Psychon Bull Rev (2017) 24:1718–1741
DOI 10.3758/s13423-017-1227-6

THEORETICAL REVIEW

The revelation effect: A meta-analytic test of hypotheses

André Aßfalg[1,2] · **Daniel M. Bernstein**[1] · **William Hockley**[3]

Published online: 1 February 2017
© Psychonomic Society, Inc. 2017

Abstract Judgments can depend on the activity directly preceding them. An example is the revelation effect whereby participants are more likely to claim that a stimulus is familiar after a preceding task, such as solving an anagram, than without a preceding task. We test conflicting predictions of four revelation-effect hypotheses in a meta-analysis of 26 years of revelation-effect research. The hypotheses' predictions refer to three subject areas: (1) the basis of judgments that are subject to the revelation effect (recollection vs. familiarity vs. fluency), (2) the degree of similarity between the task and test item, and (3) the difficulty of the preceding task. We use a hierarchical multivariate meta-analysis to account for dependent effect sizes and variance in experimental procedures. We test the revelation-effect hypotheses with a model selection procedure, where each model corresponds to a prediction of a revelation-effect hypothesis. We further quantify the amount of evidence for one model compared to another with Bayes factors. The results of this analysis suggest that none of the extant revelation-effect hypotheses can fully account for the data. The general vagueness of revelation-effect hypotheses and the scarcity of data were the major limiting factors in our analyses, emphasizing the need for formalized theories and further research into the puzzling revelation effect.

Keywords Revelation effect · Meta-analysis · Hypotheses

Electronic supplementary material The online version of this article (doi:10.3758/s13423-017-1227-6) contains supplementary material, which is available to authorized users.

✉ André Aßfalg
andre.assfalg@psychologie.uni-freiburg.de

[1] Department of Psychology, Kwantlen Polytechnic University, Surrey, BC, Canada

[2] Present address: Department of Psychology, Albert-Ludwigs University Freiburg, Engelbergerstr. 41, 79106 Freiburg, Germany

[3] Wilfrid Laurier University, Waterloo, ON, Canada

Judgments are highly malleable. For example, the activity directly preceding a judgment can bias its outcome. Watkins and Peynircioglu (1990) first described the tendency for people to consider something familiar after engaging in a brief, but cognitively demanding, task. In the original experiments, which established the standard for subsequent research, participants first studied a list of words and later received a recognition test including old, studied words and new, unstudied words. Critically, half the words in the recognition test immediately followed a preceding task (*task condition*). The remaining words in the recognition test appeared without preceding task (*no-task condition*). These preceding tasks required participants to identify a word that was revealed letter by letter (e.g., 1. *su_s_ _ne*, 2. *su_sh_ne*, 3. *sunsh_ne*, 4. *sunshine*), to solve an anagram (e.g., *esnuihsn–sunshine*), or to mentally rotate upside-down words or letters. After any of these tasks, participants were more likely to claim that they saw the test item in the study list compared to items in the no-task condition. Watkins and Peynircioglu coined the term *revelation effect* to denote the increase in "old" judgments for items in the task condition compared to items in the no-task condition. Since then, numerous other studies have replicated the revelation effect (see Aßfalg, 2017, for an overview).

A notable characteristic of the revelation effect is its generality. For example, the revelation effect occurs with various types of preceding tasks. Some experiments require participants to solve anagrams or to identify a word that is revealed letter by letter (Watkins & Peynircioglu, 1990). In other experiments, a word or picture appears rotated (e.g., upside

down). Again, the participants' task is to identify the stimulus (Peynircioglu & Tekcan, 1993). In a variation of this task, individual letters of a word appear rotated by varying degrees (Watkins & Peynircioglu, 1990). However, not all tasks include the identification of a word or a picture. Other preceding tasks include adding numbers, transposing single digits in a number sequence following a specific rule, typing a word in reverse order, or generating a synonym for a word (Luo, 1993; Niewiadomski & Hockley, 2001; Verde & Rotello, 2003; Westerman & Greene, 1998).

In other respects, the revelation effect is more limited. Watkins and Peynircioglu (1990) identified several judgment types that are not affected by the preceding task. In their experiments, the preceding task did not affect typicality judgments for category exemplars, lexical decisions, or word-frequency estimates. This led the authors to conclude that the revelation effect is restricted to episodic-memory judgments. However, other studies suggest that the revelation effect extends beyond episodic memory. For example, after solving an anagram of the word *leopard*, participants are more convinced of the veracity of the statement "*The leopard is the fastest land animal*" than without solving the anagram (Bernstein, Whittlesea, & Loftus, 2002). Further, solving an anagram of a brand name increases the preference for that brand name compared to brand names without a preceding task (Kronlund & Bernstein, 2006).

Past revelation-effect research mostly focused on recognition memory for which the signal detection model has been a useful and ubiquitous analysis method (Macmillan & Creelman, 2005). The signal detection model describes recognition memory as a function of the participant's ability to discriminate between familiar and unfamiliar stimuli and the participant's criterion—a point on the familiarity dimension above which a stimulus is called familiar. Past research suggests that participants do not alter their criterion within a test list, even when stimuli differ in their familiarity strength (Hirshman, 1995). However, more recently, some studies found evidence for criterion shifts within test lists. For example, in Rhodes and Jacoby's (2007) study, participants provided recognition judgments for words that appeared on either of two locations on the screen. While words in one location were mostly studied words, the words in the other location were mostly unstudied. Critically, participants used a more liberal criterion (more "old" responses) for the location with mostly studied words compared to the location with mostly unstudied words. Similarly, in Singer and Wixted's (2006) study, items studied immediately before testing received a more conservative criterion than items studied 2 days before testing. However, this within-list criterion shift disappeared when Singer and Wixted compared immediate testing with a 40-minute delay—still a strong experimental manipulation.

These difficulties to find within-list criterion shifts are at odds with the revelation effect which had been discovered over a decade before the aforementioned research was published (Watkins & Peynircioglu, 1990). In revelation-effect studies, task and no-task trials appear randomly intermixed in a recognition test. In this context, the criterion has been consistently shown to be more liberal in the task condition compared to the no-task condition (e.g., Verde & Rotello, 2003)—a within-list criterion shift. Compared to other manipulations that require 2-day retention intervals to demonstrate within-list criterion shifts (Singer & Wixted, 2006), typical preceding tasks in revelation-effect studies, such as solving anagrams, last only a few seconds (Aßfalg & Nadarevic, 2015). Consequently, the revelation-effect is arguably the most efficient extant method to study within-list criterion shifts.

The revelation effect also has implications for formal models of recognition memory. Any explanation of recognition memory must account for the revelation effect or fall short of its task. Historically, a similar situation occurred with the mirror effect—the simultaneous increase of hits and decrease of false alarms in one condition compared to another (Glanzer & Adams, 1985, 1990). For example, low-frequency words tend to receive more hits and fewer false alarms than high-frequency words do (e.g., Hockley, 1994). The identifying feature of global-matching models is that the familiarity of a stimulus depends on its similarity with all memory content (Clark & Gronlund, 1996). Without further assumptions, this general matching mechanism cannot account for the simultaneous increase in hits and decrease in false alarms in the mirror effect (Hintzman, 1990). Together with failed attempts to find the list-strength effect, the mirror effect led to the replacement of global-matching models by more sophisticated models of memory (Criss & Howard, 2015). More recent models, such as the retrieving-effectively-from-memory model or the context-noise model, can account for the mirror effect and a wide range of other phenomena (Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997). However, we are unaware of any publication that accounts for the revelation effect with current memory models. Thus, similar to the mirror effect, the revelation effect requires further refinement of recognition-memory models.

In a larger context, the revelation effect suggests a link between various judgment phenomena that have been—for the most part—treated as independent. These phenomena include the revelation effect, the truth effect, and the mere-exposure effect. The truth effect is an increased tendency to claim that statements are true after encountering those statements previously, compared to novel statements (Hasher, Goldstein, & Toppino, 1977). The mere-exposure effect is the tendency to prefer previously encountered stimuli over novel stimuli, even when this previous encounter is not remembered (Kunst-Wilson & Zajonc, 1980; Zajonc, 1968). Some authors have explained all three phenomena—the revelation effect, the truth effect, and the mere-exposure effect—

with the concept of processing fluency (Bernstein et al., 2002; Unkelbach, 2007; Whittlesea & Price, 2001; Whittlesea & Williams, 2001)—that is, the ease and speed of processing information. Thus, identifying the cause of the revelation effect may well serve to unite parts of the literature on cognitive illusions that have been largely treated as independent (Pohl, 2017).

However, finding the cause of the revelation effect has several challenges. Although various authors have proposed explanations of the revelation effect, our impression is that the field has not done enough to conclusively rule out some of these hypotheses. Another issue is that all extant revelation-effect hypotheses—in varying degrees—suffer from vague assumptions that make falsification attempts difficult. Finally, recent research has renewed concerns about the replicability of psychological studies (Open Science Collaboration, 2015), calling into question the validity of the empirical basis by which hypotheses are evaluated.

The purpose of this work is to reevaluate revelation-effect hypotheses based on the research of the past 26 years. To counter the methodological issues associated with underpowered studies and an overreliance on *p*-values, we chose a meta-analytic approach for this study. Our general approach was to identify contradicting predictions of revelation-effect hypotheses. We then translated these predictions into statistical models and compared these models against each other to identify the hypothesis that best accounts for the past 26 years of revelation-effect research.

Using meta-analysis to test the predictions of hypotheses has an important advantage over primary studies. Where primary research often culminates in the conclusion that an effect does or does not exist, meta-analysis takes even small effects into account that did not pass an arbitrary significance criterion. Our suspicion was that some of the conflicting results in the revelation-effect literature arose because of low statistical power. When some primary studies yield a significant test statistic and other studies do not, the same data, through the lens of meta-analysis, can provide overwhelming evidence for or against the existence of an effect (Borenstein, Hedges, Higgins, & Rothstein, 2009). This problem is exacerbated by the overuse and misinterpretation of the *p*-value coupled with low statistical power in primary studies (Rossi, 1990; Sedlmeier & Gigerenzer, 1989). Meta-analysis reduces the impact of these issues and provides effect-size estimates using formalized procedures.

In the Revelation-effect hypotheses section, we describe the hypotheses reported in the literature. For each hypothesis, we describe the core assumptions and predictions for the meta-analysis. In the Method section, we describe the selection of studies for the meta-analysis, the treatment of missing values, and the computation of effect sizes. We then introduce the meta-regression procedure we used to perform the meta-analysis and describe in detail how we coded the moderator variables. In the General analyses section, we provide an estimate of the overall effect size and identify potential confounding variables. In the Hypothesis tests section, we report the results of several model comparisons that directly relate to the hypotheses' predictions. We conclude with a discussion of the results and suggestions for future research.

Revelation-effect hypotheses

Several initially promising explanations of the revelation effect have been ruled out. For example, Luo (1993) tested whether *extended exposure* to the test item generated by some preceding tasks is responsible for the revelation effect. However, exposure time to the test item did not affect recognition judgments. This hypothesis is also incompatible with the observation that task items that are semantically unrelated to the test item produce a revelation effect (Westerman & Greene, 1996). According to a related hypothesis, the delay of a judgment by the preceding task is responsible for the revelation effect. However, if the preceding task is replaced by a 10 s period of inactivity, without exposure to the task item, the revelation effect does not occur (Westerman & Greene, 1998).

Westerman and Greene (1996) further disproved the hypothesis that the revelation effect is an *affirmation bias*—that is, a mere tendency to agree to statements such as "Did you study this word?" In one experiment, participants first studied two word lists. In the following recognition test, participants received words from both study lists and had to decide whether the words had appeared in list one. The preceding task increased the proportion of "no" responses—that is, participants were more likely to judge that revealed words had appeared in list two. Note that words from list two likely appeared more familiar than list-one words simply because participants had studied list-two words more recently. This result supports the assumption that the preceding task induces a feeling of familiarity and not a mere affirmation bias.

Several other hypotheses received various amounts of support over the years and remain largely untested. Our aim in this study was to identify areas in which these hypotheses predict opposite patterns. We identified three such areas, which we briefly summarize here. We describe these areas in more detail in the Coding of moderator variables section. The first area concerns the basis of the judgment, including familiarity, recollection, and fluency. We are aware that fluency has also been linked to judgments of familiarity (e.g., Whittlesea & Williams, 1998, 2000). Thus, to disentangle these two concepts, we define fluency-based judgments as judgments that have been linked to fluency, but not familiarity such as judgments of truth, preference, or typicality (Alter & Oppenheimer, 2009). The second area concerns the degree of similarity between the task item and the test item (same

vs. related vs. different). We consider task and test items that are the *same*, for example, if participants solve an anagram in the preceding task and then judge the familiarity of the anagram solution. Further, task and test items are *related* if they represent the same type of stimulus (e.g., both are words). Finally, task and test items are *different* if they represent different types of stimuli (e.g., one is a number, the other is a word). The third area we considered is the difficulty of the preceding task (hard vs. easy) which includes, for example, long versus short anagrams. Table 1 lists the hypotheses along with an outline of their predictions. Each hypothesis predicts a distinct data pattern regarding the moderators in the present analysis.

In the following sections, we present each hypothesis and its predictions regarding the meta-analysis. Our aim was to present the hypotheses and their predictions as formulated by their authors. In cases where we had to introduce auxiliary assumptions, we mark these assumptions as our own and describe our rationale for choosing them.

Global-matching hypothesis

According to the *global-matching hypothesis*, "the performance of a revelation task, such as solving a word fragment or carrying out a memory-span task, leads to activation of some information in memory. This activation then persists for a limited period of time. When participants see the test stimulus and try to perform a recognition decision, the activation resulting from the revelation task is still persisting. This activation then makes some contribution to the overall activation level on which the recognition decision is based." (Westerman & Greene, 1998, pp. 384–385).

Thus, following the rationale of global-matching models of memory (e.g., Clark & Gronlund, 1996), the sum of activation, caused partly by the test item and partly by the preceding task, determines the overall degree of familiarity with the test item. The activation from the preceding task produces more "old" responses in the task condition compared to the no-task condition. Figure 1 illustrates the core assumptions of the global-matching hypothesis.

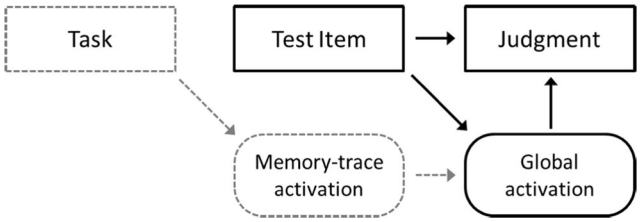

Fig. 1 Outline of the global-matching hypothesis: The preceding task activates study-list memory traces; this activation contributes to a global measure of familiarity. *Dashed lines* indicate elements only present in the task condition, *rectangles* indicate manifest variables, and *rounded rectangles* indicate latent variables

Basis of judgments (familiarity vs. recollection vs. fluency)

The global-matching hypothesis postulates a larger revelation effect for familiarity-based compared to recollection-based judgments (see Table 1). Proponents of dual-process models have likened the global similarity measure in global-matching models to familiarity rather than recollection. Westerman (2000), for example, notes, regarding the global-matching hypothesis: "If the revelation effect is caused by an increase in the familiarity of the test items that follow the revelation task, the magnitude of the revelation effect should depend on the extent to which participants use familiarity as the basis of their memory judgments." (p. 168).

Westerman and Greene (1998) also pointed out that LeCompte found that the revelation effect increases familiarity as measured with the process-dissociation (Jacoby, 1991) and remember-know procedures (Gardiner, 1988; Tulving, 1985). However, familiarity may, to a lesser degree, also contribute to judgments commonly labeled "recollection based" (e.g., associative recognition). Thus, the revelation effect could also occur in recollection-based judgments but should be smaller than the effect in familiarity-based judgments. Alternatively, one could argue that the revelation effect should be absent in recollection-based judgments. However, this assumption requires the unlikely precondition that judgments considered as recollection based are process-pure measures of recollection without familiarity. Finally, Westerman and Greene (1998) pointed out that the revelation effect seems to be limited to episodic memory. This implies that the revelation effect should be absent in fluency-based judgments that are

Table 1 Predictions in the meta-analysis about the size of the revelation effect (γ) as a function of potential moderators (rows) and revelation-effect hypotheses (columns)

	Global matching	Decrement to familiarity	Criterion flux	Discrepancy attribution
Basis of judgments	$\gamma_{familiarity} > \gamma_{recollection}, \gamma_{fluency} = 0$	$\gamma_{familiarity} > \gamma_{recollection}, \gamma_{fluency} = 0$	$\gamma_{familiarity} > \gamma_{recollection}, \gamma_{fluency} = 0$	$\gamma_{fluency} = \gamma_{familiarity} > \gamma_{recollection}$
Similarity	$\gamma_{same} = \gamma_{related} > \gamma_{different}$	$\gamma_{same} = \gamma_{related} = \gamma_{different}$	$\gamma_{same} = \gamma_{related} = \gamma_{different}$	$\gamma_{same} = \gamma_{related} = \gamma_{different}$
Difficulty	$\gamma_{hard} = \gamma_{easy}$	$\gamma_{hard} > \gamma_{easy}$	$\gamma_{hard} = \gamma_{easy}$	$\gamma_{hard} > \gamma_{easy}$

not directly addressing familiarity such as truth, preference, and typicality judgments (Alter & Oppenheimer, 2009).

Similarity (same vs. related vs. different) Regarding the moderator "similarity," Westerman and Greene (1998) write that "the magnitude of the revelation effect is determined by the extent to which the revelation task activates information from the study list that would not have been activated by the recognition-test item alone" (p. 385). The authors further maintain that the activation caused during task solving, rather than the presentation of the task item per se, is responsible for the revelation effect. They further remark that "seen from the perspective of global-matching models, it is not surprising that it is irrelevant whether the stimulus used in the revelation task is the same as the one on which the subsequent recognition decision is performed" (p. 385). This suggests that, according to the global-matching hypothesis, the revelation effect should be equally large when the test item and task item are identical compared to just similar (e.g., two unrelated words). However, Westerman and Greene remark that if task and test stimuli are sufficiently dissimilar, the revelation effect should be absent. They offer two potential explanations for this outcome: (1) Participants might be better able to "keep the activation to the revelation stimulus from contaminating the recognition process" (p. 385) because the task and test items are sufficiently different, or (2) stimuli such as words and digits might be stored in separate subsystems of memory, preventing activation to spill over from one system to another. Thus, the global-matching hypothesis predicts equally large revelation effects in the same and similar conditions, but a weaker effect in the different condition.

Difficulty (hard vs. easy) The global-matching hypothesis predicts equally large revelation effects for hard compared to easy preceding tasks. In Westerman and Greene's (1998) study, task difficulty did not moderate the revelation effect. The authors did not mention any conclusions with regard to their hypothesis, which we interpret as the assumption that task difficulty does not moderate the revelation effect. One could speculate that harder tasks likely induce more activation than easy tasks. However, even a hard arithmetic task (e.g., 248 × 372 = ?), for example, should not cause a revelation effect for words as test items, according to the global-matching hypothesis.

Decrement-to-familiarity hypothesis

Relatively early after the discovery of the revelation effect, Hicks and Marsh (1998) reviewed the then available revelation-effect data and observed that there is a positive link between the size of the revelation effect and the difficulty of the preceding task. Hicks and Marsh also observed in their own recognition experiments that if participants have to decide between an item revealed as part of an anagram and an intact item without anagram, they select the intact item as the more familiar. To account for these observations, Hicks and Marsh proposed the decrement-to-familiarity hypothesis.

Similar to the global-matching hypothesis, this hypothesis assumes that the preceding task activates memory traces. However, instead of contributing to a global measure of familiarity, this activation is thought to introduce noise to the decision process. Hicks and Marsh (1998) concluded that "the activation of competitors would reduce the signal-to-noise ratio for the item being judged" (p. 1116). With respect to the observation that participants select intact words more often than revealed words in forced-choice tasks, Hicks and Marsh further note with regard to revealed items, "given that competing representations could be activated during revelation that would not have been active if the item had been presented intact, the smaller signal-to-noise ratio might make them seem unfamiliar" (pp. 1117). Hicks and Marsh cite work by Hirshman (1995) to support their hypothesis. Hirshman showed that response criteria become more conservative if the memorability of test items increases. Hicks and Marsh argue that the opposite could be true in the revelation effect: If task items appear less familiar than no-task items, the response criterion might be more liberal in the task condition compared to the no-task condition. Figure 2 illustrates the core assumptions of the decrement-to-familiarity hypothesis.

Basis of judgments (familiarity vs. recollection vs. fluency) Table 1 summarizes the predictions of the decrement-to-familiarity hypothesis with regard to the moderators in the present meta-analysis. Regarding the moderator "basis of judgments," the decrement-to-familiarity hypothesis postulates that the preceding task affects the perceived familiarity of the test item. Consequently, the hypothesis predicts that the preceding task primarily affects familiarity-based judgments. Again, we do not assume that judgments labeled as recollection based are process-pure measures of recollection. Thus, judgments commonly labeled as recollection based might show a revelation effect as well, although to a lesser degree than familiarity-based judgments. In comparison, the

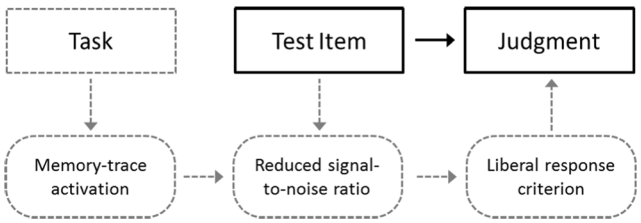

Fig. 2 Outline of the decrement-to-familiarity hypothesis: The preceding task activates memory traces. This activation reduces the test item's signal-to-noise ratio. Participants react to the now-more-difficult judgment with a liberal response criterion. *Dashed lines* indicate elements only present in the task condition, *rectangles* indicate manifest variables, and *rounded rectangles* indicate latent variables

revelation effect should not occur for fluency-based judgments such as judgments of truth, preference, and typicality. Hicks and Marsh wrote about perceptual and conceptual fluency as possible explanations of the revelation effect: "Although either of these alternatives might account for the revelation effect, current data and theory do not favor them" (p. 1107). Consequently, in the decrement-to-familiarity hypothesis, fluency is not considered as a potential cause of the revelation effect.

Similarity (same vs. related vs. different) Regarding the moderator "similarity," the decrement-to-familiarity hypothesis predicts that the revelation effect should be equally large for each similarity condition. Hicks and Marsh (1998) mention Westerman and Greene's (1998) observation that the revelation effect only occurred if task and test items were both words, but not when the task item was a number and the test item was a word. However, Hicks and Marsh do not draw any conclusions with regard to the decrement-to-familiarity hypothesis. We again interpret this as the assumption that similarity should not moderate the revelation effect. This interpretation is supported by the fact that Hicks and Marsh do not claim that only specific stimuli could decrease the signal-to-noise ratio for the recognition judgment.

Difficulty (hard vs. easy) Regarding the moderator "difficulty," the decrement-to-familiarity hypothesis predicts that hard preceding tasks produce a larger revelation effect than easy preceding tasks. Hicks and Marsh (1998) maintain that the revelation effect seems to be larger for hard compared to easy tasks. Further, with regard to letter-by-letter preceding tasks that vary the number of starting letters of a word fragment (e.g., sun_hi_e vs. _u_s_ _ _e) Hicks and Marsh remark that "if more competitors are activated by starting with fewer letters as opposed to more complete fragments, then the revelation effect should have been larger with stimuli that were more initially degraded" (p. 1117). Thus, according to the decrement-to-familiarity hypothesis, hard preceding tasks should induce a larger revelation effect than easy tasks.

Criterion-flux hypothesis

According to the *criterion-flux hypothesis* (Hockley & Niewiadomski, 2001; Niewiadomski & Hockley, 2001), the preceding task disrupts working memory, thereby replacing important information about the test list, such as the proportion of targets. In this situation, according to the hypothesis, participants temporarily rely on a default liberal response criterion. Shortly after, test-list information is reinstated in working memory, that is, the response criterion is in a constant flux between task and no-task trials. Figure 3 illustrates the core assumptions of the criterion-flux-hypothesis.

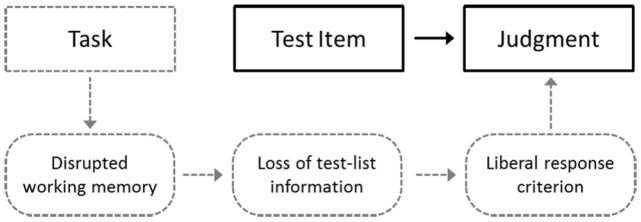

Fig. 3 Outline of the criterion-flux hypothesis: The preceding task disrupts working memory, replacing test-list information such as the proportion of targets. Participants temporarily rely on a default liberal response criterion until test-list information is restored. *Dashed lines* indicate elements only present in the task condition, *rectangles* indicate manifest variables, and *rounded rectangles* indicate latent variables

Basis of judgments (familiarity vs. recollection vs. fluency) Table 1 summarizes the predictions of the criterion-flux hypothesis with regard to the moderators in the meta-analysis. Regarding the moderator "basis of judgments," Niewiadomski and Hockley (2001) proposed the criterion-flux hypothesis to account for the revelation effect in recognition experiments. Hockley and Niewiadomski (2001) note that "[the revelation effect] is limited to episodic memory decisions based largely on familiarity, but not decisions that involve recall or recollection" (p. 1183). Thus, the criterion-flux hypothesis predicts a revelation effect for familiarity-based judgments. Assuming that judgments commonly labeled as recollection-based can sometimes rely on familiarity, a revelation effect for recollection-based judgments is possible under the criterion-flux hypothesis. However, this effect should be smaller than the revelation effect for familiarity-based judgments. Further, according to the criterion-flux hypothesis, the preceding task disrupts working memory, which induces a liberal recognition criterion. Thus, the hypothesis discards fluency as a potential cause of the revelation effect. Consequently, the criterion-flux hypothesis predicts that fluency-based judgments (e.g., truth, preference, or typicality) should not produce a revelation effect.

Similarity (same vs. related vs. different) Regarding the "similarity" moderator, the criterion-flux hypothesis predicts a revelation effect independent of the similarity between task and test items. In a series of experiments, Niewiadomski and Hockley (2001) demonstrated a revelation effect for words as recognition items independent of whether the preceding task was an anagram or an arithmetic task. The only requirement for the revelation effect is that the preceding task disrupts working memory and temporarily induces a more liberal response criterion.

Difficulty (hard vs. easy) Finally, regarding the "difficulty" moderator, the criterion-flux hypothesis predicts roughly equally sized revelation effects for hard and easy preceding tasks. For example, attempts to increase the difficulty of a preceding task by presenting two anagrams compared to just

one should not change the size of the revelation effect. Niewiadomski and Hockley (2001) note that "two problem tasks would not generally result in a larger effect than only one preceding task, because, usually, one task would be sufficient to displace list context from working memory" (p. 1137). Following this interpretation, any task that disrupts working memory will produce a revelation effect. Further attempts to increase the size of the revelation effect will fail because the disruption of working memory is already complete and cannot be increased any further.

Discrepancy-attribution hypothesis

Several authors suggested the *discrepancy-attribution hypothesis* (e.g., Whittlesea & Williams, 1998, 2000) as an explanation of the revelation effect (Bernstein et al., 2002; Leynes, Landau, Walker, & Addante, 2005; Luo, 1993; Whittlesea & Williams, 2001). These authors note that the preceding task is essentially a manipulation of fluency—that is, the ease and speed of processing information. The core assumption of this hypothesis is that participants experience low processing fluency during the preceding task. Conversely, participants process the test item relatively fluently. The discrepancy between the disfluent preceding task and fluent test item produces a feeling of familiarity that participants misattribute to the test item. Indeed, fluency, induced by priming or manipulations of stimulus clarity, can be misattributed to familiarity with a stimulus (Jacoby & Whitehouse, 1989; Whittlesea, Jacoby, & Girard, 1990). Similarly, a feeling of familiarity, induced by the discrepancy between preceding task and test-item fluency, could be the cause of the revelation effect. Figure 4 illustrates the core assumptions of the discrepancy-attribution hypothesis.

Basis of judgments (familiarity vs. recollection vs. fluency) Table 1 summarizes the predictions of the discrepancy-attribution hypothesis with regard to the moderators in the present meta-analysis. Regarding the moderator "basis of judgments," the hypothesis predicts that the revelation effect occurs for fluency-based judgments. This includes the judgments we labeled as fluency-based in the moderator analysis, such as truth, preference, and typicality, but also familiarity-based judgments, such as old–new recognition or recognition-confidence judgments (e.g., Whittlesea & Williams, 2001). Conversely, the judgments commonly labeled as recollection-based, such as associative recognition, rhyme recognition, and plurality judgments, should rely on fluency to a lesser degree and, therefore, elicit a smaller revelation effect than fluency-based or familiarity-based judgments. We again base this assumption on the idea that no judgment is a process-pure representation of its underlying cognitive construct.

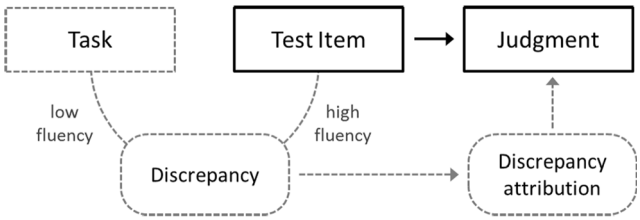

Fig. 4 Outline of the discrepancy-attribution hypothesis: Participants process the test item more fluently than the preceding task. This discrepancy in relative fluency produces a feeling of familiarity that is attributed to the test item. *Dashed lines* indicate elements only present in the task condition, *rectangles* indicate manifest variables, and *rounded rectangles* indicate latent variables

Similarity (same vs. related vs. different) Regarding the moderator "similarity," the discrepancy-attribution hypothesis predicts that the revelation effect is independent of the degree of similarity between the task and test items. The discrepancy-attribution hypothesis includes a broad definition of fluency which is closely related to mental effort. Oppenheimer (2008), for example, defines fluency as "the subjective experience of ease or difficulty with which we are able to process information" (p. 237). In this definition, high mental effort is equivalent to low processing fluency and low mental effort is equivalent to high processing fluency. Following this definition, in the discrepancy-attribution hypothesis, the degree of similarity between the task and test items should not be sufficient to moderate the revelation effect. Instead, we propose that for the revelation effect to occur, it is necessary that the preceding task requires more effort and, therefore, less fluency than the test item, irrespective of the similarity between task and test item.

Difficulty (hard vs. easy) Finally, regarding the moderator "difficulty," the discrepancy-attribution hypothesis predicts that hard preceding tasks cause a larger revelation effect than easy preceding tasks. This prediction directly follows from the definition of processing fluency applied to the discrepancy-attribution hypothesis: Hard tasks cause lower fluency than easy tasks (e.g., Oppenheimer, 2008). Assuming that the directly following judgment elicits relatively high fluency, participants should be more likely to experience a fluency discrepancy after a hard preceding task compared to an easy preceding task. It follows that the revelation effect should be larger in studies with hard preceding tasks than in studies with easy preceding tasks.

Method

Selection of studies

We identified an initial set of revelation-effect publications by searching for the term "*revelation effect*" in the PsychINFO

database and using the Google Scholar search engine.[1] We further searched for revelation-effect publications by checking all studies that cited our initial set of publications. Our inclusion criterion was that each publication contained at least one experiment in which participants performed the same type of judgment (e.g., recognition) across several trials. Additionally, a preceding task had to appear in some of these trials (within-subjects design) or, alternatively, one group of participants provided their judgments after the preceding task whereas another group provided their judgments without a preceding task (between-subjects design). We further restricted our search to publications in English; this resulted in 35 publications.

We dropped two publications from the meta-analysis that used a two-alternative-forced-choice procedure (2-AFC; Hicks & Marsh, 1998; Major & Hockley, 2007). In the 2-AFC procedure, participants first revealed one of the two choices by solving an anagram whereas the other choice appeared intact. Participants then decided which of the two choices they recognized from a previous study phase. Although this procedure met our selection criteria, as described above, the 2-AFC procedure is insensitive to the response criterion—the point on a latent memory-strength dimension above which an item is called "old." However, changes in the response criterion are a defining feature of the revelation effect (Verde & Rotello, 2004). Thus, it is doubtful whether effect sizes based on 2-AFC represent the same aspects of the revelation effect as effect sizes based on other response formats.

Next, we contacted one author of each of the remaining 33 publications (contact information was not available in four cases) and asked for any unpublished revelation-effect experiments and the raw data of the published experiments.[2] The primary reason for requesting raw data was to inform an estimate of the correlation between repeated measures, which we needed to compute effect sizes. The secondary reason was to retrieve as much information about the standard deviations of the dependent variables as possible, because several publications failed to report standard deviations and test statistics for the relevant conditions. In response to our request, we received the raw data pertaining to one unpublished manuscript and 12 published papers. Table 2 lists our final data set including 33 published papers and one unpublished manuscript.

[1] We acknowledge that this approach has limitations. Not all authors that found a revelation effect might have labeled or even considered the effect as such. Other studies may include phenomena that were labeled as revelation effect, but in truth represent other phenomena. Consequently, the challenge for theorists could be that they have to explain empirical findings that share a label rather than a common cognitive foundation. We are grateful to David Huber for raising this issue.

[2] Bornstein et al. (2015) was a late addition to the meta-analysis for which we did not try to obtain the raw data.

Table 2 List of publications included in the meta-analysis (publications for which we had the raw data appear in bold letters)

Publication	Experiments[a]	Effect sizes[b]
Aßfalg & Bernstein (2012)	3	6
Aßfalg & Nadarevic (2015)	7	14
Azimian-Faridani & Wilding (2004)	1	2
Bernstein, Godfrey, Davison, & Loftus (2004)	4	8
Bernstein, Rudd, Erdfelder, Godfrey, & Loftus (2009)	2	4
Bernstein, Whittlesea, & Loftus (2002)	3	5
Bornstein & Neely (2001)	3	12
Bornstein, Robicheaux, & Elliott (2015)	1	6
Bornstein & Wilson (2004)	2	4
Cameron & Hockley (2000)	5	20
Dewhurst & Knott (2010)	9	27
Dougal & Schooler (2007)[c]	3	12
Frigo, Reas, & LeCompte (1999)	4	4
Guttentag & Dunn (2003)	3	6
Hockley & Niewiadomski (2001)[d]	5	16
Hodgson (n.d.)	2	4
Kronlund & Bernstein (2006)	3	10
Landau (2001)	4	8
LeCompte (1995)	2	8
Leynes, Landau, Walker, & Addante (2005)	2	4
Luo (1993)	4	12
Mulligan (2007)	3	8
Mulligan & Lozito (2006)	4	8
Niewiadomski & Hockley (2001)	9	30
Peynircioglu & Tekcan (1993)	8	24
Prull, Light, Collett, & Kennison (1998)	3	6
Thapar & Sniezek (2008)	4	8
Verde & Rotello (2003)	3	12
Verde & Rotello (2004)	3	8
Watkins & Peynircioglu (1990)	11	29
Westerman (2000)	5	10
Westerman & Greene (1996)	7	14
Westerman & Greene (1998)	9	18
Young, Peynircioglu, & Hohman (2009)	2	4
Total	**143**	**371**

[a] Number of experiments refers to the number of corresponding task and no-task conditions (within and between subjects) and does not necessarily correspond to the labeling of experiments in the publications.

[b] Experiments may contribute more than one effect size, for example, one each for "old" and "new" stimuli in a recognition experiment.

[c] Several experiments without control condition or with proportion recalled as dependent variable were excluded.

[d] One experiment without a preceding task was excluded

Imputation of missing values

The revelation-effect data included two types of missing values concerning the dependent variable: correlations

between repeated measures and standard deviations. None of the publications reported the correlations between repeated measures but we were able to take correlations from the 33% of experiments for which we had the raw data. Additionally, standard deviations were only available for 64% of experiments. Regarding correlations, the problem was somewhat mitigated by experiments that manipulated task presence (task vs. no task) between subjects, but these experiments were a small minority (about 4% of all effect sizes).

Effect sizes for a within-subjects manipulation of task presence depend on the correlation between repeated measures. To substitute missing correlations, we performed a fixed-effect meta-analysis of all correlations obtained from the raw data. Because all experiments with more than two effect sizes included multiple correlations, we first computed an aggregated correlation estimate for each experiment. We then replaced the missing values with the point estimate of the population correlation ($\hat{\rho} = .29$) based on the available raw data.

Although most studies reported standard deviations or standard errors, some studies lacked any measure of variability. We decided to estimate standard deviations from the available data and impute this estimate in cases where the standard deviation was missing. Alternatively, we could have computed effect sizes based on the reported test statistics. However, this approach was impractical because some results were presented as $F < 1$. Additionally, not all studies reported separate test statistics for the revelation effect in relevant within-subject conditions such as test-item status (target vs. distractor) in recognition experiments.

To get an estimate of the standard deviations, we had to distinguish between different dependent variables. Whereas some of the judgments were dichotomous (e.g., true vs. false), reported as proportions, others involved confidences scales (e.g., 1 = *sure new* to 6 = *sure old*), reported as mean confidence ratings. We made no further distinction between confidence scales, because most scales used five to eight response options with little variation. We computed estimates of the standard deviation separately for proportions and mean confidence judgments. To this end, we used the fixed-effects model by weighing each observed variance with its sample size, thus giving more weight to large samples. We then divided the aggregated variance by the total sample size across all studies. This approach resulted in an estimate of the standard deviation for proportions of $S_p = 0.20$ and for confidence judgments of $S_c = 0.79$. In the final step, we imputed these values for the missing standard deviations of proportions and confidence judgments, respectively.

Effect-size calculation

Ignoring the authors' labeling of experiments, we defined an experiment as follows: We considered within-subject designs where all participants provided judgments with and without preceding task as an experiment. Similarly, we considered between-subjects designs that included one group of participants providing judgments with preceding task and another group without preceding task as an experiment. Using this definition, the 34 studies contained 143 experiments with 5,034 participants total. As effect size we computed the standardized mean difference between the task and no-task conditions (e.g., Cohen, 1992),

$$d = \frac{M_T - M_{NT}}{S_{pooled}},$$

where M_T is the mean judgment in the task condition, M_{NT} is the mean judgment in the no-task condition and S_{pooled} is their pooled standard deviation. We computed S_{pooled} in the within-subject case as

$$S_{pooled} = \frac{S_{diff}}{\sqrt{2(1-r)}},$$

where S_{diff} is the standard deviation of the difference scores between the task and no-task conditions and r is the correlation between judgments in the task and no-task conditions. We computed S_{diff} as

$$S_{diff} = \sqrt{S_T^2 + S_{NT}^2 - 2rS_TS_{NT}},$$

where S is the standard deviation in the task (T) and no-task (NT) conditions.

Conversely, in the between-subjects case we computed S_{pooled} as

$$S_{pooled} = \sqrt{\frac{(n_T-1)S_T^2 + (n_{NT}-1)S_{NT}^2}{n_T + n_{NT} - 2}},$$

where n is the sample size in the task (T) and no-task (NT) conditions.

We further applied the small-sample correction suggested by Hedges (1981) to the standardized mean difference d by computing

$$g = \left(1 - \frac{3}{4df - 1}\right)d,$$

where df is $n_T + n_{NT} - 2$ in the between-subjects case, df is $n - 1$ in the within-subject case, and n is the sample size (see Borenstein et al., 2009).

Most experiments included several within-subjects conditions, each with a task and no-task condition. Recognition experiments, for example, typically list the revelation effect separately for targets and distractors. On average, each experiment contributed 3.02 effect sizes for a total of 368. For each experiment, we aimed to construct as many separate effect

sizes as suggested by the experimental design. However, in four experiments with particularly complex designs, we pooled mean proportions or mean confidence judgments across some conditions. This approach kept the size and format of data tables tractable and simplified the hypothesis tests. Specifically, in two experiments by Dougal and Schooler (2007), we pooled judgments across anagram difficulty (hard vs. easy) and instead focused on whether participants were able to solve the anagrams (see the Potential confounds section).[3] In another study, Watkins and Peynircioglu (1990, Experiment 1) reported judgments as a function of test-item status (target vs. distractor) and six levels of task difficulty. Instead of coding all 12 conditions, we pooled the three hardest tasks and the three easiest tasks, separately for studied and unstudied test words. The dichotomous comparison of hard compared to easy tasks accords with other experiments in the meta-analysis and our hypothesis tests. Finally, an experiment by Bornstein and Neely (2001, Experiment 3) included four levels of study frequency and three levels of task difficulty. Instead of coding all 12 conditions, we pooled responses across study frequency conditions and kept the task difficulty levels.

Meta-analysis method

The available data place several restrictions on the choice of meta-analysis method. The revelation effect has been tested for diverse judgments such as recognition, preference, or truth (Bernstein et al., 2002; Kronlund & Bernstein, 2006). Given the diversity of revelation-effect studies, we chose the random-effects approach. This approach relies on the premise that there is more than one true effect size underlying the data. Our choice of a meta-analysis method was further influenced by the predominant within-subjects design used in revelation-effect research. Typically, task presence (task vs. no task) is crossed with one or more within-subject factors. In the context of a recognition test, for example, participants experience task and no-task trials for both targets and distractors. The data thus allowed the computation of a revelation-effect size separately for the levels of within-subject variables. However, effect sizes relying on the same participant sample and research group are dependent. To account for this dependency, we used multivariate meta-analysis and estimated covariances between effect sizes based on the available raw data (Raudenbush, Becker, & Kalain, 1988; see the Imputation of missing values section).

Note that we use the terms *effect size*, *experiment*, and *study* to distinguish between hierarchical levels of the data. We describe our computation of effect sizes in the Effect size calculation section. One or more effect sizes are nested within experiments and share the same participant sample. Finally, one or more experiments are nested within studies and originate from different participant samples but the same publication. When describing our meta-analysis model, we use the term *study* to describe the hierarchical unit that contains one or more dependent effect sizes. However, in all our analyses, we compute results based on models that assume dependent effect sizes at the study level or only at the experiment level. For the latter approach, the term *study* can be simply replaced with the term *experiment* in the description of the meta-analysis method (see also the Appendix).

In our model, the vector of observed effect sizes $\mathbf{y}_j$ in study $j = 1, \ldots, M$ follows a multivariate normal distribution given by

$$\mathbf{y}_j \sim N\left(\mathbf{X}_j \mathbf{b}_j, \Sigma_j\right), \tag{1}$$

where $\mathbf{X}_j$ is a design matrix, $\mathbf{b}_j$ is a vector of regression coefficients, and Σ_j is the variance–covariance matrix of $\mathbf{y}_j$. Estimates for the variances in the main diagonal of Σ_j are readily available (Borenstein et al., 2009). We further estimate the covariances in the off-diagonal entries of Σ_j using the approach suggested by Raudenbush et al. (1988). This approach relies on the assumption that correlations between effect sizes are homogenous across studies. We estimated this correlation based on the correlation of repeated measures in the available raw data with $\hat{\rho} = .29$ (see the Imputation of missing values section). However, throughout all analyses, we perform sensitivity analyses and report results for $\hat{\rho} = \{.1, .3, .5\}$ where $\hat{\rho} = .3$ represents the most likely scenario based on the available raw data.

We further assume that the regression coefficients $\mathbf{b}_j$ vary across studies. Let b_{jk} be the regression coefficient of predictor $k = 1,\ldots,L$ in study j, then

$$b_{jk} \sim N(\gamma_k, \tau_k), \tag{2}$$

where the γ_k and τ_k represent the mean regression coefficient and precision parameter of predictor k. For γ_k and τ_k we define the prior distributions

$$\gamma_k \sim N(\mu_0, \tau_0) \tag{3}$$

$$\tau_k \sim \Gamma(a_0, b_0), \tag{4}$$

where Γ denotes the gamma distribution. In our analyses, we selected uninformative priors with $\mu_0 = 0$, $\tau_0 = 0.01$, $a_0 = 2$, and $b_0 = 0.05$, where a_0 and b_0 are the shape and rate parameters of the gamma distribution. Consequently, the prior distribution of γ_k in Eq. 3 expresses a very weak prior belief that the revelation effect does not exist. Similarly, the prior distribution of τ_k in Eq. 4 defines a wide range of precision values as equally plausible.

Our primary aim was to test the predictions of various revelation-effect hypotheses. To that end, we defined specific statistical models for each prediction of the revelation-effect hypotheses. All tested models concern assumptions

[3] We did not include these experiments in the moderator analysis because anagram solution success proved to be a potential confound.

about the combined effect sizes γ_k in Eq. 3. The first assumption we considered is that the effect sizes of two conditions m and n are equal, $\gamma_m = \gamma_n$, which is readily implemented by choosing an appropriate design matrix. Note that this approach also enforces $b_{jm} = b_{jn}$, $\forall j$ which means that effect sizes pertaining to predictors m and n are assumed to originate from the same distribution. The second assumption we tested is that the effect sizes in a condition m are zero, $\gamma_m = 0$, which can be implemented by defining $b_{jm} \sim N(\gamma_m = 0, \tau_m)$. The third and final assumption we tested is that the effect sizes in condition m are larger than the effect sizes in condition n, $\gamma_m > \gamma_n$. This parameter restriction can be written as $\gamma_m = \gamma_n + a$, where $a > 0$. Consequently, the prior of b_{jm} is $N(\gamma_m = \gamma_n + a,\ \tau_m)$. For the auxiliary parameter a, we further defined the prior

$$a \sim N(0, 1)_{a>0}, \tag{5}$$

where Eq. 5 denotes a truncated normal distribution such that $a > 0$.

To test which model best fits the data, a model-selection scheme is required. Our aim was to select the model that provides the best balance between fit to the data and model parsimony. For this goal, we chose a Bayesian approach that samples parameters and models simultaneously (Carlin & Chib, 1995). This approach evaluates the relative evidence of one model compared to that of another model in the form of the Bayes factor. Consider the Bayes factor B_{12}, the relative evidence of Model 1 compared to Model 2. If $B_{12} > 1$, the evidence is in favor Model 1 and if $B_{12} < 1$, the evidence is in favor of Model 2. Bayes factors are convenient to interpret. A Bayes factor of $B_{12} = 10$, for example, indicates that there is 10 times more evidence in favor of Model 1 compared to Model 2. Further, the Bayes factor of Model 1 compared to Model 2, B_{12}, can be easily converted to the Bayes factor of Model 2 compared to Model 1 with $B_{21} = 1 / B_{12}$. Table 3 lists a categorical interpretation of the Bayes factor as suggested by Lee and Wagenmakers (2014). We describe the parameter-estimation and model-selection scheme for Eqs. 1 through 5 in the Appendix. We implemented the model in the R statistics environment (R Core Team, 2016).

Table 3 Interpretation of the Bayes factor B_{12}, the relative evidence in favor of Model 1 compared to Model 2 (Lee & Wagenmakers, 2014)

Bayes factor B_{12}					Interpretation
100	<	B_{12}			Extreme evidence for Model 1
30	<	B_{12}	≤	100	Very strong evidence for Model 1
10	<	B_{12}	≤	30	Strong evidence for Model 1
3	<	B_{12}	≤	10	Moderate evidence for Model 1
1	<	B_{12}	≤	3	Anecdotal evidence for Model 1

Coding of moderator variables

We took the information necessary for the coding of each moderator directly from the studies listed in Table 2. Given the straightforward coding scheme for each moderator and the objective nature of each variable, we did not assess interrater reliability.

Basis for judgments (recollection vs. familiarity vs. fluency) The hypotheses listed in Table 1 make diverging predictions about which type of judgment the revelation effect influences. According to the discrepancy-attribution hypothesis, the revelation effect results from a misattribution of fluency. Thus, the revelation effect should occur for judgments linked to fluency, such as recognition, truth, preference, or typicality judgments (Alter & Oppenheimer, 2009). However, most of the hypotheses listed in Table 1 focus on episodic memory and the recollection-familiarity dichotomy (Mandler, 1980). For example, some authors classified certain recognition judgments as primarily recollection-based (Cameron & Hockley, 2000; Westerman, 2000). Such judgments include associative recognition judgments, in which participants decide whether two test items appeared together in the study list; rhyme-recognition judgments, in which participants decide whether the test item rhymes with a study-list item; or plurality judgments, in which participants decide whether the test item appeared in the study list in singular or plural form. Conversely, recognition judgments such as old–new recognition or recognition-confidence judgments have been linked to familiarity-based recognition (Cameron & Hockley, 2000; Westerman, 2000).

In primary studies, there is evidence that the revelation effect influences familiarity-based judgments rather than recollection-based judgments. In the process-dissociation procedure (Jacoby, 1991), the preceding task only increases estimates of familiarity, not recollection. Similarly, among remember-know judgments (Tulving, 1985), the preceding task reduces "remember" responses for targets and increases "know" responses for targets and distractors (LeCompte, 1995). Further, judgments that require recollection, such as associative-recognition judgments (e.g., Did "*window*" and "*leopard*" appear together?), do not yield a revelation effect (Westerman, 2000). However, this changes if study opportunities are minimal due to very brief presentation times or stimulus masking. Without the prerequisites for recollection-based strategies, participants revert to familiarity-based strategies, producing a revelation effect in associative-recognition judgments (Cameron & Hockley, 2000). Similarly, asking participants to semantically encode study items, typically thought to aid recollection, reduces the revelation effect for old-new recognition judgments (Mulligan, 2007). Further, rhyme recognition ("*Does this word rhyme with any of the words you studied earlier?*"),

a judgment assumed to recruit recollection-based strategies, does not produce a revelation effect. Again, this changes when study opportunities are minimal, thus increasing the use of familiarity-based strategies (Mulligan, 2007). Finally, measures that increase the probability that participants rely on familiarity compared to recollection, such as long retention intervals or brief study durations, increase the revelation effect (Landau, 2001).

To determine whether the meta-analysis corroborates the findings of primary studies and whether fluency is a viable basis of the revelation effect, we defined a moderator variable *basis of judgment* with three categories, *recollection*, *familiarity*, and *fluency*. We coded this moderator as *recollection* in all studies that included associative recognition, rhyme recognition, or plurality judgments. We coded the moderator as *familiarity* in all studies that included old–new recognition or recognition-confidence judgments. We do not mean to imply that these judgments are free of recollection in typical revelation-effect studies. However, several authors argue that these judgments more likely rely on familiarity compared to associative recognition, rhyme recognition, or plurality judgments (Cameron & Hockley, 2000; Mulligan, 2007; Westerman, 2000). Finally, we coded the moderator as *fluency* in all studies that included truth, preference, or typicality judgments. Fluency is one proposed basis for these judgments (Alter & Oppenheimer, 2009). Note that familiarity-based judgments are also routinely associated with fluency (e.g., Whittlesea & Williams, 1998, 2000). We only separate familiarity-based judgments from the fluency category to simplify the hypothesis tests.

Table 4 lists the types of judgments categorized as recollection based, familiarity based, or fluency based in the present meta-analysis. Although Watkins and Peynircioglu (1990) considered some of the frequency judgments they obtained to be nonepisodic, it could be argued that judgments of frequency rely on the number of memory episodes containing a specific stimulus. Because the status of frequency judgments is debatable, we did not include them in the moderator basis of judgments.

Similarity (same vs. related vs. different) According to some hypotheses, the revelation effect occurs because the task item is similar to the test item (see the Hypotheses section). Empirically, the evidence for this claim is mixed. Westerman and Greene (1998) found evidence that the revelation effect occurs only when task and test items are of the same type (e.g., both words). For example, an anagram of a word (e.g., *esnuihsn–sunshine*) but not an arithmetic task (e.g., 248 + 372 = ?) elicited a revelation effect for words as recognition test items. Similarly, in an experiment by Bornstein, Robicheaux, and Elliot (2015), the size of the revelation effect increased as a function of the similarity between the task item and the test item. The effect was largest when the task item was identical to the test item (e.g., *cream–cream*), weaker when the task item was semantically related to the test item (e.g., *coffee–cream*), and weakest when the task item was semantically unrelated to the test item (e.g., *shore–cream*).

However, other studies found no link between the similarity of the task and test items and the size of the revelation effect. For example, some studies found that the revelation effect also appears when the preceding task involves arithmetic problems or anagrams consisting solely of numbers, followed by words as recognition test items (Niewiadomski & Hockley, 2001; Verde & Rotello, 2003). Further, high compared to low conceptual or perceptual similarity between a revealed word and the recognition test item did not increase the revelation effect in a study by Peynircioglu and Tekcan (1993); the revelation effect was equally large when the word in the preceding task and the recognition test item were orthographically and phonetically similar (e.g., *state–stare*) or dissimilar (e.g., *lion–obey*).

These conflicting results are difficult to reconcile. However, by looking beyond the dichotomous significance patterns of primary studies and relying on effect sizes instead, we hoped to identify a clear data pattern. Critically, to test hypotheses that predict a link between similarity and the size of the revelation effect, we defined a moderator variable *similarity*. We coded all effect sizes as *same* when the task item was identical to the test item, for example, when

Table 4 Overview of the judgment types coded in the basis-of-judgment moderator with an example item and the moderator category to which the judgment contributed

Judgment	Example	Category
Associative recognition	Did you see "*window*" and "*leopard*" together?	Recollection
Rhyme recognition	Does "*window*" rhyme with a word from the study list?	Recollection
Plurality	Did you see "*window*" or" *windows*" before?	Recollection
Old–new recognition	Did you see "*window*" before?	Familiarity
Recognition confidence	How confident are you that you saw "*window*" before?	Familiarity
Truth	Is the "*leopard*" the fastest animal?	Fluency
Preference	Indicate your preference for *Coke* compared to other colas such as Pepsi and Jolt.	Fluency
Typicality	How typical is "*poodle*" for the category dog?	Fluency

participants first solved an anagram (e.g., *esnuihsn*) and then judged whether the anagram solution (e.g., *sunshine*) was old or new. We coded all effect sizes as *related* when the task and test items were not identical but of the same stimulus type. As stimulus types we considered words, numbers, pictures, and nonwords. We considered a condition as related when, for example, participants solved an anagram (e.g., *esnuihsn*) and then judged whether another word (e.g., *airplane*) was old or new. Finally, we coded all effect sizes as *different* when the task and test items were not identical and of a different type, for example, when participants first solved an addition problem (e.g., 248 + 372 = ?) and then judged whether a word (e.g., *sunshine*) was old or new. Further, we excluded studies from the similarity moderator when it was unclear whether the task item and the test item were of the same type (e.g., words). For example, some studies used letters in the preceding task (e.g., memory span) and words as test items. Other studies used both numbers and words in the preceding task—for example, participants solved an anagram, followed by an arithmetic task—or words as task items and statements (e.g., "*The leopard is the fastest animal*") as test items.

Difficulty (hard vs. easy) A key prediction of some hypotheses is that hard preceding tasks should elicit a larger revelation effect than easy preceding tasks (see the Hypotheses section). To test these hypotheses, we defined a moderator *difficulty* with two levels: *hard* and *easy*. We coded the difficulty moderator only for studies which compared a relatively hard task with a relatively easy task either within-subject or between-subjects but not between experiments. For example, Watkins and Peynircioglu (1990) compared the revelation effect when participants solved eight- or five-letter anagrams. We coded eight-letter anagrams as "hard" and five-letter anagrams as "easy." Similarly, in other studies, participants transposed eight- versus three-digit numbers or completed an eight- vs. versus three-letter memory span task (Westerman & Greene, 1998; Verde & Rotello, 2003). Again, we coded eight-digit numbers and 8-letter memory spans as "hard" and three-digit numbers and three-letter memory spans as "easy." Other studies manipulated the number of starting letters in letter-by-letter revelation tasks (Bornstein & Neely, 2001; Watkins & Peynircioglu, 1990). In this task, participants had to identify fragmented words whose missing letters were revealed sequentially (e.g., 1. *su_s_ _ne*, 2. *su_sh_ne*, 3. *sunsh_ne*, 4. *sunshine*). In Bornstein and Neely's (2001) experiment, participants identified word fragments with two, four, or six starting letters. We coded word fragments with two starting letters as "hard" (e.g., *_ _n_ _i_ _*) and fragments with six starting letters (e.g., *su_sh_ne*) as "easy." Because the difficulty of the condition with four starting letters was less clear, we did not include this condition in the analysis. Similarly, Watkins and Peynircioglu (1990) varied the number of starting letters between two and seven. We coded the conditions with two, three, and four starting letters as "hard" and the conditions with five, six, or seven letters as "easy." Niewiadomski and Hockley (2001) took a more extensive approach. In several experiments, preceding tasks consisted either of single or double tasks. In single tasks, participants solved an anagram or an arithmetic task. Conversely, in double tasks, participants solved two anagrams, two arithmetic tasks, or a mixture of one anagram and one arithmetic task in quick succession. Again, the rationale for these experiments was that double tasks should be harder than single tasks. Therefore, we coded double tasks as "hard" and single tasks as "easy."

General analyses

As a prerequisite for the moderator analysis and for the sake of completeness, we report estimates of the overall effect size before turning to the hypothesis tests. We additionally attempted to identify potential confounds for the moderator analysis, in particular, experimental procedures, stimulus materials, or participant groups that in the past produced unexpected or extreme results. We then excluded studies with potential confounds from the moderator analysis to minimize the chance of spurious results. For all effect-size estimates, we report point estimates along with 95% credible intervals (95%–CI). We further tested the possibility of publication bias in the revelation-effect literature but found no problematic data patterns (see Supplemental Material).

Overall effect size

We first applied the meta-analysis to the overall data set which included 34 studies, 143 experiments, 368 effect sizes, and 5,034 participants. The results of the meta-analysis depend on whether dependent effect sizes occur within studies or only within experiments. As described before, we define studies as one or more experiments based on different participant samples but part of the same publication. We further define an experiment as one or more effect sizes based on the same participant sample. Arguably, multiple experiments by the same research group will be dependent and, consequently, effect sizes within studies are dependent. Conversely, our estimate of the correlation between dependent effect sizes explicitly relies on repeated measures within experiments (i.e., participant samples) not within studies. The correlation of effect sizes within experiments likely overestimates the true correlation between effect sizes within studies. The results of the meta-analysis further depend on the assumed correlation between dependent effect sizes. To avoid results that depend on a specific set of assumptions, we present important statistics for models that assume dependencies either on the study or the experiment level. We further present results for assumed

correlations $\hat{\rho} = \{.1, .3, .5\}$ between dependent effect sizes for all analyses.

Table 5 lists the combined effect-size estimates for the overall data set. For comparison purposes, Table 5 also lists the combined effect-size estimates based on a univariate random-effects model. This model relies on synthetic effect sizes computed across dependent effect sizes (Borenstein et al., 2009). The combined effect-size estimates in Table 5 and their credible intervals indicate a reliable revelation effect for the overall data set. When we compared Model 1 that freely estimates the overall effect size with Model 2 that assumes that the overall effect size is 0 (i.e., $\gamma = 0$), the resulting Bayes factors were overwhelmingly in favor of Model 1 (all $B_{12} > 100$). The corresponding precision estimates ($\hat{\tau}_k$) further suggested considerable variability in the regression coefficients between studies/experiments with variance estimates ranging between 0.02 and 0.05, supporting our decision to use a random-effects model.

Potential confounds

Not all attempts to elicit a revelation effect have succeeded. From a meta-analytic viewpoint, these failed attempts are particularly important. When the units of observation, such as experiments in a meta-analysis, cannot be assigned randomly to comparison groups, these groups must be matched on potential confounds to avoid spurious results.

We found four variables that represent potential confounds in the meta-regression. In recognition experiments, the revelation effect occurs because of the preceding task during the recognition test; however, some publications also included a preceding task during study. Participants studying some words by reading the words and other words by solving anagrams were later, in a recognition test, more likely to call words "new" after solving an anagram compared to words without an anagram (Dewhurst & Knott, 2010; Mulligan & Lozito, 2006). This corresponds to negative effect sizes, that is, a reversed revelation effect, and occurred for 11 out of 29 effect sizes (38%) with study anagrams compared to just 15% negative effect sizes with respect to the entire data set. The second potential confound regarded participant age. Although most studies provided only a crude indicator of age (e.g., "undergraduates"), some studies specifically investigated the revelation effect for varying age groups (Guttentag & Dunn, 2003; Hodgson, n.d.; Prull et al., 1998; Thapar & Sniezek, 2008). We found that six out of 10 effect sizes for elderly participants were negative. The third potential confound regarded stimulus materials. Experiments in which participants solved a picture puzzle of a face exclusively led to negative effect sizes (Aßfalg & Bernstein, 2012). The fourth and final potential confound regarded whether participants had successfully completed the preceding task. Dougal and Schooler (2007) found an increase in "old" responses only when participants successfully solved an anagram compared to the no-task condition. Conversely, unsolved anagrams did not produce a revelation effect. In our data set, out of six effect sizes belonging to unsolved anagrams, four were negative (67%).

To determine the impact of these potential confounds, we compared effect sizes with and without potential confounds. Because effect sizes with a potential confound were relatively rare, we pooled all potential confounds into one predictor. We then applied our meta-analysis approach to compare two models. Model 1 assumes that effect sizes with and without confounds share the same true effect size. Conversely, Model 2 estimates separate true effect sizes for the condition with and without potential confounds.

Table 6 lists the Bayes factors in favor of Model 2 (i.e., different true effect sizes with and without potential confounds). Overall, the evidence is strongly in favor of potential confounds resulting in different effect sizes than other studies. Assuming dependencies on the experiment level and a moderate correlation of $\hat{\rho} = .3$, the estimate of the true effect size for studies without potential confound, 0.30, 95%–CI = [0.26, 0.34], was larger than the estimate of the true effect size for studies with one of the potential confounds, 0.01, 95%–CI = [-0.15, 0.16]. Without additional data it is difficult to determine which of the conditions we marked as potential confound is indeed a confound. However, we chose a conservative approach and decided to remove all experiments containing a potential confound from all remaining analyses.

Table 5 Combined effect-size estimates (γ), credible intervals (multivariate), and confidence intervals (univariate), as a function of assumed dependencies between effect sizes on the study versus experiment level, the meta-analysis model, and the assumed correlation between dependent effect sizes ($\hat{\rho}$)

Dependencies	Model	Correlation between dependent effect sizes		
		$\hat{\rho} = .1$	$\hat{\rho} = .3$	$\hat{\rho} = .5$
Study	Random-effects, multivariate	0.26 [0.21, 0.32]	0.23 [0.19, 0.28]	0.21 [0.15, 0.26]
Experiment		0.27 [0.23, 0.31]	0.27 [0.23, 0.31]	0.26 [0.21, 0.3]
Study	Random-effects, univariate	0.29 [0.23, 0.34]	0.28 [0.23, 0.34]	0.28 [0.22, 0.34]
Experiment		0.28 [0.24, 0.32]	0.28 [0.24, 0.32]	0.28 [0.24, 0.32]

Table 6 Bayes factors in favor of potential confounds being different from other effect sizes as a function of assumed dependencies between effect sizes on the study versus experiment level and the assumed correlation between dependent effect sizes ($\hat{\rho}$)

Dependencies	Correlation between dependent effect sizes		
	$\hat{\rho} = .1$	$\hat{\rho} = .3$	$\hat{\rho} = .5$
Study	>100	>100	>100
Experiment	23.39	61.50	>100

Table 7 Bayes factors of Model 2 ($\gamma_{familiarity} = \gamma_{fluency} > \gamma_{recollection}$) relative to Model 1 ($\gamma_{familiarity} > \gamma_{recollection}$, $\gamma_{fluency} = 0$) as a function of assumed dependencies between effect sizes on the study versus experiment level and the assumed correlation between dependent effect sizes ($\hat{\rho}$)

Dependencies	Correlation between dependent effect sizes		
	$\hat{\rho} = .1$	$\hat{\rho} = .3$	$\hat{\rho} = .5$
Study	5.29	3.65	2.27
Experiment	3.81	3.10	2.85

Hypothesis tests

Basis of judgments (recollection vs. familiarity vs. fluency)

All hypotheses in Table 1 predict a smaller revelation effect when the judgment relies on recollection (e.g., associative recognition, rhyme recognition, or plurality judgments) rather than familiarity (e.g., old-new recognition or recognition-confidence judgments). Most hypotheses also predict the absence of a revelation effect for fluency-based judgments (e.g., truth, preference, or typicality judgments). In contrast, the discrepancy-attribution hypothesis predicts a revelation effect for fluency-based judgments and familiarity-based judgments alike, the latter being typically considered another form of fluency-based judgment (e.g., Alter & Oppenheimer, 2009). The discrepancy-attribution hypothesis also predicts a smaller revelation effect for judgments that rely on recollection rather than fluency or familiarity.

Based on these predictions, we specified two models with regard to the true effect size γ. Model 1 ($\gamma_{familiarity} > \gamma_{recollection}$, $\gamma_{fluency} = 0$) represents the prediction of most revelation-effect hypotheses and Model 2 ($\gamma_{familiarity} = \gamma_{fluency} > \gamma_{recollection}$) represents the prediction of the discrepancy-attribution hypothesis. To test these models, we performed a meta-analysis with basis of judgments as a dummy-coded moderator variable. We included only experiments in which the type of judgment was either recollection, familiarity, or fluency-based (see Coding of moderator variables section). We further excluded all experiments from the moderator analysis which contained at least one potential confound (see Potential confounds section). The final selection included 29 studies, 103 experiments, 272 effect sizes, and 3,710 participants.

Table 7 lists the Bayes factors of Model 2 ($\gamma_{familiarity} = \gamma_{fluency} > \gamma_{recollection}$) compared to Model 1 ($\gamma_{familiarity} > \gamma_{recollection}$, $\gamma_{fluency} = 0$). Evidence suggesting that Model 2 is more likely than Model 1 was mostly "moderate," sometimes bordering on "anecdotal" (see Table 3). Figure 5 contains the individual effect sizes for basis of judgments (familiarity vs. fluency vs. recollection) along with the posterior densities of the corresponding true effect sizes. Assuming dependencies on the experiment level and a moderate correlation of $\hat{\rho} = .3$, the estimate of the true effect size tended to be larger in the condition familiarity, 0.35, 95%–CI = [0.31, 0.38], than in the conditions fluency, 0.24, 95%–CI = [0.14, 0.35], and recollection, 0.14, 95%–CI = [0.04, 0.24]. Overall, the evidence favors the discrepancy-attribution hypothesis over all other revelation-effect hypotheses. However, this evidence was moderate at best.

Similarity (same vs. related vs. different)

Most hypotheses in Table 1 predict equally-sized revelation-effects across similarity conditions. Conversely, the global-matching hypothesis predicts that the revelation effect is equally large in the same (e.g., *sunshine–sunshine*) and related conditions (e.g., *airplane–sunshine*), but smaller in the different condition (e.g., [248 + 372]-*sunshine*).

Based on these predictions, we specified two models with regard to the true effect size γ. Model 1 ($\gamma_{same} = \gamma_{related} = \gamma_{different}$) represents the prediction of most revelation-effect hypotheses and Model 2 ($\gamma_{same} = \gamma_{related} > \gamma_{different}$) represents the prediction of the global-matching hypothesis. To test these models, we performed a meta-analysis with similarity as a dummy-coded moderator variable. To minimize the influence of judgment type on the results, we included only studies in which judgments were either old-new or recognition-confidence. We further excluded all studies from the moderator analysis which contained at least one potential confound (see Potential confounds section). The final selection included 27 studies, 88 experiments, 225 effect sizes, and 3,296 participants.

Table 8 lists the Bayes factors of Model 1 ($\gamma_{same} = \gamma_{related} = \gamma_{different}$) compared to Model 2 ($\gamma_{same} = \gamma_{related} > \gamma_{different}$). Evidence suggesting that Model 1 is more likely than Model 2 was again mostly "moderate" and sometimes merely "anecdotal" (see Table 3). Figure 6 contains the individual effect sizes for similarity (same vs. related vs. different) along with the posterior densities of the corresponding true effect sizes. Assuming dependencies on the experiment level and a moderate correlation of $\hat{\rho} = .3$, the estimate of the true effect

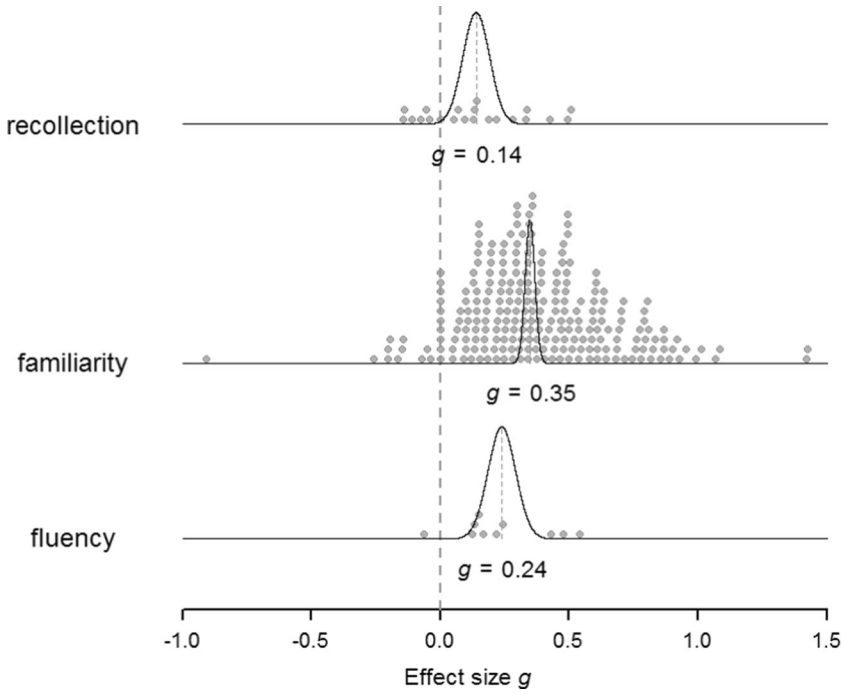

Fig. 5 Distribution of individual effect sizes (gray circles) and the corresponding posterior density of the true effect size as a function of basis of judgment (recollection vs. familiarity vs. fluency). Posterior densities rely on the assumption of dependent effect sizes on the experiment level with $\hat{\rho} = .3$

size tended to be about equal in the conditions same, 0.35, 95%–CI = [0.29, 0.41], related, 0.40, 95%–CI = [0.35, 0.45], and different, 0.31, 95%–CI = [0.20, 0.42]. Overall, the evidence contradicts the global-matching hypothesis. However, this evidence was moderate at best.

Difficulty (hard vs. easy)

The global-matching and criterion-flux hypotheses predict equally-sized revelation effects for hard compared to easy preceding tasks. Conversely, the decrement-to-familiarity and discrepancy-attribution hypotheses predict a larger revelation effect for hard compared to easy preceding tasks.

To test the moderating effect of difficulty, we performed a meta-regression with difficulty as a dummy-coded moderator variable. We included only studies in this analysis that explicitly compared a hard version of a preceding task (e.g., *__n__i__*) with a relatively easy version (e.g., *su_sh_ne*; see the Coding of moderators section). We excluded all studies from the moderator analysis that contained at least one potential confound (see Potential confounds section). The final selection included five studies, 15 experiments, 48 effect sizes, and 517 participants.

Table 9 lists the Bayes factors of Model 2 ($\gamma_{hard} = \gamma_{easy}$) compared to Model 1 ($\gamma_{hard} > \gamma_{easy}$). Evidence suggesting that Model 2 is more likely than Model 1 was "moderate" or "anecdotal" (see Table 3). Figure 7 contains the individual effect sizes for difficulty (hard vs. easy) along with the posterior densities of the corresponding true effect sizes. Assuming dependencies on the experiment level and a moderate correlation of $\hat{\rho} = .3$, the estimate of the true effect size was larger in the hard condition, 0.49, 95%–CI = [0.39, 0.61], compared to the easy condition, 0.37, 95%–CI = [0.28, 0.45]. However, the evidence favors the assumption of the global-matching and criterion-flux hypotheses that the revelation effect is about equal in the hard and easy conditions. This evidence was again moderate at best.

Table 8 Bayes factors of Model 1 ($\gamma_{same} = \gamma_{related} = \gamma_{different}$) relative to Model 2 ($\gamma_{same} = \gamma_{related} > \gamma_{different}$) as a function of assumed dependencies between effect sizes on the study versus experiment level and the assumed correlation between dependent effect sizes ($\hat{\rho}$)

Dependencies	Correlation between dependent effect sizes		
	$\hat{\rho} = .1$	$\hat{\rho} = .3$	$\hat{\rho} = .5$
Study	5.29	3.65	2.27
Experiment	3.81	3.10	2.85

Discussion

We tested the predictions of four revelation-effect hypotheses in a meta-analysis spanning 26 years: the global-matching hypothesis (Westerman & Greene, 1998), the

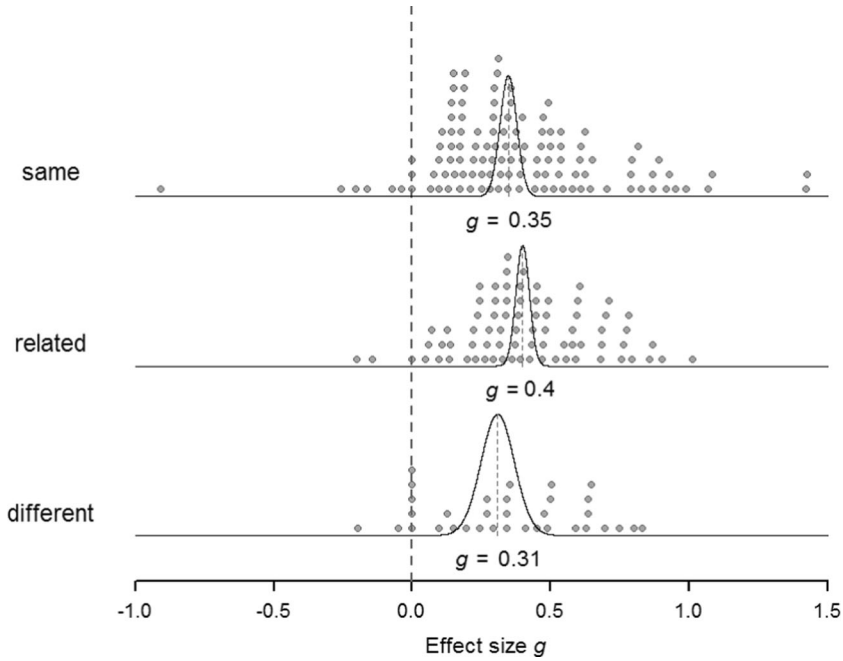

Fig. 6 Distribution of individual effect sizes (gray circles) and the corresponding posterior density of the true effect size as a function of similarity (same vs. related vs. different). Posterior densities rely on the assumption of dependent effect sizes on the experiment level with $\hat{\rho} = .3$

decrement-to-familiarity hypothesis (Hicks & Marsh, 1998), the criterion-flux hypothesis (Niewiadosmki & Hockley, 2001), and the discrepancy-attribution hypothesis (Whittlesea & Williams, 2001). We defined statistical models tailored specifically to the predictions of the four hypotheses. In a series of hypothesis tests, we assessed the amount of evidence for these models relative to each other. Each test indicated evidence in favor of one model over another. This evidence was mostly moderate, but sometimes merely anecdotal, suggesting a lack of data. The evidence for or against a statistical model is an indicator of the veracity of the corresponding revelation-effect hypothesis. None of the tested revelation-effect hypotheses predicted the entire data pattern. We observed two out of three correct predictions for the criterion-flux hypothesis and the discrepancy-attribution hypothesis, but only one out of three correct predictions for the other hypotheses.

Table 9 Bayes factors in favor of Model 2 ($\gamma_{hard} = \gamma_{easy}$) compared to Model 1 ($\gamma_{hard} > \gamma_{easy}$) as a function of assumed dependencies between effect sizes on the study vs. experiment level and the assumed correlation between dependent effect sizes ($\hat{\rho}$)

Dependencies	Correlation between dependent effect sizes		
	$\hat{\rho} = .1$	$\hat{\rho} = .3$	$\hat{\rho} = .5$
Study	5.90	4.43	2.33
Experiment	3.10	2.45	1.23

Limitations

Extant revelation-effect hypotheses tend to be vague. The hypotheses' predictions we listed in Table 1 are accurate in as far as they reflect either a core assumption of a hypothesis or a stated/implied auxiliary assumption by its authors. However, in some instances it is difficult to see how some of the predictions can be derived from a specific hypothesis. This issue results from the use of vague concepts and vague descriptions of the interplay between these concepts. Unfortunately, solving this issue is beyond the scope of the present evaluation. Ideally, hypotheses should be formalized mathematically to avoid vagueness and to provide quantitative instead of merely qualitative predictions.

Another limitation concerns the coding of variables based on the hypothesis predictions. The difficulty of the preceding task (hard vs. easy) and similarity between task and test item (same vs. related vs. different) are easier to objectively index than the variable, basis of judgment (recollection vs. familiarity vs. fluency). Whereas difficulty and similarity represent task characteristics, basis of judgment relies on theoretical constructs. Consequently, our conclusions regarding basis of judgment rely on auxiliary assumptions such as a dual-process model of memory that separates recollection and familiarity or the existence of fluency as a cognitive construct. Although the revelation effect literature in particular and the literature on memory and fluency in general consistently rely on these theoretical constructs, this does not mean that they are undisputed (Alter & Oppenheimer, 2009; Cameron & Hockley, 2000; LeCompte, 1995; Mandler, 1980; Mulligan, 2007; Westerman, 2000; Yonelinas, 2002).

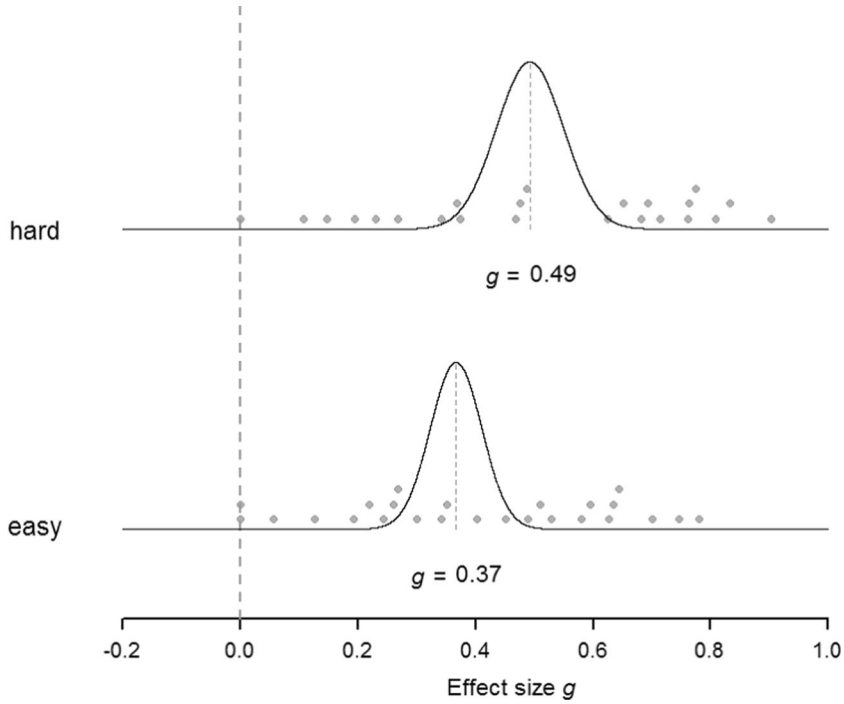

Fig. 7 Distribution of individual effect sizes (gray circles) and the corresponding posterior density of the true effect size as a function of difficulty (hard vs. easy). Posterior densities rely on the assumption of dependent effect sizes on the experiment level with $\hat{\rho} = .3$

Another limitation of the present meta-analysis concerns our statistical approach. We used a multivariate meta-analysis to account for dependent effect sizes and a random-effects model to account for the considerable variability in revelation-effect studies. We further tested revelation-effect hypotheses with a series of model selection tests which overcome some of the disadvantages of null-hypothesis testing. Using the Bayes factor for model selection further allowed us to quantify the amount of evidence for one model over another. However, no amount of sophistication in statistical modeling can fully account for the complexity of behavioral data. For example, our approach relies on the assumption that correlations between dependent effect sizes are homogeneous across studies which is, at most, a good approximation to reality.

Finally, our hypothesis tests revealed only moderate evidence at best for one model over another. This suggests a lack of data regarding some of the key predictions of revelation-effect hypotheses. Sparse data limits the conclusions that can be drawn based on this data. It also suggests areas of research that would particularly benefit from additional research.

Evaluation of hypotheses

In this study, we used meta-analytic procedures to test the predictions of revelation-effect hypotheses. All hypotheses failed to predict at least one outcome. However, these hypotheses have strengths and weaknesses that extend beyond the meta-analysis. For example, primary studies provide evidence against some hypotheses that is not reflected in the meta-analysis. In this section, we evaluate each hypothesis regarding the results of the meta-analysis, the outcome of primary studies, and methodological considerations.

Global-matching hypothesis In the present analyses, the global-matching hypothesis falsely predicted that the revelation effect should be absent for fluency-based judgments and that the revelation effect should be moderated by the degree of similarity between the task and test items. Both falsifications touch a key assumption of the global-matching hypothesis: If the preceding task is similar to the test item, then it activates study-list memory traces. This activation then contributes to the perceived familiarity of the test item and causes the revelation effect.

With regard to primary studies, the global-matching hypothesis also struggles to account for the outcome of a study by Frigo, Reas, and LeCompte (1999). These authors found that a study list is unnecessary for a revelation effect to appear. In their study, participants listened to a list of words and, later, performed a recognition test with a preceding task. Unsurprisingly, a revelation effect occurred. However, unbeknownst to participants, the study list did not contain any actual words. Instead, participants had listened to radio static and were told that the static contained words that were difficult to perceive. In the absence of any study list, the activation of study-list memory traces cannot be responsible for the revelation effect.

Decrement-to-familiarity hypothesis In the present analyses, the decrement-to-familiarity hypothesis falsely predicted that the revelation effect is absent for fluency-based judgments and that the revelation effect is larger for hard compared to easy preceding tasks. According to the decrement-to-familiarity hypothesis, the preceding task activates memory traces that render the recognition judgment more difficult than in the no-task condition. This increased difficulty then triggers a more liberal response criterion, which in turn causes the revelation effect. The observation that task difficulty does not influence the revelation effect is troubling for the decrement-to-familiarity hypothesis. However, it could be argued that the manipulations of task difficulty that contributed to the meta-analysis are flawed. For example, eight- compared to five-letter anagrams might be too similar with regard to task difficulty to detect a difference in the revelation effect. Similarly, it is not guaranteed that solving two anagrams in a row is perceived as harder compared to a single anagram. The lack of manipulation checks for the task difficulty manipulations in these studies further makes drawing conclusions difficult.

With regard to primary studies, the decrement-to-familiarity hypothesis received support by studies measuring event-related potentials. Familiarity, measured as activity in the anterior, frontal scalp region 300–500 ms after the onset of the test item, is lower after a preceding task than without a preceding task (Azimian-Faridani & Wilding, 2004). Later research, however, found this data pattern only when the task item was identical to the test item, not when it was different (Leynes, Landau, Walker, & Adante, 2005). Thus, there is no evidence of a familiarity decrease when the task item and the test item are different, despite a clear revelation effect in the meta-analysis. This casts doubt on the assumption that a familiarity decrement could be responsible for all extant revelation effect data.

Lending support to the idea that the preceding task activates memory-traces, the effect of the preceding task on subsequent judgments seems to disappear over time. Bernstein and colleagues (Bernstein et al., 2009) found a revelation effect when participants first solved anagrams and then judged autobiographical questions. For example, directly after solving an anagram of the word *window*, participants were more likely to claim having broken a window playing ball as a child. However, this revelation effect disappeared when participants had to wait 20 s between solving the anagram and answering the autobiographical-memory question.

Criterion-flux hypothesis In the present analyses, the criterion-flux hypothesis falsely predicted that the revelation effect would be absent for fluency-based judgments. According to this hypothesis, the preceding task removes task-relevant information from working memory. This temporarily causes participants to adopt a more liberal response criterion which in turn causes the revelation effect.

A problematic aspect of the criterion-flux hypothesis is its vagueness in some areas. The hypothesis does not specify which information the preceding task removes from working memory. It is also unclear why and how this information is recovered after the recognition judgment. Further, the degree of working-memory involvement in a preceding task is difficult to assess. This hinders the primary means to test the criterion-flux hypothesis. Finally, with regard to primary studies, the criterion-flux hypothesis is incompatible with the observation that a memory span task involving numbers, a classic working-memory task, did not induce a revelation effect in Westerman and Greene's (1998) study.

Discrepancy-attribution hypothesis In the present analyses, the discrepancy-attribution hypothesis falsely predicted a larger revelation effect for hard compared to easy preceding tasks. However, as the only hypothesis, it correctly predicted a revelation effect for judgments that are commonly associated with fluency (e.g., truth, preference, typicality). According to the discrepancy-attribution hypothesis, the preceding task is perceived as relatively disfluent. This perception induces the expectation of continuing disfluency. However, if the test item is then perceived as relatively fluent, this fluency is misattributed to familiarity, truth, or preference, depending on the judgment required by the participants.

Fluency-based hypotheses have been popular (see Alter & Oppenheimer, 2009, for a review) but suffer, in our opinion, from offering too many degrees of freedom for researchers to fit the hypothesis to their results. A brief overview of the literature on fluency reveals in what diverse ways fluency is thought to affect judgments. For example, some authors argue that primed words are processed more fluently (e.g., Whittlesea et al., 1990). However, Huber, Clark, Curran, and Winkielman (2008) argue that long-duration primes appearing for about 1 s cause habituation rather than a fluency increase. This would explain why short-duration primes cause an increase of the reported familiarity in matching recognition test items whereas long-duration test items cause a decrease (Huber et al., 2008; Jacoby & Whitehouse, 1989). Other authors argue that not fluency per se, but a surprising discrepancy between expected and observed fluency influences judgments (Whittlesea & Williams, 1998). Further, Oppenheimer (2006) argues that participants can discount fluency when they become aware of the source of the fluency and realize that this source is not diagnostic for their judgment. Because the process of discounting is not well calibrated, participants tend to overdiscount. This represents a reversal of fluency-based effects. In sum, the mechanisms summarized here can account for fluency increasing familiarity, fluency not affecting familiarity (e.g., if fluency was not surprising given the context) or fluency reducing familiarity (e.g., if

participants over-discounted fluency). Without independent measures of fluency, surprise, and awareness, it is unclear what observation would falsify fluency-based hypotheses.

Future directions

Extant revelation-effect hypotheses tend to offer relatively vague predictions. This situation could be improved with hypotheses that are formalized in mathematical models. The following points have been made before but they are worth reiterating (Bjork, 1973; Hintzman, 1991). Mathematical models would increase the precision of predictions by offering specific context-dependent estimates of the size of the revelation effect. These models would be easier to falsify than extant hypotheses. Consequently, the process of identifying critical experiments to adjudicate between competing hypotheses would be easier. In the present study, only the global-matching hypothesis relies on a formal model of memory. However, we found that it was not the formal part of the hypothesis but relatively vague verbal extensions of the model to accommodate the revelation effect that proved to be problematic. Another formal model that has been used extensively in revelation-effect research is the signal detection theory. Although this model helps to conceptualize the revelation effect as a criterion shift, it does not explain why this criterion shift occurs. Thus, formal models will likely have to go beyond current hypotheses by formally defining knowledge structures and decision processes that operate on them. Besides the challenge of explaining the revelation effect, theories will also have to account for the observation in the present meta-analysis that recollection-based judgments are less susceptible to the revelation effect than other judgments.

Conclusions

Twenty-six years of research shows that the revelation effect represents a robust and general cognitive phenomenon. In this meta-analysis, all extant revelation-effect hypotheses were unable to account for the observed data patterns. However, the available data is more compatible with some hypotheses (criterion flux and discrepancy attribution) than with others (global matching and decrement to familiarity). A general theme in this work is that revelation-effect hypotheses tend to be vague. A move from imprecise verbal accounts to mathematical models would increase the testability of hypotheses. We further observed a lack of data, particularly regarding fluency-based judgments and manipulations of task difficulty. This limits the conclusions that can be drawn from tests of revelation-effect hypotheses. Such future research would further illuminate the nature of the revelation effect.

Author Note Work on this manuscript was funded by a German Research Foundation grant (AS 427/1-1) and a Kwantlen Polytechnic University grant (HSS 651201) to André Aßfalg and a Canada Research Chair Award (950-228407) to Daniel M. Bernstein. We are grateful to David E. Huber, Christoph Klauer, Eric Mah, and an anonymous reviewer for commenting on an earlier version of this manuscript.

Appendix

Here, we briefly review our meta-analysis model and describe a parameter-estimation procedure for the model based on the Markov-chain-Monte-Carlo (MCMC) approach. We further propose a model-selection procedure using the Bayes factor combined with the model-sampling approach suggested by Carlin and Chib (1995). We conclude with a small simulation study that illustrates the statistical power of this model-selection approach for several combinations of parameter restrictions, effect sizes, and sample sizes.

The model

Let y_{ij} be the $i = 1, \ldots, N_j$ effect sizes in study $j = 1, \ldots, M$. Further, let $k = 1, \ldots, L$ denote the predictors in a regression analysis. Then, using matrix notation, the regression equation can be written as

$$\mathbf{y}_j \sim N\left(\mathbf{X}_j \mathbf{b}_j, \Sigma_{ej}\right), \tag{6}$$

where the vector of N_j effect sizes $\mathbf{y}_j$ depends on a $N_j \times L$ design matrix $\mathbf{X}_j$, a $L \times 1$ vector of regression coefficients $\mathbf{b}_j$, and a $N_j \times N_j$ variance-covariance matrix Σ_{ej}. The entries in Σ_{ej} can be estimated from the data and are assumed to be known (Raudenbush, Becker, & Kalain, 1988).

The prior distribution of the regression coefficients b_{jk} is

$$b_{jk} \sim N(\gamma_k, \tau_k), \tag{7}$$

where the γ_k and τ_k represent the mean regression coefficient and precision parameter of predictor k. The prior distributions of γ_k and τ_k are

$$\gamma_k \sim N(\mu_0, \tau_0) \tag{8}$$

$$\tau_k \sim \Gamma(a_0, b_0), \tag{9}$$

where Γ denotes the gamma distribution. In our analyses, we selected uninformative priors with $\tau_0 = 0.01$, $a_0 = 2$, and $b_0 = 0.05$.

Parameter restrictions

We consider three types of parameter restrictions for γ_k. The restriction $\gamma_m = \gamma_n$ can be implemented by choosing an appropriate design matrix $\mathbf{X}_j$. Note that this approach also enforces $b_{jm} = b_{jn}$, $\forall j$, which means that effect sizes pertaining to

predictors m and are n assumed to originate from the same distribution. We implement the restriction $\gamma_m = 0$ by defining the prior of b_{jm} as $N(\gamma_m = 0, \tau_m)$. Finally, the restriction $\gamma_m > \gamma_n$ is equivalent to $\gamma_m = \gamma_n + a$, where $a > 0$. Consequently, the prior of b_{jm} is $N(\gamma_m = \gamma_n + a, \tau_m)$. For the auxiliary parameter a, we define the prior

$$a \sim N(0,1)_{a>0}, \tag{10}$$

where Eq. 10 denotes a truncated normal distribution such that $a > 0$.

Parameter estimation

Parameter estimation for Eqs. 6–10 can be accomplished with several iterations of the MCMC procedure (Gelman & Hill, 2006). Each iteration includes the following steps:

Step 1: Draw the b_{jk} parameters with prior $N(\gamma_k, \tau_k)$ from their full conditionals. Although this prior distribution assumes independent b_{jk} parameters, the dependent effect sizes require that the b_{jk} parameters are sampled from a multivariate normal distribution. For Step 1, we use the generalized least-squares estimators for the vector of regression coefficients $\mathbf{b}_j$ and their precision matrix Λ_j in study j given as

$$\hat{\mathbf{b}}_j = \left(\mathbf{X}_j' \Sigma_{ej}^{-1} \mathbf{X}_j\right)^{-1} \mathbf{X}_j' \Sigma_{ej}^{-1} \mathbf{y}_j \tag{11}$$

and

$$\hat{\Lambda}_j = \mathbf{X}_j' \Sigma_{ej}^{-1} \mathbf{X}_j \tag{12}$$

(Raudenbush et al., 1988). Note that $(\mathbf{X}_j' \Sigma_{ej}^{-1} \mathbf{X}_j)$ must be invertible to estimate $\hat{\mathbf{b}}_j$. This condition is violated, for example, if a study contains more predictors than effect sizes. Further, let γ be the vector of L effect sizes γ_k and Λ_γ a diagonal matrix of their corresponding precisions τ_k. Then the vector of regression coefficients $\mathbf{b}_j$ is drawn from its full conditional with

$$\mathbf{b}_j \Big| \hat{\mathbf{b}}_j, \gamma, \hat{\Lambda}_j, \Lambda_\gamma \sim N\Bigg(\left(\Lambda_\gamma + \mathbf{N}_j \circ \hat{\Lambda}_j\right)^{-1} \left(\Lambda_\gamma \gamma + \left(\mathbf{N}_j \circ \hat{\Lambda}_j\right) \hat{\mathbf{b}}_j\right), \left(\Lambda_\gamma + \mathbf{N}_j \circ \hat{\Lambda}_j\right)^{-1} \Bigg),$$

where $\mathbf{N}_j$ is a $L \times L$ matrix whose entry in row m and column n is $(n_m\, n_n)^{-0.5}$ with n_m being the number of effect sizes contributing to the estimate of b_{jm} and n_n being the number of effect sizes contributing to the estimate of b_{jn}. Further, $\mathbf{N}_j \circ \hat{\Lambda}_j$ denotes the entrywise product of $\mathbf{N}_j$ and $\hat{\Lambda}_j$.

Step 2: In Step 2, draw the γ_k parameters with prior $N(0, \tau_0)$ from their full conditionals with

$$\gamma_k \Big| b_{jk}, \tau_0 \sim N\left(\left(\tau_b \sum\nolimits_{j=1}^{M} b_{jk}\right) (\tau_0 + M\tau_b)^{-1}, (\tau_0 + M\tau_b)^{-1} \right), \tag{14}$$

where τ_b is the observed precision of the b_{jk} parameters.

Step 3: Draw the τ_k parameters with prior $\Gamma(a_0, b_0)$ from their full conditionals with

$$\tau_k \Big| b_{jk}, a_0, b_0 \sim \Gamma\left(a_0 + \frac{M}{2}, b_0 + \frac{\sum\nolimits_{j=1}^{M} \left(b_{jk} - \overline{b}_k\right)^2}{2} \right), \tag{15}$$

where $\overline{b}_k$ is the observed mean of the b_{jk} parameters across all $j = 1, \ldots, M$.

Step 4: For every parameter restriction which requires that $\gamma_m = \gamma_n + a$, an auxiliary parameter a with prior $N(0, 1)_{a>0}$, must be drawn from its full conditional distribution. Ideally, the estimate of a would be based on M difference scores of b_{jm} and b_{jn}, $j = 1,\ldots,M$. In practice, however, study j will not always contribute both parameters b_{jm} and b_{jn}, for example, because a specific predictor was not included in the study. To deal with this issue, we propose the following solution. Let $\overline{b}_m$ and $\overline{b}_n$ be the means of the b parameters of predictor m and n, respectively. Then we propose drawing a from the conditional distribution

$$a \Big| b_{jm}, b_{jn} \sim N\left(\left(M_{min}\tau_d \left[\overline{b}_m - \overline{b}_n\right]\right) (1 + M_{min}\tau_d)^{-1}, (1 + M_{min}\tau_d)^{-1} \right)_{a>0}, \tag{16}$$

where $\tau_d = \left(\sigma^2_{b_{jm}} + \sigma^2_{b_{jn}}\right)^{-1}$ and M_{min} is the smaller of the two sample sizes of b_{jm} and b_{jn}.

The MCMC procedure repeats Steps 1–4 in each iteration of the algorithm. The MCMC procedure starts with random parameter values, and the conditional distributions in Eqs. 13–16 need several iterations to reach their stationary distribution—the posterior distribution of the model parameters given the data. By repeatedly sampling from the posterior distribution, point estimates and credible intervals of the model parameters can be obtained.

Model selection

We perform model selection as part of the MCMC procedure using the approach suggested by Carlin and Chib (1995). Let X_1 and X_2 denote the design matrices of two

competing models $M = 1$ and $M = 2$, respectively. Further, let Eq. 6 define the likelihood function of the observed effect sizes $\mathbf{y}$ given the model parameters $\theta = \{b_{<j=1,\ldots,M><k=1,\ldots,L>}, \gamma_{<k=1,\ldots,L>}, \tau_{<k=1,\ldots,L>}\}$, denoted as $L(\mathbf{y}|\theta, X)$. Then the probability of $M = 1$ is given as

$$p_{M1} = \frac{L\left(\mathbf{y}\middle|\theta_1, X_1, M=1\right)\left\{\prod_{i=1}^{2} P\left(\theta_i\middle|M=1\right)\right\}\pi_1}{\sum_{j=1}^{2} L\left(\mathbf{y}\middle|\theta_j, X_j, M=j\right)\left\{\prod_{i=1}^{2} P\left(\theta_i\middle|M=j\right)\right\}\pi_j}, \tag{17}$$

where θ_i denotes the parameters of model i, $P(\theta_i|M = i)$ are the priors of the model parameters if $M = i$ and pseudo priors if $M \neq i$. Further, π_i is the prior probability of model i. In our analyses, we assume that both models are equally likely a priori—that is, $\pi_1 = \pi_2$=.5. Following Carlin and Chib's (1995) suggestion, we further set pseudo priors $P(\theta_i|M = j)$, $i \neq j$ to normal (or gamma in the case of τ_k) approximations of their posterior distributions observed in an initial run of the MCMC algorithm. The Bayes factor for $M = 1$ relative to $M = 2$ is then defined as $B_{12} = \overline{p}_{M=1}/(1-\overline{p}_{M=1})$, where $\overline{p}_{M=1}$ is the proportion of cases where M was sampled as $M = 1$ across all iterations of the MCMC algorithm.

Simulation study

To assess the validity of our model-selection approach, we performed a simulation study. We simulated effect sizes with the help of Eqs. 6–9. The effect sizes were sampled using Eqs. 6 and 7 with means $\gamma_1 = 0$ and $\gamma_2 = d$. The difference between γ_1 and γ_2 was simulated to be $d = \{0, 0.1, 0.3, 0.5\}$. In the condition $d = 0$, we generated four effect sizes per study and all studies originated from a single group with zero mean ($\gamma = 0$). For all other values of d, we generated two effect sizes per study for each group ($\gamma_1 = 0$, $\gamma_2 = d$). Thus, γ_1 and γ_2 contributed an equal number of effect sizes to all studies. We further varied the number of studies $M = \{5, 10\}$ in the simulation. We deliberately chose small values for M to evaluate the performance of the model selection when data were sparse. We then simulated 100 data sets for each combination of effect-size difference d and study sample size M. Finally, we analyzed each data set with three pairs of models. Each pair included the three types of parameter restrictions: The first pair included an equality restriction, the second pair included order restrictions, and the third pair included a zero-value restriction (i.e., $\gamma_2 = 0$). In all three pairs, the first model was at least approximately true when $d = 0$. Conversely, the second model was true when $d > 0$. Thus, d represents the effect size in a model comparison between Model 1 and Model 2. We estimated the model parameters for all data sets using 1,000 iterations of the MCMC procedure of which we discarded the first 500 to remove the influence of random starting values for the parameters.

Table 10 lists the mean Bayes factors of Model 2 compared to Model 1 for all simulation conditions. Ideally, when Model 1 and Model 2 are about equally suited to account for the data ($\gamma = 0$, first data column in Table 10), the more parsimonious model should be preferred by the model-selection procedure. This was the case for the equality restriction when Model 1 had considerably fewer parameters than Model 2. Conversely, the order restriction compared about equally complex and appropriate models ($\gamma_1 > \gamma_2$ vs. $\gamma_1 < \gamma_2$). Consequently, the Bayes factor was close to one, indicating no preference for either model. Finally, the zero-value restriction included a more parsimonious Model 1 (γ_1 free, $\gamma_2 = 0$) compared to Model 2 (γ_1 free, γ_2 free), while both models were about equally suited to fit the data. The resulting Bayes factor slightly favored the more parsimonious Model 1. However, with increasing values of γ_2, Model 2 is more adequate than Model 1, which was accurately reflected by the Bayes factor. The Bayes factor indicated more evidence for Model 2 compared to Model 1 as the effect size and the sample size increased. The simulation also revealed diverging statistical power for the three parameter restrictions, which was larger with equality compared to order and zero-value restrictions.

Table 10 Median Bayes factor of Model 2 relative to Model 1 as a function of the number of studies in the meta-analysis (M), the data-generating parameters (columns), and the models fitted to the data (rows)

		$M = 5$ studies			
Model 1	Model 2	$\gamma = 0$	$\gamma_1 = 0$ $\gamma_2 = .1$	$\gamma_1 = 0$ $\gamma_2 = .3$	$\gamma_1 = 0$ $\gamma_2 = .5$
$\gamma_1 = \gamma_2$	γ_1 free γ_2 free	0.09	6.81	>100	>100
$\gamma_1 > \gamma_2$	$\gamma_1 < \gamma_2$	1.07	1.92	8.93	36.09
γ_1 free $\gamma_2 = 0$	γ_1 free γ_2 free	0.63	0.80	6.14	70.43
		$M = 10$ studies			
Model 1	Model 2	$\gamma = 0$	$\gamma_1 = 0$ $\gamma_2 = .1$	$\gamma_1 = 0$ $\gamma_2 = .3$	$\gamma_1 = 0$ $\gamma_2 = .5$
$\gamma_1 = \gamma_2$	γ_1 free γ_2 free	0.06	55.25	>100	>100
$\gamma_1 > \gamma_2$	$\gamma_1 < \gamma_2$	1.04	2.14	24.66	76.38
γ_1 free $\gamma_2 = 0$	γ_1 free γ_2 free	0.56	0.75	15.40	>100

References

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

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13,* 219–235.

*Azimian-Faridani, N., & Wilding, E. L. (2004). An event-related potential study of the revelation effect. *Psychonomic Bulletin & Review, 11,* 926–931.

Aßfalg, A. (2017). Revelation effect. In R. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in judgment, thinking, and memory* (pp. 339–356). Abingdon, UK: Routledge.

*Aßfalg, A., & Bernstein, D. M. (2012). Puzzles produce strangers: A puzzling result for revelation-effect theories. *Journal of Memory and Language, 67,* 86–92.

*Aßfalg, A., & Nadarevic. L. (2015). A word of warning: Instructions and feedback cannot prevent the revelation effect. *Consciousness and Cognition, 34,* 75–86.

*Bernstein, D. M., Godfrey, R. D., Davison, A., & Loftus, E. F. (2004). Conditions affecting the revelation effect for autobiographical memory. *Memory & Cognition, 32,* 455–462.

*Bernstein, D. M., Rudd, M. E., Erdfelder, E., Godfrey, R., & Loftus, E. F. (2009). The revelation effect for autobiographical memory: A mixture-model analysis. *Psychonomic Bulletin & Review, 16,* 463–468.

*Bernstein, D. M., Whittlesea, B. W., & Loftus, E. F. (2002). Increasing confidence in remote autobiographical memory and general knowledge: Extensions of the revelation effect. *Memory & Cognition, 30,* 432–438.

Bjork, R. A. (1973). Why mathematical models? *American Psychologist, 28,* 426–433.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.

*Bornstein, B. H., & Neely, C. B. (2001). The revelation effect in frequency judgment. *Memory & Cognition, 29,* 209–213.

*Bornstein, B. H., Robicheaux, T. R., & Elliott, E. M. (2015). The role of semantic relatedness in producing the revelation effect: A test of the global-matching model. *Journal of Cognitive Psychology, 27,* 207–217.

*Bornstein, B., & Wilson, J. (2004). Extending the revelation effect to faces: Haven't we met before? *Memory, 12,* 140–146.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B,* 473–484.

*Cameron, T. E., & Hockley, W. E. (2000). The revelation effect for item and associative recognition: Familiarity versus recollection. *Memory & Cognition, 28,* 176–183.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Criss, A. H., & Howard, M. W. (2015). Models of episodic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 165–183). Oxford, UK: Oxford University Press.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478.

*Dewhurst, S. A., & Knott, L. M. (2010). Investigating the encoding—retrieval match in recognition memory: Effects of experimental design, specificity, and retention interval. *Memory & Cognition, 38,* 1101–1109.

*Dougal, S., & Schooler, J. W. (2007). Discovery misattribution: When solving is confused with remembering. *Journal of Experimental Psychology: General, 136,* 577.

*Frigo, L. C., Reas, D. L., & LeCompte, D. C. (1999). Revelation without presentation: Counterfeit study list yields robust revelation effect. *Memory & Cognition, 27,* 339–343.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16,* 309–313.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition, 13,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

*Guttentag, R., & Dunn, J. (2003). Judgments of remembering: The revelation effect in children and adults. *Journal of Experimental Child Psychology, 86,* 153–167.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16,* 107–112.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6,* 107–128.

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1105–1120.

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology, 41,* 109–139.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Hillsdale, NJ: Erlbaum.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.

Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition, 22,* 713–722.

*Hockley, W. E., & Niewiadomski, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition, 29,* 1176–1184.

Hodgson, O. R. (n.d.). *On the generality of the revelation effect in young and older adults.* Unpublished manuscript.

Huber, D. E., Clark, T. F., Curran, T., & Winkielman, P. (2008). Effects of repetition priming on recognition memory: Testing a perceptual fluency-disfluency model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1305–1324.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118,* 126–135.

*Kronlund, A., & Bernstein, D. M. (2006). Unscrambling words increases brand name recognition and preference. *Applied Cognitive Psychology, 20,* 681–687.

Kunst-Wilson, W., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*, 557–558.

*Landau, J. D. (2001). Altering the balance of recollection and familiarity influences the revelation effect. *The American Journal of Psychology, 114*(3), 425–437.

*LeCompte, D. C. (1995). Recollective experience in the revelation effect: Separating the contributions of recollection and familiarity. *Memory & Cognition, 23,* 324–334.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

*Leynes, P. A., Landau, J., Walker, J., & Addante, R. J. (2005). Event-related potential evidence for multiple causes of the revelation effect. *Consciousness and Cognition, 14,* 327–350.

Luo, C. R. (1993). Enhanced feeling of recognition: Effects of identifying and manipulating test items on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 405.

Macmillan, N. A., & Creelman, C. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Major, J. C., & Hockley, W. E. (2007). A test of two different revelation effects using forced-choice recognition. *Psychonomic Bulletin & Review, 14,* 1096–1100.

*Mulligan, N. W. (2007). The revelation effect: Moderating influences of encoding conditions and type of recognition test. *Psychonomic Bulletin & Review, 14,* 866–870.

*Mulligan, N. W., & Lozito, J. P. (2006). An asymmetry between memory encoding and retrieval revelation, generation, and transfer-appropriate processing. *Psychological Science, 17,* 7–11.

*Niewiadomski, M. W., & Hockley, W. E. (2001). Interrupting recognition memory: Tests of familiarity-based accounts of the revelation effect. *Memory & Cognition, 29,* 1130–1138.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology, 20,* 139–156.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*, 237-241.

*Peynircioglu, Z. F., & Tekcan, A. I. (1993). Revelation effect: Effort or priming does not create the sense of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 382–388.

Pohl, R. (2017). *Cognitive illusions: Intriguing phenomena in judgment, thinking, and memory*. Abingdon, UK: Routledge.

*Prull, M. W., Light, L. L., Collett, M. E., & Kennison, R. F. (1998). Age-related differences in memory illusions: Revelation effect. *Aging, Neuropsychology, and Cognition, 5,* 147–165.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103,* 111–120.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 305–320.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34,* 125–137.

*Thapar, A., & Sniezek, S. M. (2008). Aging and the revelation effect. *Psychology and Aging, 23,* 473–477.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26,* 1–12.

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 219–230.

*Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 739–746.

*Verde, M. F., & Rotello, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review, 11,* 560–566.

*Watkins, M. J., & Peynircioglu, Z. F. (1990). The revelation effect: When disguising test items induces recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1012–1020.

*Westerman, D. L. (2000). Recollection-based recognition eliminates the revelation effect in memory. *Memory & Cognition, 28,* 167–175.

*Westerman, D. L., & Greene, R. L. (1996). On the generality of the revelation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1147.

*Westerman, D. L., & Greene, R. L. (1998). The revelation that the revelation effect is not due to revelation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 377.

Whittlesea, B. W., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language, 29,* 716–732.

Whittlesea, B. W., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition, 29,* 234–246.

Whittlesea, B. W., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica, 98,* 141–165.

Whittlesea, B. W., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 547–565.

Whittlesea, B. W., & Williams, L. D. (2001). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 3–13.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517.

*Young, K. D., Peynircioglu, Z. F., & Hohman, T. J. (2009). Revelation effect in metamemory. *Psychonomic Bulletin & Review, 16,* 952–956.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9,* 1–27.