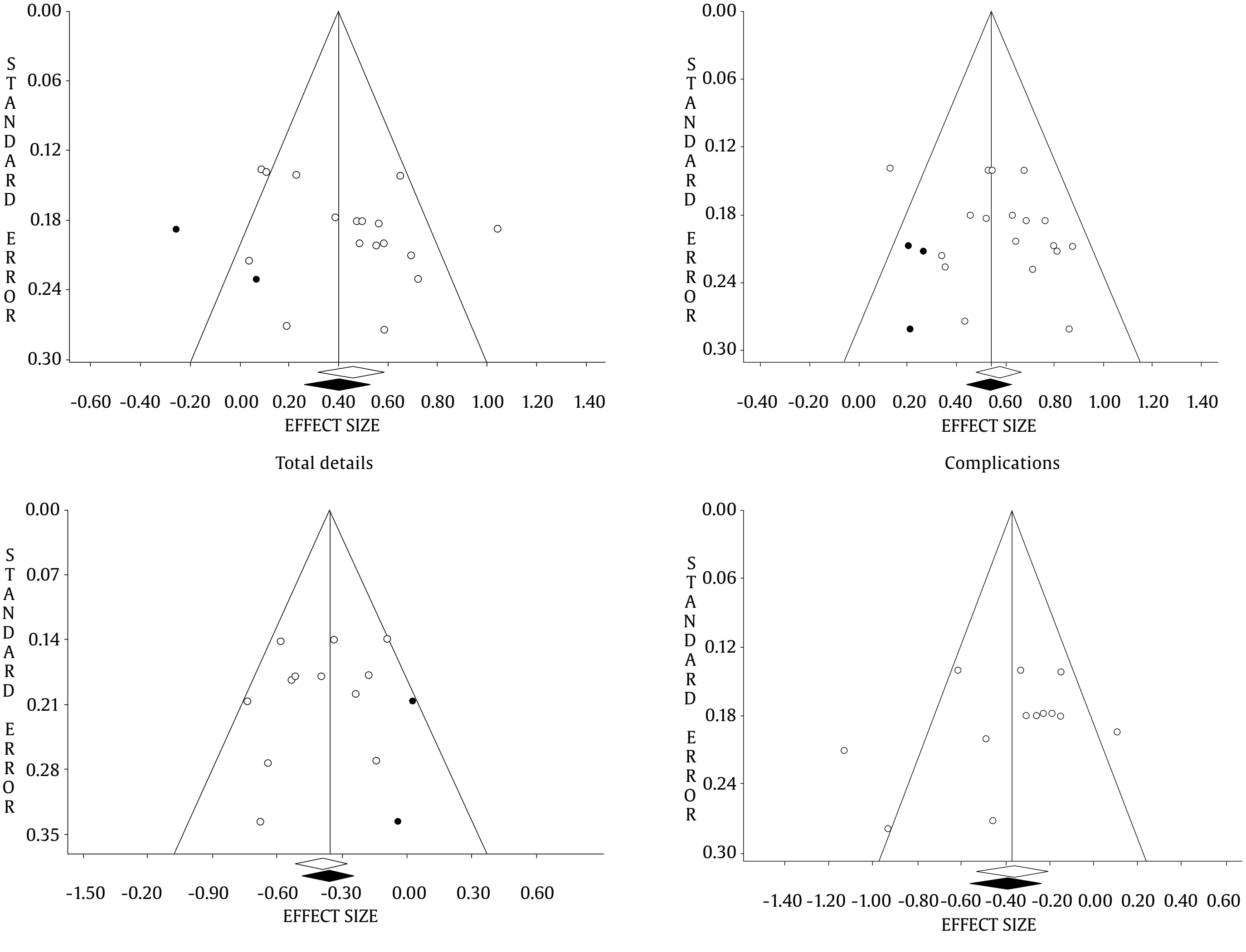


| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | -0.15 | -0.50 , 0.21 | 121 |
| Deeb et al. 2020/Subgroup 2 | -0.30 | -0.65 , 0.06 | 122 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | -0.26 | -0.61 , 0.10 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | -0.19 | -0.54 , 0.16 | 123 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 1 | 0.11 | -0.27 , 0.49 | 99 |
| Leal, Vrji, Deeb, and Kamerman 2019/Subgroup 3 | -0.45 | -0.99 , 0.08 | 51 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | -0.14 | -0.42 , 0.14 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | -0.33 | -0.60 , -0.06 | 205 |
| Vrij, Leal, Jupe, and Harvey 2018 | -0.93 | -1.47 , -0.38 | 53 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 1 | -0.49 | -0.88 , -0.09 | 100 |
| Vrij, Leal, Mann, Dalton, et al. 2017/Subgroup 2 | -1.13 | -1.54 , -0.72 | 99 |
| Vrij, Leal, Mann, Fisher, et al. 2018 | -0.62 | -0.89 , -0.34 | 204 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | -0.22 | -0.57 , 0.13 | 122 |
| Overall (random-effects model) | -0.37 | -0.53 , -0.20 | 1620 |

Self-handicapping strategies

## Appendix E

Funnel Plots for Initial Recall Data



Total details

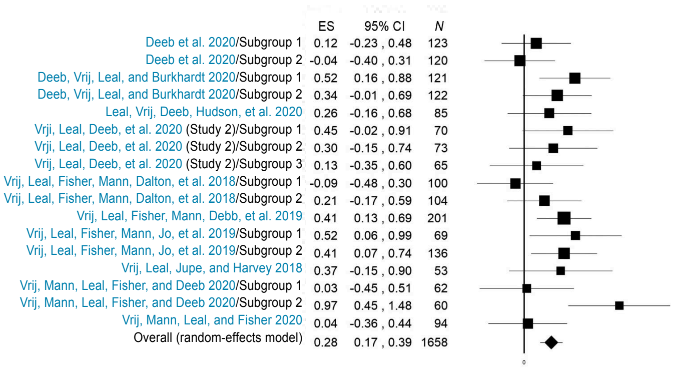
Complications

Common knowledge details

Self-handicapping strategies

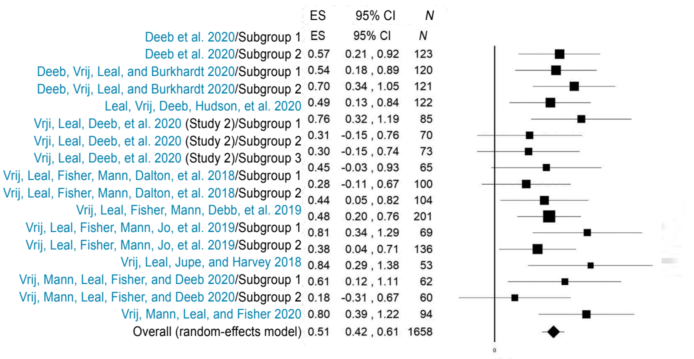
## Appendix F

### Second Recall Data - Forest Plots



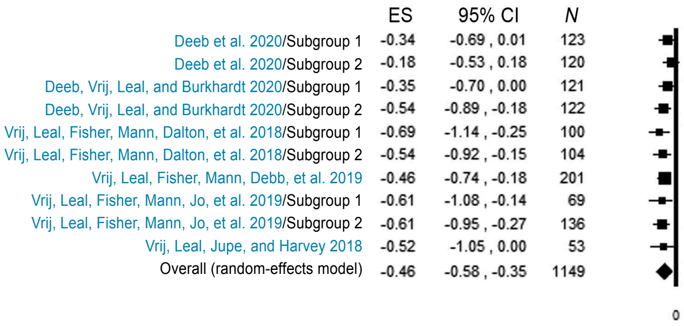
| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | 0.12 | -0.23 , 0.48 | 123 |
| Deeb et al. 2020/Subgroup 2 | -0.04 | -0.40 , 0.31 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | 0.52 | 0.16 , 0.88 | 121 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.34 | -0.01 , 0.69 | 122 |
| Leal, Vrij, Deeb, Hudson, et al. 2020 | 0.26 | -0.16 , 0.68 | 85 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 1 | 0.45 | -0.02 , 0.91 | 70 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 2 | 0.30 | -0.15 , 0.74 | 73 |
| Vrij, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 3 | 0.13 | -0.35 , 0.60 | 65 |
| Vrij, Leal, Fisher, Mann, Dalton, et al. 2018/Subgroup 1 | -0.09 | -0.48 , 0.30 | 100 |
| Vrij, Leal, Fisher, Mann, Dalton, et al. 2018/Subgroup 2 | 0.21 | -0.17 , 0.59 | 104 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | 0.41 | 0.13 , 0.69 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.52 | 0.06 , 0.99 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | 0.41 | 0.07 , 0.74 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | 0.37 | -0.15 , 0.90 | 53 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | 0.03 | -0.45 , 0.51 | 62 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 2 | 0.97 | 0.45 , 1.48 | 60 |
| Vrij, Mann, Leal, and Fisher 2020 | 0.04 | -0.36 , 0.44 | 94 |
| Overall (random-effects model) | 0.28 | 0.17 , 0.39 | 1658 |

New total details



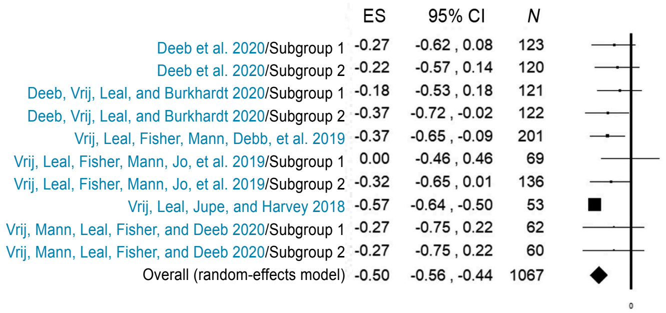
| | ES | 95% CI | *N* |
|---|---|---|---|
| | ES | 95% CI | *N* |
| Deeb et al. 2020/Subgroup 1 | 0.57 | 0.21 , 0.92 | 123 |
| Deeb et al. 2020/Subgroup 2 | 0.54 | 0.18 , 0.89 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | 0.70 | 0.34 , 1.05 | 121 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.49 | 0.13 , 0.84 | 122 |
| Leal, Vrij, Deeb, Hudson, et al. 2020 | 0.76 | 0.32 , 1.19 | 85 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 1 | 0.31 | -0.15 , 0.76 | 70 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 2 | 0.30 | -0.15 , 0.74 | 73 |
| Vrij, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 3 | 0.45 | -0.03 , 0.93 | 65 |
| Vrij, Leal, Fisher, Mann, Dalton, et al. 2018/Subgroup 1 | 0.28 | -0.11 , 0.67 | 100 |
| Vrij, Leal, Fisher, Mann, Dalton, et al. 2018/Subgroup 2 | 0.44 | 0.05 , 0.82 | 104 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | 0.48 | 0.20 , 0.76 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.81 | 0.34 , 1.29 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | 0.38 | 0.04 , 0.71 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | 0.84 | 0.29 , 1.38 | 53 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | 0.61 | 0.12 , 1.11 | 62 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 2 | 0.18 | -0.31 , 0.67 | 60 |
| Vrij, Mann, Leal, and Fisher 2020 | 0.80 | 0.39 , 1.22 | 94 |
| Overall (random-effects model) | 0.51 | 0.42 , 0.61 | 1658 |

New complications



| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | -0.34 | -0.69 , 0.01 | 123 |
| Deeb et al. 2020/Subgroup 2 | -0.18 | -0.53 , 0.18 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | -0.35 | -0.70 , 0.00 | 121 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | -0.54 | -0.89 , -0.18 | 122 |
| Vrij, Leal, Fisher, Mann, Dalton, et al. 2018/Subgroup 1 | -0.69 | -1.14 , -0.25 | 100 |
| Vrij, Leal, Fisher, Mann, Dalton, et al. 2018/Subgroup 2 | -0.54 | -0.92 , -0.15 | 104 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | -0.46 | -0.74 , -0.18 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | -0.61 | -1.08 , -0.14 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | -0.61 | -0.95 , -0.27 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | -0.52 | -1.05 , 0.00 | 53 |
| Overall (random-effects model) | -0.46 | -0.58 , -0.35 | 1149 |

New common knowledge details

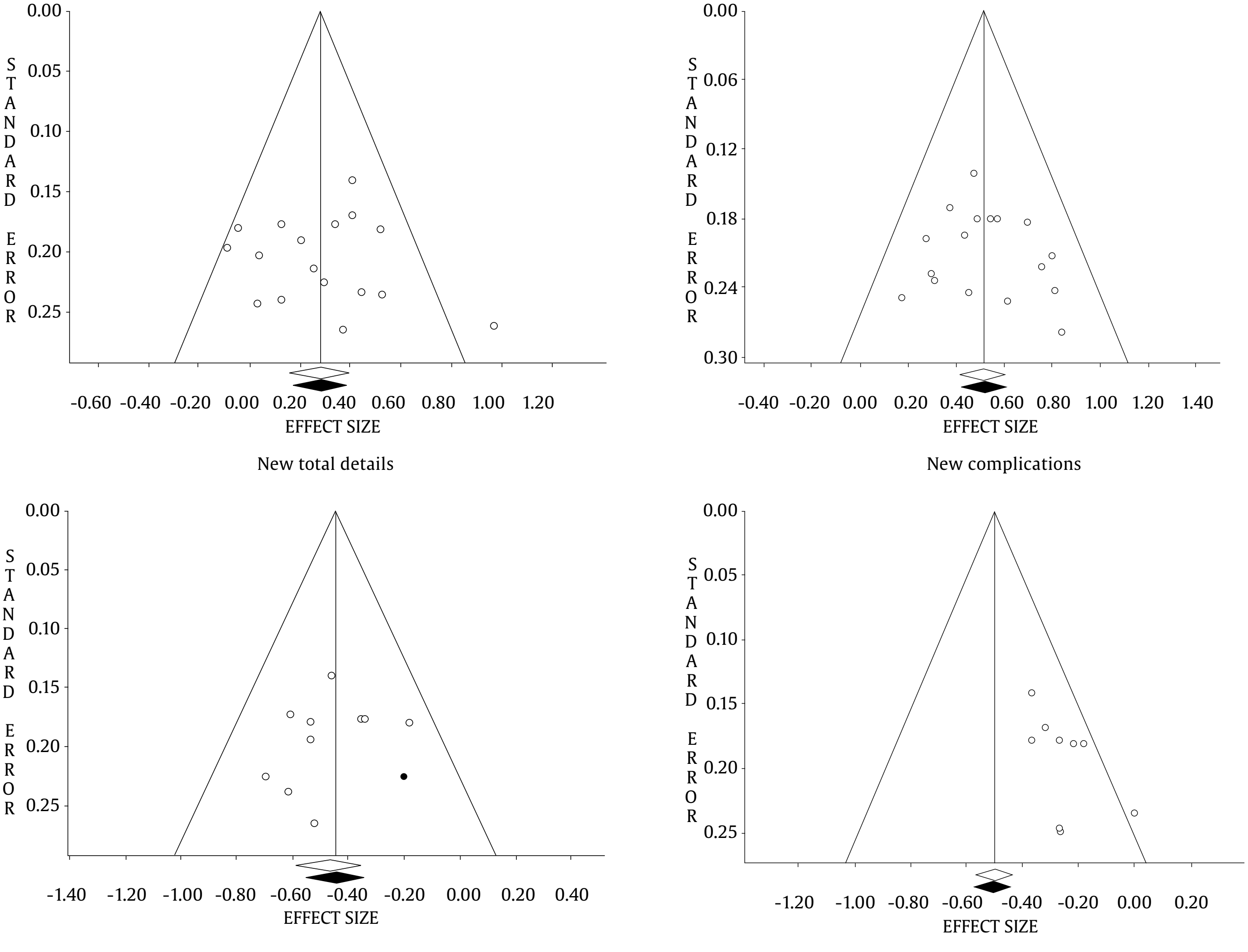


| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | -0.27 | -0.62 , 0.08 | 123 |
| Deeb et al. 2020/Subgroup 2 | -0.22 | -0.57 , 0.14 | 120 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | -0.18 | -0.53 , 0.18 | 121 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | -0.37 | -0.72 , -0.02 | 122 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | -0.37 | -0.65 , -0.09 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.00 | -0.46 , 0.46 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | -0.32 | -0.65 , 0.01 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | -0.57 | -0.64 , -0.50 | 53 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | -0.27 | -0.75 , 0.22 | 62 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 2 | -0.27 | -0.75 , 0.22 | 60 |
| Overall (random-effects model) | -0.50 | -0.56 , -0.44 | 1067 |

New self-handicapping strategies

## Appendix G

Funnel Plots for Second Recall Data



New total details

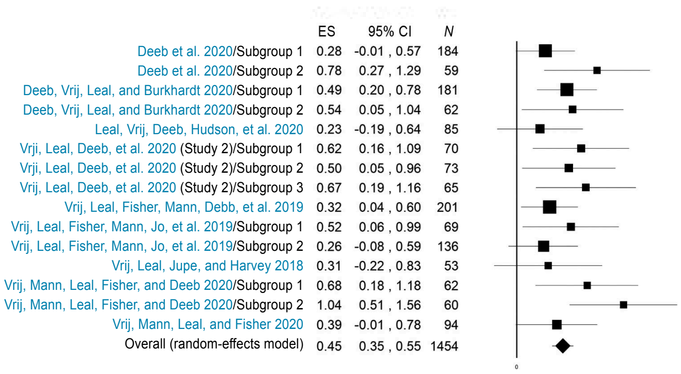
New complications

New common knowledge details

New self-handicapping strategies

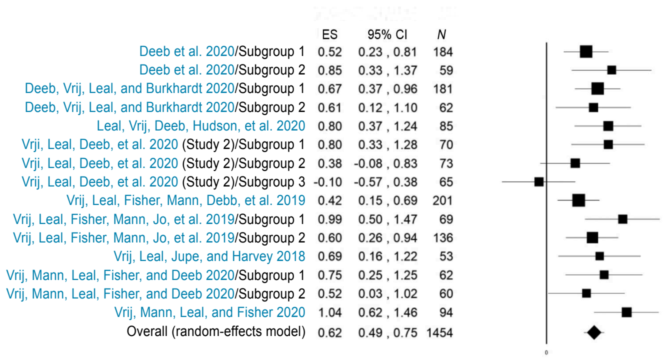
## Appendix H

### Total Recall Data - Forest Plots



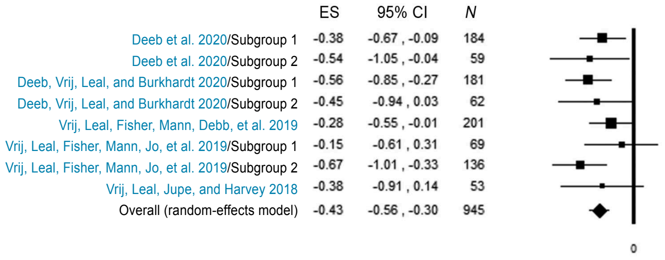
| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | 0.28 | -0.01 , 0.57 | 184 |
| Deeb et al. 2020/Subgroup 2 | 0.78 | 0.27 , 1.29 | 59 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | 0.49 | 0.20 , 0.78 | 181 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.54 | 0.05 , 1.04 | 62 |
| Leal, Vrij, Deeb, Hudson, et al. 2020 | 0.23 | -0.19 , 0.64 | 85 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 1 | 0.62 | 0.16 , 1.09 | 70 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 2 | 0.50 | 0.05 , 0.96 | 73 |
| Vrij, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 3 | 0.67 | 0.19 , 1.16 | 65 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | 0.32 | 0.04 , 0.60 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.52 | 0.06 , 0.99 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | 0.26 | -0.08 , 0.59 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | 0.31 | -0.22 , 0.83 | 53 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | 0.68 | 0.18 , 1.18 | 62 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 2 | 1.04 | 0.51 , 1.56 | 60 |
| Vrij, Mann, Leal, and Fisher 2020 | 0.39 | -0.01 , 0.78 | 94 |
| Overall (random-effects model) | 0.45 | 0.35 , 0.55 | 1454 |

Unique total details



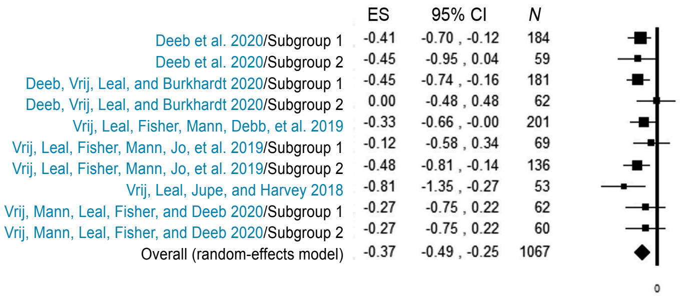
| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | 0.52 | 0.23 , 0.81 | 184 |
| Deeb et al. 2020/Subgroup 2 | 0.85 | 0.33 , 1.37 | 59 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | 0.67 | 0.37 , 0.96 | 181 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.61 | 0.12 , 1.10 | 62 |
| Leal, Vrij, Deeb, Hudson, et al. 2020 | 0.80 | 0.37 , 1.24 | 85 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 1 | 0.80 | 0.33 , 1.28 | 70 |
| Vrji, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 2 | 0.38 | -0.08 , 0.83 | 73 |
| Vrij, Leal, Deeb, et al. 2020 (Study 2)/Subgroup 3 | -0.10 | -0.57 , 0.38 | 65 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | 0.42 | 0.15 , 0.69 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | 0.99 | 0.50 , 1.47 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | 0.60 | 0.26 , 0.94 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | 0.69 | 0.16 , 1.22 | 53 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | 0.75 | 0.25 , 1.25 | 62 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 2 | 0.52 | 0.03 , 1.02 | 60 |
| Vrij, Mann, Leal, and Fisher 2020 | 1.04 | 0.62 , 1.46 | 94 |
| Overall (random-effects model) | 0.62 | 0.49 , 0.75 | 1454 |

Unique complications



| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | -0.38 | -0.67 , -0.09 | 184 |
| Deeb et al. 2020/Subgroup 2 | -0.54 | -1.05 , -0.04 | 59 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | -0.56 | -0.85 , -0.27 | 181 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | -0.45 | -0.94 , 0.03 | 62 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | -0.28 | -0.55 , -0.01 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | -0.15 | -0.61 , 0.31 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | -0.67 | -1.01 , -0.33 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | -0.38 | -0.91 , 0.14 | 53 |
| Overall (random-effects model) | -0.43 | -0.56 , -0.30 | 945 |

Unique common knowledge details

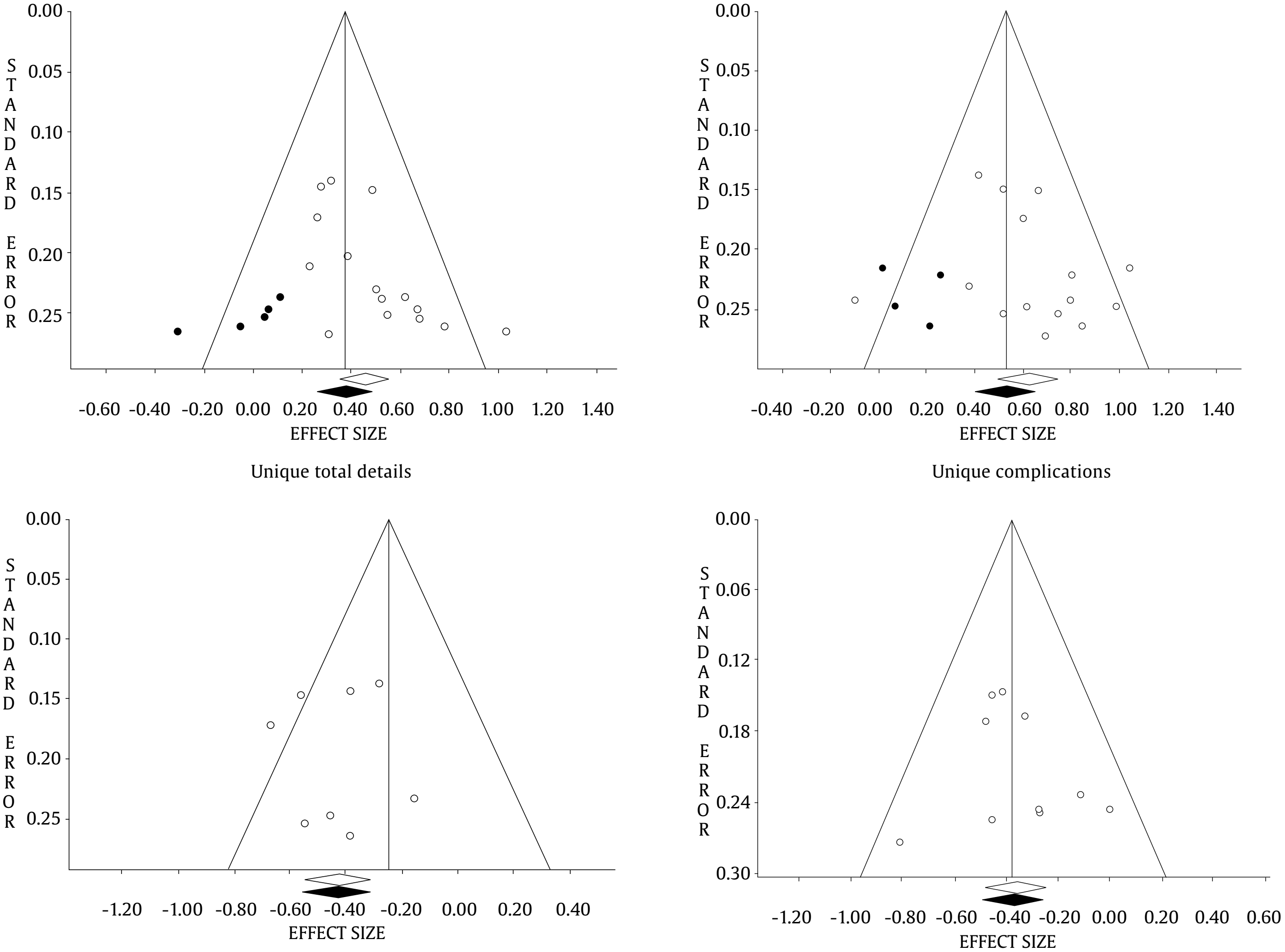


| | ES | 95% CI | *N* |
|---|---|---|---|
| Deeb et al. 2020/Subgroup 1 | -0.41 | -0.70 , -0.12 | 184 |
| Deeb et al. 2020/Subgroup 2 | -0.45 | -0.95 , 0.04 | 59 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 1 | -0.45 | -0.74 , -0.16 | 181 |
| Deeb, Vrij, Leal, and Burkhardt 2020/Subgroup 2 | 0.00 | -0.48 , 0.48 | 62 |
| Vrij, Leal, Fisher, Mann, Debb, et al. 2019 | -0.33 | -0.66 , -0.00 | 201 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 1 | -0.12 | -0.58 , 0.34 | 69 |
| Vrij, Leal, Fisher, Mann, Jo, et al. 2019/Subgroup 2 | -0.48 | -0.81 , -0.14 | 136 |
| Vrij, Leal, Jupe, and Harvey 2018 | -0.81 | -1.35 , -0.27 | 53 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 1 | -0.27 | -0.75 , 0.22 | 62 |
| Vrij, Mann, Leal, Fisher, and Deeb 2020/Subgroup 2 | -0.27 | -0.75 , 0.22 | 60 |
| Overall (random-effects model) | -0.37 | -0.49 , -0.25 | 1067 |

Unique self-handicapping strategies

## Appendix I

Funnel Plots for Total Recall Data



Unique total details

Unique complications

Unique common knowledge details

Unique self-handicapping strategies

## Appendix J

## Bayesian Analyses

There is some debate about whether the frequentist and the Bayesian approaches should be employed together. On the one hand, it has been suggested that the two frameworks might be epistemologically incompatible (Mayo, 1996, 2018). Indeed, whilst the frequentist approach makes use of conditional distributions of the data given the hypothesis, the Bayesian makes use of probability for the data as well as for the hypotheses. On the other hand, it is believed that it might be fruitful to employ both, and there are indeed examples where this has been done (Hong et al., 2013; Wu et al., 2020). Indeed, a comparison of the two frameworks can shed light on the consistency of the results when applying different methods. The merits here are that it is possible to explore two different answers to a same question, each with its own peculiarities.

We carried out a Bayesian meta-analysis for the following reasons. First, with the Bayesian approach there is no need to select either a fixed-effect or a random-effects model as it accounts for model uncertainty. Hence, the Bayesian approach can be applied when there is no certainty about heterogeneity (in contrast to a random-effects model which presumes that it is non-zero), which can be particularly the case when there is no solid background to assume the presence (vs. absence) of heterogeneity.

Second, it permits to explore the plausibility of our a priori choices concerning the application of either the fixed effect (for common knowledge details and self-handicapping strategies relative to second and total recalls) or random-effects (all remaining analyses) standard meta-analyses as outlined in the main text of this article after taking into account the observed data.

Third, in addition to the significance testing of the global effect size obtained via the standard meta-analytic approach, a Bayesian meta-analysis returns a Bayes factor, which provides the amount of evidence in support of the alternative hypothesis (the presence of an effect) against the null hypothesis (the absence of an effect). We interpreted Bayes factors cut-offs as outlined by Jeffreys (1961): values between 1 and 3 indicate weak evidence for $H1$, values between 3 and 10 indicate substantial evidence for $H1$, values between 10 and 20 indicate strong evidence for $H1$, and values beyond 20 indicate very strong evidence for $H1$. Vice versa, values between 1/3 and 1 indicate weak evidence for $H0$, scores between 1/3 and 1/10 indicate substantial evidence for $H0$ etc.

Fourth, a Bayesian approach provides a posterior distribution of the global effect size, whose utility is twofold: i) the probability of each value of the global effect size can be calculated and is shown graphically and ii) future research on the topic can use the resulting posterior distribution of both the effect size and of $\tau$ as an informative prior distribution. Therefore, any new experiment can build on previous evidence in a cumulative manner.

Last, with the Bayesian approach it is possible to obtain the probability of the presence of an effect (in our case, a difference between truth tellers and lie tellers in the examined variables), rather than a dichotomic answer "significant/not significant".

We conducted a Bayesian model averaging meta-analysis (Gronau et al., 2020) and used default priors as recommended by Gronau et al. (2017): A Cauchy prior with scale $1/\sqrt{2}$ centred at zero for the effect size $\mu$, and an inverse-gamma (1, 0.15) prior for between studies standard deviation $\tau$. With this procedure, four Bayesian meta-analysis models are considered: (a) a fixed-effect null-hypothesis, (b) a fixed-effect alternative hypothesis, (c) a random-effects null hypothesis, and (d) a random-effects alternative hypothesis. The four models above are obtained by fixing at zero both $\mu$ and $\tau$ (model a), only $\tau$ (model b), only $\mu$ (model c), or neither (model d). Then, an inclusion Bayes factor for the presence of an effect is obtained by contrasting the two models that predict an effect ($\mu \neq 0$, models b and d, $H1$) at the numerator of the formula to the two models that predict a null effect ($\mu = 0$, models a and c, $H0$) at the denominator,

$$BF_{10\ \text{effect}} = \frac{p\ (\text{model b}|\text{data}) + p\ (\text{model d}|\text{data})}{p\ (\text{model a}|\text{data}) + p\ (\text{model c}|\text{data})} \Bigg/ \frac{p\ (\text{model b}) + p\ (\text{model d})}{p\ (\text{model a}) + p\ (\text{model c})}$$

with the left-hand side of the formula relating to posterior inclusion odds and the right-hand side of the formula relating to the prior inclusion odds. In essence, a Bayesian model averaged meta-analysis: i) can quantify the evidence in support for the presence of an effect while accounting for uncertainty relative to choosing a fixed effect or a random-effects meta-analysis, ii) can provide evidence for the presence/absence of between-studies heterogeneity, and iii) returns posterior odds for each of the four models once the observer data are taken into account. The higher a posterior odd, the more plausible a specific model.

## Results

We had 12 meta-analyses, the result of three interviews (first recall, second recall and total recall), and four dependent variables (total details, complications, common knowledge details, and self-handicapping strategies). When conducting the standard meta-analyses, we opted (a priori) for a fixed-effect model when analysing common knowledge details and self-handicapping strategies concerning "second recall" and "total recall" (four analyses), and a random-effects model for the remaining eight analyses. However, especially when research on a specific topic is still in its development, it is not always possible to rule out uncertainty concerning the correct model selection (fixed-effect vs. random-effects).

A series of Bayesian model averaging meta-analyses was conducted to explore such uncertainty and to evaluate our choices. The results showed that we chose, a priori, the model with higher posterior probabilities in five out of 12 cases. The seven cases where the analyses showed that the model we chose was less probable than its alternative (Table J1, superscript 1) concerned: i) complications in initial and second and total recall, ii) total details in second and total recall, iii) common knowledge details in initial recall, and iv) self-

**Table J1.** Posterior Model Probabilities for the Outcome Variables

| | Posterior models probabilities | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total details | | | Complications | | | Common knowledge details | | | Self-handicapping strategies | | |
| Model | IR | SR[1] | TR[1] | IR[1] | SR[a] | TR[1] | IR[1] | SR | TR | IR | SR[1] | TR |
| Fixed H0 | 1.842e-21 | 9.723e-7 | 1.738e-15 | 5.322e-37 | 2.908e-23 | 1.808e-28 | 1.764e-11 | 8.241e-13 | 3.376e-9 | 2.419e-11 | 9.163e-59 | 2.422e-7 |
| Fixed H1 | 0.047 | 0.613 | 0.739 | 0.558 | 0.802 | 0.560 | 0.557 | 0.776 | 0.714 | 0.066 | 0.016 | 0.765 |
| Random H0 | 4.688e -5 | 0.001 | 1.058e -6 | 4.925e -9 | 1.766e -8 | 8.817e -7 | 2.648e -4 | 2.369e -5 | 8.725e-4 | 0.013 | 0.017 | 8.222e-4 |
| Random H1 | 0.953 | 0.386 | 0.261 | 0.442 | 0.198 | 0.440 | 0.443 | 0.224 | 0.285 | 0.921 | 0.967 | 0.234 |

*Note.* [1]Indicates that the posterior probability of the model we selected (fixed-effect vs. random-effects) was lower than that of the model we did not select. For example, for the meta-analysis concerning new total details (second recall) we ran a random-effects standard meta-analysis, but the posterior model probabilities indicate that a fixed-effect model was the most plausible (≈61% probability) after having taken into account the observed data than the random-effects model (≈38% probability). IR = Initial recall, SR = Second recall, TR = Total recall.

**Table J2.** Effect Size Estimates, Standard Deviations, 95% Credible Intervals, and Bayes Factors

| | Averaged model | | | | |
|---|---|---|---|---|---|
| | Effect size posterior μ | Effect size posterior *SD* | Effect size 95% CI | $BF_{10\mu}$ | $BF_{10\tau}$ |
| Initial recall | | | | | |
| Total details | .45 | .07 | [.32, .58] | 21329.45 | 20.40 |
| Complications | .57 | .05 | [.48, .67] | 2.03 e +8 | 0.79 |
| Common knowledge details | -.39 | .06 | [-.50, -.27] | 3774.82 | 0.79 |
| Self-handicapping strategies | -.36 | .08 | [-.53, -.20] | 74.08 | 14.19 |
| Second recall | | | | | |
| New total details | .28 | .05 | [.17, .38] | 771.06 | 0.63 |
| New complications | .51 | .05 | [.41, .60] | 5.66 e +7 | 0.25 |
| New common knowledge details | -.46 | .06 | [-.58, -.34] | 42217.14 | 0.29 |
| New self-handicapping strategies | -.34 | .08 | [-.49, -.17] | 58.33 | 61.26 |
| Total recall | | | | | |
| Unique details | .45 | .05 | [.33, .55] | 944736.14 | 0.35 |
| Unique complications | .61 | .06 | [.49, .72] | 1.13 e +6 | 0.79 |
| Unique common knowledge details | -.42 | .07 | [-.56, -.29] | 1145.19 | 0.40 |
| Unique self-handicapping strategies | -.37 | .06 | [-.49, -.24] | 1214.91 | 0.30 |

handicapping strategies in second recall. For i), ii), and iii) the fixed-effect model obtained higher posterior probability than the random-effects model we chose. However, Table J2 shows that the Bayes factor for the presence of heterogeneity ranged from $BF_{10}\tau = 0.25$ to $BF_{10}\tau = 0.79$, indicating only weak to moderate evidence for the absence of heterogeneity. Due to this uncertainty, a model averaging analysis is particularly appropriate. For iv) the random-effects model was more plausible (≈96.7 probability) after observing the data than the fixed-effect model (≈1.6% probability) we chose. In this case, there was strong evidence for the presence of heterogeneity ($BF_{10}\tau = 61.26$).

Concerning the presence of an effect, Table J2 shows very strong to extreme support for *H*1 for the four outcome variables (total details, complications, common knowledge details, and self-handicapping strategies), indicating that there is evidence for the hypothesis that such variables can discriminate truth tellers from lie tellers (*H*1). Put differently, the data were more likely under the alternative hypothesis than under the null hypothesis to a large extent, supporting Hypotheses 1 to 3. Table J2 also shows that complications obtained the highest Bayes factors.

## Discussion

On the one hand, the results of this Bayesian meta-analysis reflect the conclusions obtained via the frequentist framework (Table J2). First, complications, common knowledge details, and self-handicapping strategies can discriminate truth tellers from lie tellers, as the presence of an effect was always more likely than the lack of it. Second, via the Bayesian meta-analysis, complications emerged as a stronger veracity indicator than the other two variables, as the Bayes factor analyses show (Table J2).

On the other hand, results concerning heterogeneity were less clear-cut. The Bayesian framework found the posterior probabilities of the examined models generally supporting the lack of between-study variance. Indeed, after observing the data, model b ($\tau$ fixed at 0) received higher posterior probability than model d ($\tau \neq 0$) in most cases- except for self-handicapping strategies in initial and second recall. Yet, as Table J2 shows, the Bayes factors of $\tau$ were mostly inconclusive, indicating that at this stage it is difficult to draw any conclusion concerning the degree of between-study variance. In this regard, the Bayesian analysis reported here further strengthens the idea that future studies are needed to explore more in depth why the analysed studies seemed to show low heterogeneity, and whether this is actually the case. For example, futures studies could compare the results of different meta-analysis models obtained by varying the prior distribution of $\tau$.

Notwithstanding this, it is possible that the results obtained here are due to method invariance across experiments, as all studies come from Vrij's lab and imply a consistent coding system and a shared research design.

In conclusion, it is essential to understand why the selected studies showed no heterogeneity, as well as to explore the external validity of the obtained results, by answering questions as: Do complications, common knowledge details and self-handicapping strategies still work in scenarios that are different from those included in this meta-analysis? Is the lack of heterogeneity due to method invariance and to the fact that all studies come from the same lab or to other reasons? And, lastly, do the conclusions obtained here apply to real-life material? We hope that future studies will shed light on the potential of the complications approach.

## References

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2020). *A primer on Bayesian model-averaged meta-analysis*. https://doi.org/10.31234/osf.io/97qup

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. *Comprehensive Results in Social Psychology, 2*(1), 123-138. https://doi.org/10.1080/23743603.2017.1326760

Hong, H., Carlin, B. P., Shamliyan, T. A., Wyman, J. F., Ramakrishnan, R., Sainfort, F., & Kane, R. L. (2013). Comparing Bayesian and frequentist approaches for multiple outcome mixed treatment comparisons. *Medical Decision Making, 33*(5), 702-714. https://doi.org/10.1177/0272989X13481110

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.

Wu, Q., Xu, Y., Bao, Y., Alvarez, J., & Gonzales, M. L. (2020). Tricyclic antidepressant use and risk of fractures: A meta-analysis of cohort studies through the use of both frequentist and Bayesian approaches. *Journal of Clinical Medicine, 9*(8), 2584. https://doi.org/10.3390/jcm9082584