Routledge
Taylor & Francis Group

Memory

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

A novel paradigm to assess storage of sources in memory: the source recognition test with reinstatement

Nikoletta Symeonidou & Beatrice G. Kuhlmann

To cite this article: Nikoletta Symeonidou & Beatrice G. Kuhlmann (2021) A novel paradigm to assess storage of sources in memory: the source recognition test with reinstatement, Memory, 29:4, 507-523, DOI: 10.1080/09658211.2021.1910310

To link to this article: https://doi.org/10.1080/09658211.2021.1910310

Published online: 13 Apr 2021.

MEMORY
2021, VOL. 29, NO. 4, 507–523
https://doi.org/10.1080/09658211.2021.1910310

Routledge
Taylor & Francis Group

A novel paradigm to assess storage of sources in memory: the source recognition test with reinstatement*

Nikoletta Symeonidou and Beatrice G. Kuhlmann

Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany

ABSTRACT

The present research aimed to devise a test of source recognition that facilitates access to source information stored in memory. Therefore, we extended the standard source-monitoring paradigm, in which items are presented in a source-neutral manner during test, by a second, subsequent test with source reinstatement. In this second test, items (i.e., words) were presented with both study sources (i.e., two speakers) consecutively such that for originally studied words, one test presentation was the exact reinstatement of the original source. To validate our assumption that the test with reinstatement primarily assesses source storage, we manipulated source storage by varying encoding frequency between-participants (repetition vs. no repetition of each item-source-pair). Additionally, we varied source similarity between-participants (similar vs. dissimilar speakers). Data analyses ($N = 146$) based on multinomial and signal detection models showed a source memory enhancement in the second test with reinstatement compared to the first standard test, especially for similar sources. Additionally, repetition selectively benefited source memory in the second test, validating our interpretation of the second test as a measure for source storage. Altogether, our novel source recognition test offers a promising method for investigating various well-known source memory phenomena more comprehensively.

ARTICLE HISTORY
Received 15 September 2020
Accepted 24 March 2021

KEYWORDS
Source memory; source storage; source reinstatement; recognition; multinomial modelling

Who told me that sleep benefits memory? *Where* did I read that the computer game Tetris reduces intrusive thoughts? In various everyday life situations, we are required to remember original contextual (i.e., temporal, spatial, social, and emotional) features of an event or information in order to make inferences about its credibility or, more generally, to adequately function in our social environment. Memory for these contextual features of an event or information is referred to as *source memory* (Johnson et al., 1993; Lindsay, 1994; Mitchell & Johnson, 2009; Old & Naveh-Benjamin, 2008). Considering the importance of source memory in everyday life, it is crucial to understand the cause of source memory failures and to identify conditions under which source memory can be facilitated. Although there is a fairly large amount of studies investigating influences on source memory (for overviews, see Johnson et al., 1993; Mitchell & Johnson, 2009), the majority of studies manipulated source features during encoding (e.g., Bell & Buchner, 2010; Buchner et al., 2009; Conway & Dewhurst, 1995; Kuhlmann & Boywitt, 2016; Mather et al., 1999; May et al., 2005; Meiser & Sattler, 2007) and employed a source memory test with source-neutral presentation (e.g., only showing sources' names in the test, instead of presenting the voices of the sources, as done in the study phase, cf. Dodson & Shimamura, 2000), which is the standard test format in source-monitoring research (Johnson et al., 1993). In the current study, we aim to show that this standard source memory test often relies on successful *retrieval* of stored source information and thus potentially underestimates people's actual source memory level. Based on this reasoning, we propose an extension of the standard source memory paradigm by a newly devised test that facilitates access to source information stored in memory by manipulating source features at test.

Generally speaking, correctly recalling an information depends on both, its successful storage in memory and its successful retrieval from memory (Rouder & Batchelder, 1998; Bjork & Bjork, 1992). Consequently, any memory failure can result from difficulties in storage or in retrieval of to-be-remembered information (Batchelder & Riefer, 1986; Glisky et al., 2001; Riefer & Batchelder, 1995; Riefer

CONTACT Nikoletta Symeonidou nsymeoni@mail.uni-mannheim.de Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim D-68131, Germany

*Parts of this research were presented at the 2019 Tagung Experimentell Arbeitender Psychologen (TeaP Meeting of Experimental Psychologists) in London, United Kingdom. We do not have any financial or non-financial competing interests that might influence the results reported in this article.

& Rouder, 1992). Accordingly, failing to remember the source of an information can be due to poor storage of source features during encoding or rather due to problems in retrieving the stored source information from memory (for a specific discussion of storage and retrieval processes in source memory, see Glisky et al., 2001). Acknowledging this differentiation and identifying the prevailing mechanism of a memory failure is essential for deriving appropriate implications in theory (How does memory function?) and practice (How can we improve memory under specific circumstances?). For example, in the case of a retrieval failure, modifying the memory test can considerably foster access to the temporarily unavailable, however still stored information (Lindsay, 1994; McCloskey & Zaragoza, 1985), because different types of tests capture different memory processes. This can be straight-forwardly illustrated by comparing people's memory performance in a free-recall test versus a cued-recall and recognition test for previously encoded items. Whereas free recall heavily relies on successful *retrieval* of the stored items from memory, the latter two test formats, especially recognition, provide essential cues that facilitate retrieval and thus rather mirror the number of items actually *stored* in memory (Craik, 1986; Riefer & Rouder, 1992; Rouder & Batchelder, 1998). Consequently, memory performance in a free-recall test is typically lower than in a cued-recall and, especially, recognition test because retrieval failures impact the former more than the latter.

The idea that different memory tests tap into different memory processes (i.e., free recall: retrieval; cued recall/recognition: storage), has been systematically used by many researchers already decades ago and has since then advanced to an established method to estimate the contribution of retrieval and storage processes in *item* memory (e.g., Drachman & Leavitt, 1972; Hirshman et al., 1989; Hogan & Kintsch, 1971; Küpper-Tetzel & Erdfelder, 2012; Nadarevic, 2017; Riefer & Rouder, 1992; Schonfield & Robertson, 1966; Tulving & Pearlstone, 1966). For example, based on the finding that the bizarreness effect on item memory (i.e., better memory for bizarre than common sentences) occurs on free but not cued recall, Hirshman et al. and Riefer and Rouder concluded that there is an advantage in retrieval, but not storage, of bizarre items.

Given that such techniques have been fairly well evaluated, refined and advanced for investigating *item memory,* it is all the more surprising, that there is only little comparable research for source memory (Dodson & Shimamura, 2000; Starns & Hicks, 2013). In the standard laboratory source-monitoring task, participants study items presented in one of two sources (e.g., words spoken by one of two voices). Crucially, at test, the items are presented in a *source-neutral* manner (e.g., printed instead of spoken or spoken by a novel voice) and participants have to judge if the item was previously presented in one of the two sources (and, if so, indicate which one) or if the item is new. Given this source-neutral presentation at test, participants have to actively retrieve the stored source information from memory to make the source attribution, rendering this standard test largely dependent on source *retrieval*. Thus, even if source information is stored in memory, participants may make a false source attribution in this test due to difficulties in retrieving this stored source information.

To tackle this issue, the main goal of our current study was to adapt the standard source-monitoring test in such a way that it facilitates source retrieval and thus provides a better measure for actual source storage. In other words, we wanted to devise a test of source *recognition*.

Using source reinstatement to facilitate source retrieval

How can we minimise retrieval demands in a source memory test and thus ensure that the test primarily captures memory for *stored* source information? We suggest that source retrieval can be facilitated by reinstating or re-establishing the original source at test. This idea grounds on the encoding specificity principle (Smith & Vela, 2001; Tulving & Thomson, 1973), which states that retrieval of information from memory is most successful if retrieval conditions match encoding conditions. Much research on context reinstatement has focused on its facilitation of item memory (see Smith & Vela, 2001, for a meta-analysis). Only few studies have additionally investigated the effects of context reinstatement on source memory for such context features (Craik & Kirsner, 1974; Dodson & Shimamura, 2000; Kelley et al., 1989; Kirsner, 1974; Leynes et al., 2003; Naveh-Benjamin & Craik, 1995; Palmeri et al., 1993; Starns & Hicks, 2005, 2013; Vogt & Bröder, 2007). For instance, Dodson and Shimamura (2000) used a source memory test, in which participants were presented with words that had been previously heard in a male or in a female voice (or were new). Crucially, in this test the words were either presented by the same voice (match trials) or the respective other voice (mismatch trial). Over a series of four experiments, they found that presenting words by the matching source (i.e., reinstating the original source) resulted in substantially better source memory compared to source memory for words presented by the mismatching source at test. Similar facilitating effects of source reinstatement on source memory were reported in a few other studies (Leynes et al., 2003; Starns & Hicks, 2005, 2013; Vogt & Bröder, 2007; but see Craik & Kirsner, 1974; Kelley et al., 1989; Naveh-Benjamin & Craik, 1995).

Notably, most previous studies (with the exception of Starns & Hicks, 2013) on reinstatement in source memory used the just described match-mismatch test procedure, meaning that items were either presented by the correct (i.e., matching) or by the incorrect (i.e., mismatching) source at test. While this procedure can establish a reinstatement benefit in source memory, it has several important shortcomings that render the interpretation of the

causes of the observed reinstatement benefit difficult: For one, it might encourage a strong response bias towards the source that presents the item at test, independently of whether it is correct or not. This may result in an overestimation of source memory in the matching and an underestimation in the mismatching trials and thus artificially inflate the found reinstatement benefit (see Dodson & Shimamura, 2000; Vogt & Bröder, 2007 for a discussion of this issue). Furthermore, the reinstatement benefit is estimated in comparison to an orthogonal mismatch of the study source, which likely underestimates normal source memory, rather than in comparison to a source-neutral control test condition. Notably, Dodson and Shimamura (2000) included such source-neutral control conditions in two of their experiments by presenting some items in a novel voice or no voice (only printed on the screen) at test. However, like the match and mismatch condition these control conditions were manipulated between items so that each item is only tested in one of these conditions. Any such manipulation of test conditions between items prevents truly assessing source *recognition*, that is, whether people are able to discriminate the correct study source from the incorrect source for a given item. Further, a between-item manipulation of reinstatement versus control test conditions prevents assessing whether for the same given item for which the source is not retrievable in the standard test (control condition) it becomes accessible with reinstatement.

Notably, Starns and Hicks (2013) used a quite different approach to investigate reinstatement effects in source memory, which circumvents some of the above-mentioned limitations of the match-mismatch procedure. Specifically, they used a source memory test, in which half of the words were presented with both the correct (matching) and incorrect (mismatching) source (i.e., male or female face picture) simultaneously. That is, the test item was presented in both sources side-by-side on the screen. This manipulation prevents any guessing bias towards only one specific reinstated source. For the other half of tests words, neither source was presented at test, allowing a neutral (rather than mismatching) control condition (i.e., standard source memory test). They observed better source memory when the sources were re-presented during test but only when internal source reinstatement was difficult (i.e., when there were many different source faces that could potentially be reinstated). The fact that there was a reinstatement benefit on source memory even when both sources were reinstated side-by-side suggests that people can discriminate between the correct and incorrect source (i.e., have source recognition). However, because of the simultaneous side-by-side source presentation at test, the test presentations did not completely match the study presentations in these experiments, leading to a potential underestimation of any reinstatement effects. Further, the between-item manipulation of control versus reinstatement presentation at tests continues to prevent assessing whether for a given item for which the source is inaccessible in the control (standard) test, the source becomes accessible with source reinstatement.

A test of source recognition

To tackle the outlined issues of both approaches, we propose an alternative source recognition test, which employs full source reinstatement. That is, instead of a side-by-side source presentation, we suggest to present each test item with both study sources consecutively such that for originally studied items, one test presentation is the exact reinstatement of the original source (match) whereas the other is not (mismatch). Put differently, this test perfectly mimics the study conditions, ensuring that both sources are fully reinstated. Thus, participants merely have to recognise (rather than retrieve) the correct one out of the two source options. Therefore, akin to a typical recognition test for item memory, our proposed test for source recognition should primarily assess source storage in memory because it provides more specific retrieval cues via reinstatement and thus reduces reliance on retrieval processes.

Moreover, unlike the match-mismatch paradigm, our approach does not induce a response bias towards the matching or mismatching source: By keeping the order of both source presentations constant (i.e., test items are always presented in Source A first and then Source B or *vice versa*), both the first and second presentation is equally likely to be correct. Further, just like standard source-monitoring data, the responses from this source recognition test can be analysed with a multinomial model (e.g., the two-high-threshold multinomial model of source monitoring [2HTSM], Bayen et al., 1996) or a signal detection model (Banks, 2000; DeCarlo, 2003) to correct for any guessing bias towards one of the sources. Crucially, for source recognition data, the memory parameter estimates from these models can be interpreted as source *storage* specifically, whereas the memory parameter estimates from the standard test confound storage and retrieval. Finally, we propose to combine the standard source memory test and the proposed novel test of source recognition to a new two-test paradigm to allow estimating the contribution of retrieval processes in source memory.

A novel two-test paradigm to assess source storage and retrieval

Following previous research contrasting free recall with cued recall and recognition to infer about storage versus retrieval (Batchelder & Riefer, 1986; Danckert & Craik, 2013; Küpper-Tetzel & Erdfelder, 2012; Riefer et al., 2002; Riefer & Batchelder, 1995; Tulving & Pearlstone, 1966), we propose a similar two-test procedure with the standard source memory test followed by our novel source

recognition test. This allows us to not only measure source storage via the source recognition test but also to assess the role of retrieval processes by comparing source memory performance between both tests: As elaborated earlier, in contrast to our novel source recognition test, source memory in the standard test (no reinstatement of source features) additionally depends on self-initiated, successful source retrieval, rather than source storage only. Hence, by comparing source memory between the first, retrieval-demanding standard test and the second, retrieval-facilitating source recognition test, we can estimate the overall reinstatement benefit. Crucially, this reinstatement benefit reflects the proportion of items for which the source was stored in memory, and thus recognised on the second test, but was not retrievable in the standard test. Thus, it is indicative of source retrieval processes. Note that we conducted both tests in direct succession (with a negligibly small delay between Test 1 and Test 2) to prevent the occurrence of any testing effects (Roediger & Karpicke, 2006; see Discussion section for a detailed discussion of this issue).

The current study

In our experiment, we aimed at evaluating the outlined, novel two-test procedure. In particular, we intended to validate that the second test with source reinstatement predominantly assesses source *storage*. To this end, we employed a manipulation of source storage by manipulating encoding frequency between participants. More specifically, auditory recordings of words were either played once at a two-seconds-presentation rate per word (no repetition) or twice resulting in a four-seconds-presentation rate per word (repetition). Based on previous literature showing a beneficial effect of repeated encoding on memory in general (Hockley & Cristi, 1996; Naveh-Benjamin et al., 2000), and on multinomial model-based estimates of encoding/storage processes in particular (Riefer et al., 2002), we expected repeated encoding to result in better source storage. If the second test is indeed more storage-dependent than the first test, repetition should primarily affect source memory in this second test.

In addition, we manipulated the similarity of the two sources (high vs. low; cf. Bayen et al., 1996; Bayen & Murnane, 1996) between participants to explore if and how this influences the reinstatement effect on source memory. Starns and Hicks (2013) suggested that participants spontaneously engage in internal source reinstatement on a standard source memory test and do so successfully if internal reinstatement of the sources is easy. Following this logic, we assumed that participants might be less successful in internally reinstating two highly similar sources (or even reluctant to do so at all), since similar sources – as in the case of numerous sources (cf., Starns & Hicks, 2013) – should be rather difficult to internally reinstate. Accordingly, we expected the reinstatement benefit to be more pronounced for highly similar sources due to less successful (self-initiated) internal reinstatement compared to dissimilar sources.

Taken together, we expected that firstly, source memory would be higher in the second compared to the first test due to retrieval facilitation (i.e., reinstatement effect). Secondly, we expected that encoding frequency primarily influences source memory in the second test, which we deemed to be a test of source *recognition* and thus primarily storage-dependent. Thirdly, following Starns and Hicks (2013), we predicted that reinstatement benefits source memory most when sources are difficult to internally reinstate (here due to high source similarity).

Method

Design and participants

The design was a 2 (source) × 2 (test) × 2 (encoding frequency) × 2 (source similarity) mixed factorial. Source and test were manipulated within-participants such that participants studied items (words) which were presented with one of two sources (spoken by speakers additionally represented with a picture and name) and were first tested with a standard source-monitoring test (Johnson et al., 1993) with the items presented visually (written on the screen) followed by a second test with source reinstatement (i.e., each item presented with both sources). Between-participants we manipulated source storage by repeating each item-source pair at study for half of the participants. In addition, source similarity was manipulated between-participants such that for half of the participants the two sources were very similar (i.e., both male) whereas for the remaining half they were dissimilar (i.e., different gender; cf. Bayen et al., 1996). The full crossing of these two between-subjects factors resulted in four experimental conditions (no-repetition & dissimilar-sources; repetition & dissimilar-sources; no-repetition & similar-sources; repetition & similar-sources).

Based on previous studies assessing source monitoring processes with the 2HTSM (Bayen et al., 1996, 2000; Kuhlmann et al., 2016) we aimed for at least 24 participants per condition (i.e., 96 participants in total). However, the precision of multinomial model parameters further back in the model, such as source memory in the 2HTSM, depends on the level of preceding parameters (e.g., item memory). Because item recognition was somewhat low, especially in the second test (discussed later), estimation precision of source memory was not satisfactory after recruiting 96 eligible participants at the University of Mannheim. Thus, we extended recruitment for another week at nearby Heidelberg University. Precisely, we tracked the precision of estimation of the source memory parameter *d* via its 95% CI and aimed for a CI width below .20. Ultimately, we recruited 155 students at both universities via leaflets and the participant recruitment systems of the respective Psychology departments. Nine participants were excluded from data analysis, either because they

did not meet pre-defined eligibility requirements (German as native language [i.e., learned before the age of six]; aged 18–30 years) or because their demographic data pertaining to eligibility was missing. Thus, final analyses were based on data from 146 participants (119 women; M_{age} = 21.58 years, SD_{age} = 2.72 years), which were approximately equally distributed across the four experimental conditions (36 participants in the no-repetition & dissimilar-sources and in the repetition & similar-sources condition, 37 in each of the two remaining conditions). With at least 3,780 observations (i.e., 36 participants × 105 items) per condition, the power to detect even small effect sizes of $w = .10$ in the goodness-of-fit test of the MPT model was very high (i.e., 1-$\beta > .99$; power analysis based on G*Power 3.1.9.7; Faul et al., 2007).

Materials

Word stimuli and recordings were taken from (Meiser & Sattler, 2007; see also Kuhlmann & Boywitt, 2016). The pool consisted of 167 concrete German nouns (4–7 letters) recorded in two male and two female voices. For our study, we utilised one of the female and one of the male voice recordings. We additionally recorded a second male voice for the similar-sources condition.[1] For each participant anew, 109 words were randomly drawn from the pool to be used in the experiment. Out of these, 74 words served as target items (i.e., were presented in the study phase). The first four were primacy items (equally split between the two sources) that were not tested and thus not included in the final data analyses. The remaining 35 words served as distractors (i.e., new words) in the memory tests. Words were randomly assigned to serve as targets (half to each of the two sources) versus distractors for each participant anew.

Depending on the experimental condition, either the female and one male voice (dissimilar-sources conditions) or both male voices (similar-sources conditions) were used for the source manipulation. We additionally linked each voice to a common German name (i.e., "Jakob" and "Susanne" as male and female for the dissimilar-sources conditions; "Jakob" and "Johan" for the similar-sources conditions) and to a face picture taken from Bayen et al. (1996, Figure 2, "high similarity" and "low similarity" pictures for the similar and dissimilar sources, respectively). The face of the respective source was presented centred on the screen with the name placed right below while the recording of the study item was played. Thus, across all participants, each source consisted of a fixed voice-name-face-combination. The source (dis)similarity manipulation affected all three source features (similarity of the voice, name, and face).

Procedure

All participants were tested in groups up to six people in laboratory rooms with separated computer cubicles. Participants first provided written informed consent and were then randomly assigned to one of the four experimental conditions. The experiment was administered via the software OpenSesame (Mathôt et al., 2012). Audio was presented via headphones, which participants wore throughout the entire task. It started with a volume-regulation procedure: Participants were presented with the word *Auto* (German for *car*), spoken by both source voices consecutively (order counterbalanced between participants, contingent on the order of source presentation in Test 2) at a medium starting value. Participants then could adapt the volume to their personal preference.

Then, the study phase followed. Participants were instructed to memorise the words and their sources for a later memory test. Seventy (+ 4 primacy) items were spoken by one of the two source voices one at a time in a pseudorandom order, which constrained item presentation to a maximum of three words spoken by the same source in direct succession. In total, half of the target words were presented by one source, and the other half by the other source. Each auditory item was accompanied by the corresponding name and face of the source, which was shown in the middle of the screen. Depending on the condition, the two sources (voice + name + picture) consisted of either one female and one male speaker (dissimilar conditions) or two male speakers (similar conditions). In the no-repetition conditions, each word recording was played once together with the picture and name of the respective source, which were presented simultaneously on the screen for 2000 ms. In the repetition conditions, each item was spoken twice in immediate succession by the same speaker, and the picture and name of the respective speaker stayed on the screen for 4000 ms. Each study trial was initiated by a fixation circle that persisted for 500 ms.

After the study phase, participants performed a filler task (i.e., verifying simple mathematical equations by pressing the keys "c" and "m" for correct and false equations, respectively) for three minutes before turning to the test phase.

In the test phase, participants first performed a standard source-monitoring test (Test 1), where all 70 target words from the study phase (35 per source) plus 35 new words, randomly selected from the word pool to serve as distractors (i.e., 105 words in total), were presented consecutively on the top centre of the screen in a randomised order. Below, the names of the two sources were presented to the left and right (counterbalanced across participants) and the option "new" was presented at the centre bottom. Participants' task was to decide self-paced for each word, if it was previously spoken by Source A, Source B, or not presented at all during the study phase (i.e., was new) by pressing the corresponding key ("d" or "k" for the source presented on the left or right side of the screen, respectively; space key for new items).

Directly afterwards participants were asked whether they had used internal source reinstatement as a retrieval

strategy in this first test. If they affirmed, they were additionally asked to indicate their frequency of using internal reinstatement, that is, for how many test words (in percent) they used this strategy. They were also asked to describe any other retrieval strategies they may have used.[2] The rationale behind these questions was to explore whether participants indeed spontaneously engage in internal source reinstatement (or other retrieval strategies), as proposed but not explicitly measured by Starns and Hicks (2013), and whether retrieval-strategy use differs dependent on source similarity (i.e., less internal reinstatement when the sources are highly similar).

Subsequently, the second retrieval-facilitating source-monitoring test followed (Test 2). All target and distractor items from the previous test were presented again in a newly randomised order. This time, however, each item was first presented aurally by both speakers in direct succession (e.g., first spoken by Jakob, then by Susanne), visually accompanied by the respective faces and names of the sources (i.e., all three source features were reinstated). The presentation screen in this second test were set up exactly like the study screen, with a fixation circle (500 ms) initiating each trial and a presentation duration of 2000 ms for each source. Thus, for target ("old") items, one presentation was an exact reinstatement of the original study context. Following both presentations, participants were then asked to decide again if the word was previously spoken by Source A, Source B, or was not presented during the study phase (i.e., was new) by pressing the corresponding key. This test screen and the response key assignment were identical to the first test for each participant. To prevent that participants missed an item due to inattention or distraction, they were allowed to rehear the test item, again spoken by both sources consecutively (by pressing the key "w"), before making their decision. The presentation order of the two sources stayed constant throughout this test but was counterbalanced across participants (contingent on their screen position in Test 1). This precluded any dependencies between the presentation order and the correctness of the source. Importantly, we carefully instructed participants to not simply repeat their answers from the first test but rather to remember for each word anew by which source it had been presented or if it was a new word – even if this resulted in a response different from their previous one. Furthermore, we emphasised that the distractor words still were to be classified as new, although they had already appeared in the first test. We encouraged participants to view the second test as a separate memory test, independent from the first one.

For exploratory reasons, participants were then asked to indicate if they used any strategies to better memorise the words and their speakers in the study phase on a paper-pencil questionnaire. Specifically, we used an adapted version of (Kuhlmann & Touron, 2012) strategy questionnaire that assesses the frequency of having used imagery, sentence generation, clustering, and rote repetition separately for words and for words and their sources, respectively. Finally, participants provided demographic information, were debriefed and compensated for their participation, either by course credit or by payment.

Results

The alpha level was set to .05 for all analyses. To obtain memory measures that are independent from response bias, we applied the 2HTSM model (Bayen et al., 1996) to our data and estimated source memory via the model's parameter *d*. The 2HTSM model belongs to the model family of multinomial processing tree (MPT) models (see Erdfelder et al., 2009 for an overview) and allows disentangling the contribution of memory processes versus guessing biases to the observed test responses, providing a "purer" measure for item and source memory. Therefore, the 2HTSM is favoured over other commonly used measures of item and source memory (e.g., hit & false alarm rates, averaged conditional source identification measures/ ACSIMs; cf. Murnane & Bayen, 1996) and has gained considerable popularity in source memory research over the last decades (Bayen & Kuhlmann, 2011; Bell et al., 2017; Bröder & Meiser, 2007; Erdfelder et al., 2009; Vogt & Bröder, 2007). Additionally, to corroborate that our results hold independently of the analysis method, we used the bivariate signal detection model (SDT model; Banks, 2000; DeCarlo, 2003) as well as empirical measures (corrected hit rates and ACSIMs) to analyse the data at hand. The results of these additional analyses are reported in the Appendix (Appendix A for SDT and Appendix B for the empirical measures). To anticipate, the results converge for all three methods of analysis.

Model description and fit

Figure 1 depicts Submodel 4 of the 2HTSM (Bayen et al., 1996) which contains four parameters that represent the probabilities of different cognitive processes that underlie participants' responses in a source memory test: The probability of item memory (i.e., recognising an item or detecting that a distractor is new) is measured by parameter *D*. If a word is recognised in the test phase, the original source may also be correctly recalled with probability *d*. If the source cannot be recalled (i.e., 1-*d*), guessing processes take place. Specifically, parameter *g* measures the probability to guess that an item was presented by Source A (Source B is guessed with the complementary probability 1-*g*). Because source memory is conditional on item recognition, source memory for unrecognised items cannot emerge (Bell et al., 2017; Malejka & Bröder, 2016). Thus, if item recognition fails (i.e., 1-*D*), participants' answers are solely based on guessing processes: With probability *b*, participants guess that a word was previously presented in the study phase (i.e., is "old"), followed by guessing that the word was spoken by either Source A (*g*) or

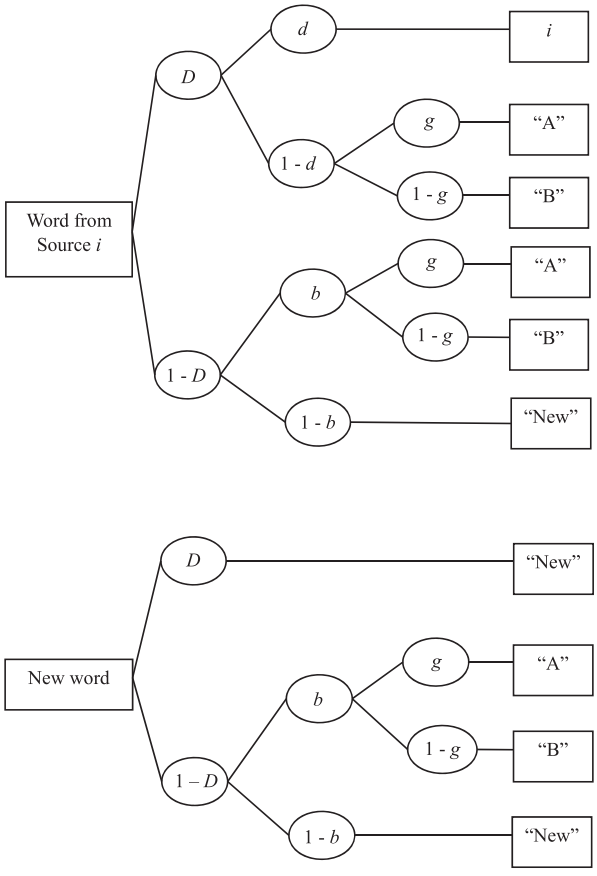

Figure 1. Graphical representation of the two-high-threshold multinomial model of source monitoring (2HTSM).

Note: The figure shows Submodel 4 of the 2HTSM for target words (upper tree) and for new words (lower tree). *i* denotes words from Source *i*, $i \in \{A, B\}$. Source A = Jakob, Source B = Susanne/ Johan, dependent on the experimental condition. Boxes on the right represent participants' answers in the source memory test. *D* = probability of detecting a word as previously presented or not presented; *d* = probability of correctly recalling the source (speaker) of a recognised word; *b* = probability of guessing that a word was previously presented; *g* = probability of guessing that a detected or undetected word was spoken by Source A (i.e., speaker "Jakob"). Adapted from "Source discrimination, item detection, and multinomial models of source monitoring", by Bayen et al. (1996, p. 202).

Source B (1-*g*). With the complementary probability 1-*b*, participants guess that the item was new.

Note that the depicted Submodel 4 is the most parsimonious submodel of the 2HTSM as it implements the strict, but often applicable, assumptions that neither item nor source memory differ between sources and that source guessing does not differ between recognised and unrecognised items. To assess fit of this submodel to our data and estimate its parameters, we used the software multiTree (Moshagen, 2010). Based on the aggregated observed response frequencies in both source monitoring tests (cf., Table C1 in Appendix C), we evaluated fit of this submodel and estimated its parameters via maximum likelihood (ML) estimation methods as implemented in the software *multiTree* (Moshagen, 2010). Note that because we administered two memory tests, we aggregated frequencies of all distinguishable response outcomes across participants for each test independently (and separately for each experimental condition).[3] Overall, the model fit the data well, $G^2(16) = 17.63$, $p = .346$ (across all four conditions and both tests). Parameter estimates are listed in Table 1.

Item memory

As apparent from Table 1, the item memory parameter *D* of the 2HTSM was numerically higher in the repetition compared to the non-repetition conditions (for each test type and level of source similarity), higher for the dissimilar compared to the similar conditions (for each test type and level of repetition), and higher for the first, standard test compared to the second test with reinstatement (for each level of repetition and source similarity), respectively. To test whether these numerical differences were substantial, we set the *D* parameters of the respective conditions equal. For all equality restrictions, the model fit declined significantly, indicating that item memory was better in the repetition compared to the no-repetition conditions, $\Delta G^2\text{s}(1) \geq 19.42$, $p < .001$, and worse for similar sources compared to dissimilar ones, $\Delta G^2\text{s}(1) \geq 12.31$, $p < .001$. In addition, item memory declined from the first to the second test in all four between-subjects conditions, $\Delta G^2\text{s}(1) \geq 7.58$, $p < .006$. This was presumably due to repetition of the distractor items from the first test in our second test, which increased their familiarity and thus made them

Table 1. Parameter estimates of the two-high-threshold multinomial model of source monitoring (2HTSM).

		Parameter estimates				
Test	Condition	*D*	*d*	*b*	*g*	*a*
Test 1 (standard)	no-repetition & dissimilar-sources	.48 [.45; .50]	.82 [.75; .89]	.46 [.44; .50]	.50 [.46; .53]	.95 [.83; 1.08]
	repetition & dissimilar-sources	.57 [.55; .60]	.82 [.77; .88]	.49 [.46; .52]	.51 [.48; .55]	.85 [.77; .94]
	no-repetition & similar-sources	.40 [.37; .43]	.42 [.34; .50]	.50 [.48; .53]	.48 [.46; .51]	.65 [.48; .82]
	repetition & similar-sources	.50 [.47; .53]	.48 [.42; .55]	.46 [.43; .49]	.49 [.46; .52]	.56 [.46; .66]
Test 2 (with reinstatement)	no-repetition & dissimilar-sources	.42 [.38; .45]	.86 [.77; .94]	.51 [.48; .53]	.54 [.51; .57]	
	repetition & dissimilar-sources	.51 [.48; .54]	.97 [.89; 1.04]	.58 [.55; .61]	.52 [.49; .56]	
	no-repetition & similar-sources	.32 [.29; .35]	.65 [.54; .76]	.55 [.53; .58]	.47 [.44; .49]	
	repetition & similar-sources	.42 [.39; .45]	.86 [.77; .95]	.60 [.57; .62]	.53 [.50; .55]	

Note: Brackets indicate 95% confidence intervals. Test 1 is the standard source memory tests (high retrieval demands), Test 2 is the novel source recognition test with source reinstatement. *D* = probability of detecting a word as previously presented or not presented; *d* = probability of correctly recalling the source of a recognised word; *b* = probability of guessing that a word was previously presented; *g* = probability of guessing that a detected or undetected word was spoken by Source A (i.e., speaker "Jakob"); *a* = proportional change in *d* from the first to the second test (quantifies the size of the reinstatement effect: the higher *a* the lower the reinstatement effect). Overall, the model fit the data well, $G^2(16) = 17.63$, $p = .346$ (across all four conditions and both tests).

more difficult to discriminate from the actual target (i.e., previously studied) words.

Source memory

In line with our first prediction and in contrast to the effects observed in item memory, source memory (parameter *d*) was numerically higher in the second test (with reinstatement) compared to the first (standard) test in all four conditions (see Figure 2), and pronouncedly so in the similar-sources conditions. To test for the significance of this difference, we set parameter *d* of Test 1 (d_{Test_1}) equal to parameter *d* of Test 2 (d_{Test_2}) within each experimental condition. This equality restriction led to worse model fit, indicating that source memory in the second test was significantly higher than in the first test, for both conditions with similar sources, $\Delta G^2(1) = 10.86$, $p < .001$, and $\Delta G^2(1) = 46.73$, $p < .001$ without and with repetition, respectively. For dissimilar sources, the small numerical difference between tests was significant in the repetition condition only, $\Delta G^2(1) = 9.32$, $p = .002$, but not in the condition without repetition, $\Delta G^2(1) = 0.46$, $p = .496$. Other analyses of source memory presented in Appendix A and B replicated the reinstatement effect for similar sources but not for repeated dissimilar sources (and confirmed the null effect for dissimilar non-repeated sources). Taken together, we found a robust reinstatement effect on memory for similar sources with our novel experimental paradigm comparing against a neutral standard source memory test.

Regarding our second hypothesis, we predicted that repeated encoding should more strongly influence source memory in the primarily storage-dependent second test rather than in the more retrieval-dependent first test. Thus, we tested whether source memory was substantially larger in the repetition compared to the no-repetition conditions separately for each test and level of source similarity by setting the respective source memory parameters equal across both encoding conditions (e.g., $d_{Test\_1\text{–no-repetition \& similar}} = d_{Test\_1\text{–repetition \& similar}}$). In line with our prediction, this equality constraint had no impact on model fit in the first test, $\Delta G^2(1) = .02$, $p = .881$ for the dissimilar-sources, and $\Delta G^2(1) = 1.29$, $p = .256$ for the similar-sources condition (see also Figure 2). In contrast, and as expected, it significantly worsened model fit in the second test for the similar-sources condition, $\Delta G^2(1) = 8.22$, $p = .004$. For the dissimilar-sources condition, source memory in Test 2 was numerically higher in the repetition condition compared to the no-repetition condition (cf. Table 1), however the comparison just missed significance, $\Delta G^2(1) = 3.48$, $p = .062$. Taken together, these findings corroborate our assumption that source memory in the second test with reinstatement more clearly reflects source storage – and is thus influenced by a manipulation of encoding frequency – compared to source memory in the first standard test.

Finally, we predicted that source similarity might influence the reinstatement effect, that is, the degree of the increase in source memory from the first to the second test. In particular, we expected the reinstatement effect to be weaker for highly discriminable dissimilar sources, presumably because they are easier to internally reinstate. To test this interaction of test and source similarity, we imposed parametric order constraints (Knapp & Batchelder, 2004; Kuhlmann et al., 2019) to measure the proportional difference between source memory in Test 1 relative to Test 2, allowing for a direct comparison of this difference (i.e., the reinstatement effect) between the source similarity conditions. In concrete terms, we imposed the order constraint that source memory in the first test is worse than in the second test in all four conditions (i.e., $d_{Test\_1} < d_{Test\_2}$) and reparametrized the model such that source memory in Test 1 is modelled as a proportion of source memory in Test 2 (i.e., $d_{Test\_1} = a * d_{Test\_2}$). The novel parameter *a* represents the proportional change in source memory between the first and the second test and thus inversely quantifies the size of the reinstatement effect. The higher *a* is, the lower the difference between d_{Test_1} and d_{Test_2}, and the smaller the reinstatement effect. The estimated *a* parameters for each condition are reported in Table 1.[4]

To test the influence of source similarity on the reinstatement effect in source memory, we equated parameter *a*

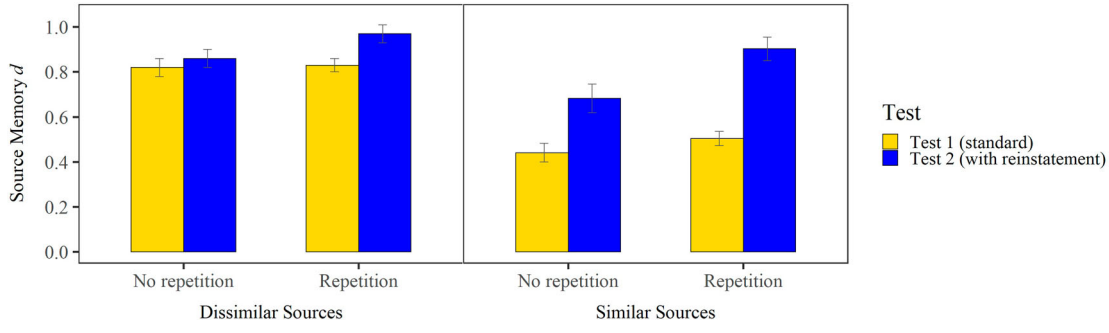

Figure 2. Probability of source memory on both tests in the four experimental conditions.

Note: Probability of source memory is measured by parameter *d* of the two-high-threshold multinomial model of source monitoring (2HTSM; Bayen et al., 1996). Error bars indicate one standard error of the estimate. Test 1 is the standard source memory tests (high retrieval demands), Test 2 is the novel source recognition test with source reinstatement.

across the dissimilar and similar conditions, separately for the no-repetition and repetition conditions, respectively (e.g., $a_{\text{no-repetition\& dissimilar}} = a_{\text{no-repetition \& similar}}$). For both levels of encoding frequency, this led to a significantly worse model fit, $\Delta G^2(1) = 6.78$, $p = .009$ for the no-repetition conditions, $\Delta G^2(1) = 18.25$, $p < .001$ for the repetition conditions. As evident in Figure 2 and in the *a* estimates provided in Table 1, the reinstatement effect was indeed more pronounced (i.e., lower *a*) in the similar compared to the dissimilar conditions. Thus, as expected, the reinstatement effect was significantly stronger when the sources were highly similar and thus difficult to discriminate.

Reported internal reinstatement use

To further explore whether difficulty of internal reinstatement (due to low source distinctiveness) not only reduced participants' ability but also (or even primarily) their willingness (i.e., likelihood) to engage in internal reinstatement, we compared the number of participants who indicated that they had engaged in internal source reinstatement in the first test between the dissimilar- and similar-sources conditions (we combined the two encoding-frequency groups of the same level of source similarity to one group). The analysis yielded that participants in the similar-sources conditions reported being as likely to engage in internal source reinstatement (58.90%) as participants in the dissimilar-sources conditions (60.27%), $\chi^2(1, N = 73) = 0.33$, $p = .566$, for the no-repetition and $\chi^2(1, N = 73) = 0.66$, $p = .417$, for the repetition conditions, combined $\chi^2(1, N = 146) = 0.28$, $p = .866$. Thus, the greater reinstatement benefit in the similar-sources compared to dissimilar-sources conditions seems to ground on participants inability (rather than lacking willingness) to engage in effective internal source reinstatement when made difficult, which supports Starns and Hicks' (2013) proposal.

Discussion

The main purpose of this research was to devise a primarily storage-dependent test of source memory. More specifically, our goal was to investigate source *recognition* – that is, how well people can discriminate between an item reinstated in its original source from the presentation of that item in a different source. Compared to a standard source memory test in which test items were presented in a source-neutral manner (and thus, requiring source retrieval), our novel source recognition test with reinstatement indeed revealed substantially higher source memory, implying that more source information was stored in memory than the standard test suggested. Further bolstering our interpretation of our source recognition test with reinstatement as primarily dependent on source storage in memory, we found that an encoding manipulation only affected source memory in this test but not in the standard test. Collectively, our results support the idea that source reinstatement at test can be employed to measure primarily storage-dependent source recognition.

Advantages of the novel source recognition test and paradigm

The consistent finding that the encoding frequency manipulation selectively affected source memory (independent of the memory model underlying its measurement) in the source recognition test with source reinstatement but not in the standard source memory test, speaks for the validity of our newly developed test as a measure for source storage. This novel test opens up the possibility to investigate several, well-known source memory phenomena, such as the age-related decline in source memory (Brown et al., 1995; Johnson et al., 1993; Kuhlmann & Boywitt, 2016; Old & Naveh-Benjamin, 2008), more comprehensively in terms of underlying processes. Crucially, our source recognition test with reinstatement has some notable advantages over the previously used match-mismatch approach (i.e., presenting each item either with their original, matching or another, mismatching source; Craik & Kirsner, 1974): It more directly assesses participants ability to recognise (i.e., discriminate) the matching from the mismatching source and it does not induce a response bias towards one specific source, because all items are presented with both sources consecutively. Admittedly, our procedure does not preclude participants from developing a response bias towards the first (or second) presented source in Test 2. Additional analyses reported in Appendix E however, showed that there was no guessing bias towards the source presented first versus second (see *g* parameters of Test 2 in Table E1). This analysis also showed that the reinstatement effect did not depend on the source order. That is, reinstatement effects were comparably large independent of whether the correct source was reinstated first versus second in all four conditions (see *a* parameter in Table E1).

While of course the novel source recognition test with reinstatement can be used on its own, we see several advantages in combining it with a standard source monitoring test as done here. Adding to the source recognition test's above-discussed advantages over the match-mismatch procedure, the first, standard source memory test serves as a neutral control condition. This comparison allows for an assessment of the difficulty of source retrieval. That is, the more source recognition exceeds source memory in the standard test, the more difficult source retrieval is. In the MPT modelling approach, the parametric order constraints allow to directly quantify this retrieval component via the proportional change in source memory from the first, standard test to the second, retrieval-facilitating test. Hence, combining our proposed paradigm with a multinomial modelling approach allows us to not only derive a measure for source storage (i.e., *d* in Test 2) but also to estimate the contribution of retrieval processes (change parameter *a*) in source memory.

An at first glance uncommon feature of our paradigm may be the two-test procedure, which naturally raises the question if participants' memory in the source recognition test with reinstatement was biased due to this test always occurring second. In other words, critics might argue that performance in the second test is artificially inflated due to re-testing. However, this is highly unlikely, because both tests in our paradigm were administered in direct succession. Thus, the source memory enhancement from the first to the second test cannot be explained in terms of a testing effect, which typically occurs at a delay of two (or more days) between the two tests (Roediger & Karpicke, 2006; Rowland, 2014). Indeed, we observed no testing effect on item memory which in contrast declined on the second test.

Furthermore, previous studies on item memory showed that administering a free-recall test does not influence performance on an immediately following cued-recall test (Marevic & Rummel, 2020; Riefer & Rouder, 1992) or recognition test (Darley & Murdock, 1971; Wenger et al., 1980), respectively. Notably, this only holds when the first test is not biased towards retrieval-practicing a subset of items in a systematic way which causes systematic forgetting of the non-practiced items (Anderson et al., 1994; Hicks & Starns, 2004). Thus, it is an important aspect of our paradigm that all items are tested in the first standard test, not a subset. Finally, it must be noted that the opposite order, that is, first recognition then free recall, might inflate performance in the second test, suggesting that testing primarily fosters retrieval (i.e., accessibility) rather than storage (i.e., availability; see Rowland, 2014 for a detailed discussion). In short, retesting memory within a short time interval inflates memory only if the first test is biased towards some items and/ or is easier (i.e., less retrieval-dependent) than the second test, which is both not the case in our paradigm.

Considering these findings, we carefully decided on the implemented order of the two tests – first source recall then source recognition – and against counterbalancing of test order to preclude that the source-recognition test biases performance in the standard source memory test. Further crucial, there are several patters in our data that clearly contradict the notion that memory for the same material is always better when re-tested. For example, the observed decline in item memory from Test 1 to Test 2 (see Table 1) and the lack of a consistent reinstatement effect for dissimilar sources (in line with our predictions based on Starns & Hicks, 2013) proves that memory is not necessarily improved on the second test. A general test order effect also cannot account for the differential sensitivity of the source memory parameters in Test 2 to our encoding manipulations (e.g., more pronounced source memory benefit due to repetition in Test 2 compared to Test 1). This interaction pattern, however, fits perfectly to our notion, that the retrieval cues provided in Test 2 via source reinstatement make this test less retrieval- and more storage-dependent and hence more sensitive to storage-manipulations.

That source memory performance, like other memory performance, crucially depends on the type of test and is not generally better or worse on a second test, was also shown by Dodson and Shimamura (2000). The authors employed two source-monitoring tests in two of their experiments. Although their first test included the match-mismatch manipulation strongly affecting source memory, memory on the second standard test was comparable for all items independent of their previous test condition. Compared to the first test, source memory on the second test was thus both better (mismatch items) but also worse (match items). Taken together, we deem it unlikely that the first, standard source memory test artificially boosted performance in the second, source recognition test in our proposed paradigm.

Having said that, there is one noteworthy shortcoming associated with our novel test and paradigm: In all four conditions, item memory substantially decreased from the first to the second test. While this again speaks against a general memory improvement effect of a second test, it unfortunately renders the source memory estimation, which depends on the number of recognised items, less reliable in the second test (and the same applies for empirical measures of source memory conditionalizing on item recognition like ACSIM; Murnane & Bayen, 1996). This reduction in item recognition in the second test with reinstatement also seems to contradict the encoding specificity principle and previous findings of beneficial effects of context reinstatement on item recognition (see Smith & Vela, 2001, for a meta-analysis). Although these previous studies did not present the same item in multiple contexts but either in a matching, mismatching or a novel context, we do not believe that context reinstatement benefits to item recognition are confined to tests with only one context per item. Rather, we believe that these detrimental effects on item recognition result from the repetition of the distractor items from the first test in our second test (see also Dodson & Shimamura, 2000, Experiment 4). Due to the repetition of the distractors, their level of familiarity increases (cf., Jennings & Jacoby, 1997). Thus, participants cannot longer rely on familiarity in their item recognition on Test 2 and are instead taxed with making another source recollection even for item recognition, namely deciding whether an item is familiar because it was shown in the study phase or because it appeared in the previous test. As, however, we were primarily interested in source memory differences between both tests and our sample was large enough to ensure reliable *d* parameter estimation in both tests (see confidence intervals for *d* in Table 1), even with the lower item recognition (*D*) level on Test 2, this drawback does not limit the validity of our main results. One possible solution to this issue could be to only include those items in the second test that participants classified as "old" in the first test (i.e., hits and false alarms). However, this would have restricted item-presentation in Test 2 to only "old"-items from Test 1, rendering the second test a pure

source attribution test (no "new" option) and thus preventing the use of established measures of source memory (Murnane & Bayen, 1996; Bröder & Meiser, 2007). Another possibility would be to test a (random) half of the items with the standard source monitoring test and the other half with the source recognition test (including reinstatement) and to randomise (or counterbalance) the order in which both tests occur. This would, however, require doubling (or at least substantially increasing) the number of to-be-learned items or the number of participants as the model parameter estimation for each type of test would rely only on half of the overall available items.

Moderators of the reinstatement effect in source memory

Our study provides additional evidence that the reinstatement effect is stronger for similar compared to dissimilar sources. Although the analyses based on the 2HTSM suggested a reinstatement effect for repeated dissimilar sources, this effect was rather small and did not replicate in the SDT-based (see Appendix A) or ACSIM-based (see Appendix B) analyses. Thus, overall, there was no robust evidence for a reinstatement effect for dissimilar sources whereas this effect was robustly obtained for similar sources across all analyses. Taking up Starns and Hicks' (2013) reasoning, one possible explanation for this result pattern is that participants are less able to engage in effective internal source reinstatement in the similar-sources compared to the dissimilar-sources conditions. With less effective internal reinstatement of similar sources, source memory profits more from an external reinstatement. Interestingly, our additional exploratory analysis based on participants' self-reports indicated, that participants in the similar-sources conditions were comparably inclined to engage in internal source reinstatement as participants in the dissimilar-sources conditions. This further supports Starns and Hicks' assumption that low source discriminability (due to high similarity in our study vs. due to high numerosity in Starns & Hicks, 2013) primarily reduces the success or effectiveness (rather than the mere likelihood) of internal reinstatement. Note however, that our question on the use of internal reinstatement may have been phrased in such a way that people were inclined to agree with it to make a good impression. Thus, the frequency reports of both conditions may be overestimated. A possible solution to reduce such demand characteristics in future studies might be to ask participants first about their general retrieval strategies before specifically asking about internal reinstatement. The challenge in this is to ensure participants' understanding of what a retrieval strategy actually is and to provide a clear distinction to encoding strategies, as, in our experience, participants tend to confuse both (i.e., most answers to our open question on retrieval strategies described common encoding strategies). Altogether, our results corroborate the findings of Starns and Hicks (2013) that participants spontaneously engage in internal source reinstatement and that externally provided source cues are particularly helpful when internal reinstatement is difficult and thus less effective due to low source distinctiveness.

Another potential reason for the higher reinstatement effects in the similar conditions might be that reinstatement reactivates both, item-to-source and source-to-source associations, and this is particularly helpful in the similar-sources conditions. More specifically, in order to make a correct source judgment in the similar-sources conditions, participants do not only have to remember the pairing of the item to a specific source feature (e.g., specific voice), but also to bind the individual source features to each other (e.g., male voice belongs to Jakob versus Johan). In contrast, in the dissimilar-sources conditions, it is enough to remember only one source feature and simply infer the remaining features (e.g., female voice must belong to Susanne not Jakob). Note that both outlined explanations (internal reinstatement versus role of source-to-source binding) are not opposed to each other, but rather can complement each other: Internal reinstatement in the similar-sources conditions might be less efficient because participants (additionally) fail to reinstate the source-to-source associations. In any case, our goal was to manipulate difficulty of source discrimination, which we successfully achieved through varying source similarity. Future research should explore the mechanisms behind these difficulty differences.

Finally, the size of the reinstatement effect should also be more pronounced the more the first standard source memory test actually depends on source retrieval. More specifically, there is a variation in how certain source features are typically tested in the standard test and this influences demands on source retrieval. For certain sources, such as faces or pictures, those faces (or pictures) are typically shown again in the source test but not in the same set-up with the item as in the study phase (i.e., partial reinstatement), providing important retrieval cues and thus making source retrieval easier. In contrast, for other sources no such partial reinstatement is provided at test, putting more burden on source retrieval. For example, for voices as sources (without an accompanying picture of the speaker) only the names of the speakers can be displayed but the voices cannot be (partially) reinstated during the standard test. Future studies could systematically investigate how the size of the reinstatement effect might differ dependent on the source feature and the chosen methodological paradigm.

Conclusion

The two-test procedure of our novel paradigm effectively makes use of source reinstatement as a retrieval-facilitating technique and thus provides a better measure for actual source *storage* in memory. The overall considerable size of the reinstatement effect we observed in source

memory points to the importance of retrieval processes in source memory, at least when sources are similar and thus difficult to discriminate: The use of the standard paradigm (source-neutral presentation at test, i.e., no reinstatement) leads to a substantial underestimation of actually stored source information, which in turn might result in erroneous conclusions or false (practical) recommendations. For example, studies have repeatedly shown that source memory in the standard test is impaired in some populations (e.g., older compared to younger adults, Brubaker & Naveh-Benjamin, 2014; Glisky et al., 2001; Naveh-Benjamin, 2000; Naveh-Benjamin et al., 2007; people with depression or schizophrenia compared to healthy controls, Mitchell & Johnson, 2009). So far, however, it is not entirely clear, whether this impairment is actually due to retrieval problems only or reflects issues with source storage. Thus, by using our novel paradigm of source recognition, future studies could investigate if the described source memory impairments are minimised or even levelled in Test 2. Naturally, dependent on the prevailing process, different recommendations are sensible (e.g., using associative strategies during study in the case of a storage problem, cf. Glisky et al., 2001; Kuhlmann & Touron, 2012, 2017; using internal reinstatement as a strategy in the case of retrieval problems, Starns & Hicks, 2013).

To conclude, our proposed extension to the standard source memory paradigm provides a fruitful basis for new studies in various domains of source memory research. Its use would significantly contribute to a more thorough understanding of the cognitive processes that underlie different source memory phenomena, as in the case of applied clinical and aging research as well as basic research on episodic memory.

Notes

1. One of the male voices of the original recordings was that of a current psychology professor at the University of Mannheim. To avoid bias due to familiarity with this voice for some of our participants, we recorded the 167 words by a male student assistant from another university and used these recordings alongside with the other (unfamiliar) male recordings for the similar-sources condition.
2. Because participants predominantly described encoding instead of retrieval strategies when openly asked about their use of retrieval strategies other than internal reinstatement, we did not include these descriptions in our analysis.
3. This complete pooling approach assumes parameter homogeneity across items and participants. To ensure that our parameter estimates were not biased due to possible participant homogeneity, we additionally used a Bayesian-hierarchical approach (latent-trait approach [Klauer, 2010], as implemented in the R package TreeBUGS [Heck et al., 2018]) to derive individual and group-level parameter estimates based on Markov-chain Monte Carlo (MCMC) sampling. Table D1 in Appendix D shows that the estimated group-mean parameters from this latent-trait estimation were very similar to the estimates based on the aggregated data, indicating no systematic bias due to participant heterogeneity. For relatively homogeneous samples, as in the case of our student sample, estimating parameters based on aggregated frequencies has several advantages over averaging individual parameter estimates (cf. Chechile, 2009). It results in more precise (group-level) parameter estimates, reducing the risk of underestimating between-group differences, and more generally, it enables the use of inferential statistics (G^2) to derive clear-cut answers to the research questions at hand. We thus report the analyses based on Maximum-Likelihood estimation from the aggregated data as our main results.
4. The reparametrized model is of the same dimensionality (i.e., has the same degrees of freedom) and thus yields the same model fit and identical parameter estimates for all other parameters as the original model.

Acknowledgments

We thank Katja Bitz, Michelle Dörnte, Jule Schilling, Paula Schmelzer, Liliane Wulff and Selina Zaidler for their help with participant recruitment and data collection. We thank Jan Rummel for the opportunity to recruit participants in his lab and his student assistants for their help in collecting data at Heidelberg University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – KU 3329/1-1 and GRK 2277 "Statistical Modeling in Psychology".

Data availability statement

The data and datasets that support the findings of this study are openly available in the OSF repository at https://osf.io/6nzjs/.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, *11*(4), 267–273. https://doi.org/10.1111/1467-9280.00254

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*(2), 129–149. https://doi.org/10.1111/j.2044-8317.1986.tb00852.x

Bayen, U. J., & Kuhlmann, B. G. (2011). Influences of source–item contingency and schematic knowledge on source monitoring: Tests of the probability-matching account. *Journal of Memory and Language*, *64*(1), 1–17. https://doi.org/10.1016/j.jml.2010.09.001

Bayen, U. J., & Murnane, K. (1996). Aging and the use of perceptual and temporal information in source memory tasks. *Psychology and Aging*, *11*(2), 293–303. https://doi.org/10.1037/0882-7974.11.2.293

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 197–215. https://doi.org/10.1037/0278-7393.22.1.197

Bayen, U. J., Nakamura, G. V., Dupuis, S. E., & Yang, C.-L. (2000). The use of schematic knowledge about sources in source monitoring. *Memory & Cognition*, *28*(3), 480–500. https://doi.org/10.3758/BF03198562

Bell, R., & Buchner, A. (2010). Valence modulates source memory for faces. *Memory & Cognition*, *38*(1), 29–41. https://doi.org/10.3758/MC.38.1.29

Bell, R., Mieth, L., & Buchner, A. (2017). Emotional memory: No source memory without old-new recognition. *Emotion*, *17*(1), 120–130. https://doi.org/10.1037/emo0000211

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Erlbaum.

Bröder, A., & Meiser, T. (2007). Measuring source memory. *Zeitschrift Für Psychologie / Journal of Psychology*, *215*(1), 52–60. https://doi.org/10.1027/0044-3409.215.1.52

Brown, A. S., Jones, E. M., & Davis, T. L. (1995). Age differences in conversational source monitoring. *Psychology and Aging*, *10*(1), 111–122. https://doi.org/10.1037/0882-7974.10.1.111

Brubaker, M. S., & Naveh-Benjamin, M. (2014). The effects of presentation rate and retention interval on memory for items and associations in younger adults: A simulation of older adults' associative memory deficit. *Aging, Neuropsychology, and Cognition*, *21*(1), 1–26. https://doi.org/10.1080/13825585.2013.772558

Buchner, A., Bell, R., Mehl, B., & Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evolution and Human Behavior*, *30*(3), 212–224. https://doi.org/10.1016/j.evolhumbehav.2009.01.004

Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, *53*(6), 562–576. https://doi.org/10.1016/j.jmp.2009.06.005

Conway, M. A., & Dewhurst, S. A. (1995). Remembering, familiarity, and source monitoring. *The Quarterly Journal of Experimental Psychology Section A*, *48*(1), 125–140. https://doi.org/10.1080/14640749508401380

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 409–421). Amsterdam: North-Holland.

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284. https://doi.org/10.1080/14640747408400413

Danckert, S. L., & Craik, F. I. M. (2013). Does aging affect recall more than recognition memory? *Psychology and Aging*, *28*(4), 902–909. https://doi.org/10.1037/a0033263

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*(1), 66–73. https://doi.org/10.1037/h0031836

DeCarlo, L. T. (2003). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology*, *47*(3), 292–303. https://doi.org/10.1016/S0022-2496(03)00005-1

Dodson, C. S., & Shimamura, A. P. (2000). Differential effects of cue dependency on item and source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 1023–1044. https://doi.org/10.1037/0278-7393.26.4.1023

Drachman, D. A., & Leavitt, J. (1972). Memory impairment in the aged: Storage versus retrieval deficit. *Journal of Experimental Psychology*, *93*(2), 302–308. https://doi.org/10.1037/h0032489

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Glisky, E. L., Rubin, S. R., & Davidson, P. S. R. (2001). Source memory in older adults: An encoding or retrieval problem? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1131–1146. https://doi.org/10.1037/0278-7393.27.5.1131

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). Treebugs: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review*, *11*(1), 125–130. https://doi.org/10.3758/BF03206471

Hirshman, E., Whelley, M. M., & Palij, M. (1989). An investigation of paradoxical memory effects. *Journal of Memory and Language*, *28*(5), 594–609. https://doi.org/10.1016/0749-596X(89)90015-6

Hockley, W. E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition*, *24*(2), 202–216. https://doi.org/10.3758/BF03200881

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562–567. https://doi.org/10.1016/S0022-5371(71)80029-4

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging*, *12*(2), 352–361. https://doi.org/10.1037/0882-7974.12.2.352

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kelley, C. M., Jacoby, L. L., & Hollingshead, A. (1989). Direct versus indirect tests of memory for source: Judgments of modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1101–1108. https://doi.org/10.1037/0278-7393.15.6.1101

Kirsner, K. (1974). Modality differences in recognition memory for words and their attributes. *Journal of Experimental Psychology*, *102*(4), 579–584. https://doi.org/10.1037/h0036112

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*(4), 215–229. https://doi.org/10.1016/j.jmp.2004.03.002

Kuhlmann, B. G., Bayen, U. J., Meuser, K., & Kornadt, A. E. (2016). The impact of age stereotypes on source monitoring in younger and older adults. *Psychology and Aging*, *31*(8), 875–889. https://doi.org/10.1037/pag0000140

Kuhlmann, B. G., & Boywitt, C. D. (2016). Aging, source memory, and the experience of "remembering". *Aging, Neuropsychology, and Cognition*, *23*(4), 477–498. https://doi.org/10.1080/13825585.2015.1120270

Kuhlmann, B. G., Erdfelder, E., & Moshagen, M. (2019). Testing interactions in multinomial processing tree models. *Frontiers in Psychology*, *10*, 2364. https://doi.org/10.3389/fpsyg.2019.02364

Kuhlmann, B. G., & Touron, D. R. (2012). Mediator-based encoding strategies in source monitoring in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1352–1364. https://doi.org/10.1037/a0027863

Kuhlmann, B. G., & Touron, D. R. (2017). Relate it! Objective and subjective evaluation of mediator-based strategies for improving source memory in younger and older adults. *Cortex*, *91*, 25–39. https://doi.org/10.1016/j.cortex.2016.11.015

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*(1), 37–47. https://doi.org/10.1080/09658211.2011.631550

Leynes, P. A., Bink, M. L., Marsh, R. L., Allen, J. D., & May, J. C. (2003). Test modality affects source monitoring and event-related potentials. *The American Journal of Psychology*, *116*(3), 389. https://doi.org/10.2307/1423500

Lindsay, D. S. (1994). Memory source monitoring and eyewitness testimony. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony* (pp. 27–55). Cambridge University Press. https://doi.org/10.1017/CBO9780511759192.003

Malejka, S., & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition*, *44* (1), 63–72. https://doi.org/10.3758/s13421-015-0549-8

Marevic, I., & Rummel, J. (2020). Retrieval-mediated directed forgetting in the item-method paradigm: The effect of semantic cues. *Psychological Research*, *84*(3), 685–705. https://doi.org/10.1007/s00426-018-1085-5

Mather, M., Johnson, M. K., & de Leonardis, D. M. (1999). Stereotype reliance in source monitoring: Age differences and neuropsychological test correlates. *Cognitive Neuropsychology*, *16*(3-5), 437–458. https://doi.org/10.1080/026432999380870

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

May, C. P., Rahhal, T., Berry, E. M., & Leighton, E. A. (2005). Aging, source memory, and emotion. *Psychology and Aging*, *20*(4), 571–578. https://doi.org/10.1037/0882-7974.20.4.571

McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, *114*(1), 1–16. https://doi.org/10.1037/0096-3445.114.1.1

Meiser, T., & Sattler, C. (2007). Boundaries of the relation between conscious recollection and source memory for perceptual details. *Consciousness and Cognition*, *16*(1), 189–210. https://doi.org/10.1016/j.concog.2006.04.003

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin*, *135*(4), 638–677. https://doi.org/10.1037/a0015849

Moshagen, M. (2010). Multitree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Murnane, K., & Bayen, U. J. (1996). An evaluation of empirical measures of source identification. *Memory & Cognition*, *24*(4), 417–428. https://doi.org/10.3758/BF03200931

Nadarevic, L. (2017). Emotionally enhanced memory for negatively arousing words: Storage or retrieval advantage? *Cognition and Emotion*, *31*(8), 1557–1570. https://doi.org/10.1080/02699931.2016.1242477

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187. https://doi.org/10.1037/0278-7393.26.5.1170

Naveh-Benjamin, M., Brav, T. K., & Levy, O. (2007). The associative memory deficit of older adults: The role of strategy utilization. *Psychology and Aging*, *22*(1), 202–208. https://doi.org/10.1037/0882-7974.22.1.202

Naveh-Benjamin, M., & Craik, F. I. M. (1995). Memory for context and its use in item memory: Comparisons of younger and older persons. *Psychology and Aging*, *10*(2), 284–293. https://doi.org/10.1037/0882-7974.10.2.284

Naveh-Benjamin, M., Craik, F. I. M., Gavrilescu, D., & Anderson, N. D. (2000). Asymmetry between encoding and retrieval processes: Evidence from divided attention and a calibration analysis. *Memory & Cognition*, *28*(6), 965–976. https://doi.org/10.3758/BF03209344

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging*, *23*(1), 104–118. https://doi.org/10.1037/0882-7974.23.1.104

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328. https://doi.org/10.1037/0278-7393.19.2.309

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–1203. https://doi.org/10.1037/a0033044

Riefer, D. M., & Batchelder, W. H. (1995). A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition*, *23*(5), 611–630. https://doi.org/10.3758/BF03197263

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*(2), 184–201. https://doi.org/10.1037/1040-3590.14.2.184

Riefer, D. M., & Rouder, J. N. (1992). A multinomial modeling analysis of the mnemonic benefits of bizarre imagery. *Memory & Cognition*, *20* (6), 601–611. https://doi.org/10.3758/BF03202710

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rouder, J. N., & Batchelder, W. H. (1998). Multinomial models for measuring storage and retrieval processes in paired associate learning. In C. Dowling, F. Roberts, & P. Theuns (Eds.), *Recent progress in mathematical psychology* (pp. 195–225). Erlbaum.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *20*(2), 228–236. https://doi.org/10.1037/h0082941

Schütz, J., & Bröder, A. (2011). Signal detection and threshold models of source memory. *Experimental Psychology*, *58*(4), 293–311. https://doi.org/10.1027/1618-3169/a000097

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203–220. https://doi.org/10.3758/BF03196157

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Starns, J. J., & Hicks, J. L. (2005). Source dimensions are retrieved independently in multidimensional monitoring tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1213–1220. https://doi.org/10.1037/0278-7393.31.6.1213

Starns, J. J., & Hicks, J. L. (2013). Internal reinstatement hides cuing effects in source memory tasks. *Memory & Cognition*, *41*(7), 953–966. https://doi.org/10.3758/s13421-013-0325-6

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381–391. https://doi.org/10.1016/S0022-5371(66)80048-8

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

Vogt, V., & Bröder, A. (2007). Independent retrieval of source dimensions: An extension of results by Starns and Hicks (2005) and a comment on the ACSIM measure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 443–450. https://doi.org/10.1037/0278-7393.33.2.443

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(2), 135–144. https://doi.org/10.1037/0278-7393.6.2.135

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characterstics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1415–1434. https://doi.org/10.1037/0278-7393.25.6.1415

Appendices

Appendix A. Results based on the bivariate signal detection model

Item memory

The crucial difference between the 2HTSM and the bivariate signal detection model (or threshold and signal detection models, in general) is that they rest upon different assumptions about the nature of memory (Schütz & Bröder, 2011; Stanislaw & Todorov, 1999; Starns & Hicks, 2013). In the 2HTSM, source and item memory represent discrete cognitive states that are reached as soon as a specific threshold is passed. By comparison, source and item memory in the signal detection model depend on continuous memory strength signals. Thus, in the 2HTSM, decisions are based on discrete states, whereas in the signal detection model, they are based on whether the accumulated continuous evidence (i.e., memory strength) surpasses a response/ decision criterion (for both, item- and source-decisions). There is an ongoing debate about the validity of these two outlined, opposing views on memory and their corresponding model families (Pazzaglia et al., 2013; Schütz & Bröder, 2011). In source monitoring research, this issue is further complicated as it has been proposed that item memory is continuous but source memory is a threshold process (Yonelinas, 1999). In our present study we did not aim to test the nature of item and source memory and for example did not assess confidence judgments which might allow further insights into continuous versus threshold processes. To ensure that our conclusions about the reinstatement effect on source memory do not depend on the specific multinomial threshold measure we used to assess source memory, we confirmed that our results replicate in SDT-based analyses.

We computed *d'* scores for item memory by subtracting the inverse-normal transform of the false alarm rate from the inverse-normal transform of the hit rate ($d'_{item} = z(HR) - z(FAR)$). The hit rate referred to the proportion of items assigned to one of the two sources (i.e., deemed "old") among all studied items and the false alarm rate referred to the proportion of items assigned to one of the two sources among all distractor items. For HR and FAR of 0 we applied the adjustment proposed by Hautus (1995); that is, a constant of .5 was added to the number of hits and false alarms and the number of detection and signal trials was increased by 1 (see also Stanislaw & Todorov, 1999). Means and standard deviations of d'_{item} for each experimental condition can be found in Table A1, separately for each test. We submitted d'_{item} to a 2 (test) × 2 (encoding frequency) × 2 (source similarity) mixed analysis of variance (ANOVA). The overall results indicated main effects of repetition, similarity and test type, and no interactions. In line with the results based on the 2HTSM, item memory measured in d'_{item} was better in the repetition compared to the no-repetition conditions, $F(1, 142) = 10.98$, $p = .001$, $\eta_p^2 = .07$, worse for similar sources compared to dissimilar ones, $F(1, 142) = 7.82$, $p = .006$, $\eta_p^2 = .05$, and worse in the second test with reinstatement compared to the first, standard source memory test, $F(1, 142) = 70.34$, $p < .001$, $\eta_p^2 = .33$. Thus, the item memory analyses based on the bivariate signal detection model fully replicate those obtained on the 2HTSM measure of item recognition.

Source memory

We computed *d'* scores for source memory by subtracting the inverse-normal transform of the false alarm rate from the inverse-normal transform of the hit rate ($d' = z(HR) - z(FAR)$; cf., DeCarlo, 2003; Starns & Hicks, 2013). The hit rate referred to the proportion of items assigned to source "Jakob" among all Jakob-items and the false alarm rate referred to the proportion of items assigned to source "Jakob" among all Susanne-items (or Johan-items for the similar-sources conditions). Because of cases with HR of 1 and FAR of 0 we applied the adjustment proposed by Hautus (1995). Means and standard deviations of the *d'* scores separately for each test and each experimental condition can be found in Table A1. We submitted the *d'* scores to a 2 (test) × 2 (encoding frequency) × 2 (source similarity) mixed analysis of variance (ANOVA). The overall results indicated main effects of repetition, similarity, and test type. As expected, source memory was better in the repetition compared to the no-repetition conditions, $F(1, 142) = 6.72$, $p = .011$, $\eta_p^2 = .05$, and worse for similar sources compared to dissimilar ones, $F(1, 142) = 27.49$, $p < .001$, $\eta_p^2 = .16$. Crucially, in line with the results based on the 2HTSM, source memory performance in the source recognition test (with source reinstatement) was higher compared to performance in the standard test, $F(1, 142) = 4.47$, $p = .036$, $\eta_p^2 = .03$. Further crucial, there was an interaction between repetition and test type, $F(1, 142) = 3.96$, $p = .049$, $\eta_p^2 = .03$, and between similarity and test type, $F(1, 142) = 7.99$, $p = .005$, $\eta_p^2 = .05$. Simple main effect analyses following up on the repetition × test type interaction, revealed that repeated encoding substantially benefitted source memory in the second, source recognition test, $F(1, 142) = 11.16$, $p = .001$, $\eta_p^2 = .07$, but not in the first, standard test, $F(1, 142) = 2.38$, $p = .125$. Thus, the SDT-based analyses further corroborate our assumption that the source recognition test more strongly relies on source storage (and hence is more affected by an encoding manipulation) compared to the standard source memory test. Simple main effect analyses following up on the similarity × test type interaction, revealed that the source memory increase from the standard test to the source recognition test (i.e., reinstatement effect) was only significant when the sources were similar and thus difficult to discriminate, $F(1, 142) = 12.20$, $p = .001$, $\eta_p^2 = .08$. In contrast, there was no significant reinstatement effect for dissimilar sources, $F < 1$.

At large, these results mirror the above reported results based on the 2HTSM estimate of source memory. The only exception is that in the pairwise comparisons conducted in the 2HTSM-based analyses, there was a significant (but small) reinstatement effect for dissimilar sources in the repetition condition. The SDT-based source memory measure did not replicate this small effect, $F < 1$ for the similarity × test type × encoding frequency three-way interaction and $p = .634$ for the pair-wise comparison in the repetition & dissimilar-sources condition, and rather supports the conclusion that there is no reinstatement effect on memory for easy to discriminate sources. Overall, the findings based on the bivariate signal detection model replicate the results from the 2HTSM, thus ensuring that the conclusion drawn for the research question at hand are independent of the underlying memory model.

Table A1. Means and standard deviations of *d'* based on the bivariate signal detection model.

Test	Condition	d'_{item}	d'_{source}
Test 1	no-repetition & dissimilar-sources	1.43 (.69)	1.54 (0.87)
(standard)	repetition & dissimilar-sources	1.81 (.89)	1.76 (1.05)
	no-repetition & similar-sources	1.16 (.55)	0.63 (0.79)
	repetition & similar-sources	1.49 (.71)	0.90 (1.05)
Test 2	no-repetition & dissimilar-sources	1.18 (.62)	1.40 (0.98)
(with reinstatement)	repetition & dissimilar-sources	1.56 (.80)	1.81 (0.99)
	no-repetition & similar-sources	.89 (.43)	0.78 (.69)
	repetition & similar-sources	1.23 (.65)	1.32 (.69)

Note: Standard deviation in brackets.

Appendix B. Results based on empirical measures

Corrected hit rates (item memory)

Corrected hit rates (i.e., PRs) for each participant were computed by subtracting their false alarm rates (i.e., proportion of distractor items assigned to one of the two sources) from their hit rates (i.e., proportion of studied items assigned to one of the two sources). Means and standard deviations of PR for each experimental condition can be found in Table B1, separately for each test. We submitted PRs to a 2 (test) × 2 (encoding frequency) × 2 (source similarity) mixed analysis of variance (ANOVA). The overall results indicated main effects of repetition, similarity and test type, and no interactions. Replicating the model-based analysis (2HTSM and bivariate SDT), item recognition (i.e., PRs) was better in the repetition compared to no-repetition conditions, $F(1, 142) = 9.39$, $p = .003$, $\eta_p^2 = .06$, worse for similar sources compared to dissimilar ones, $F(1, 142) = 6.53$, $p = .012$, $\eta_p^2 = .04$, and worse in the second test with reinstatement compared to the first, standard source memory test, $F(1, 142) = 58.92$, $p < .001$, $\eta_p^2 = .29$. Thus, results on corrected hit rates fully mirrored the model-based results on item memory.

Averaged conditional source identification measures (source memory)

For the averaged conditional source identification measures (ACSIMs) only studied items were considered. ACSIMs were computed by averaging the proportion of correct source attributions among hits across both sources. Means and standard deviations of the ACSIMs for each experimental condition can be found in Table A1, separately for each test. We submitted ACSIMs to a 2 (test) × 2 (encoding frequency) × 2 (source similarity) mixed analysis of variance (ANOVA). The overall results indicated main effects of repetition, similarity and test type. Replicating the model-based analyses (2HTSM and bivariate SDT), source memory (i.e., ACSIMs) was better in the repetition compared to the no-repetition conditions, $F(1, 142) = 7.38$, $p = .007$, $\eta_p^2 = .05$, worse for similar sources compared to dissimilar ones, $F(1, 142) = 28.06$, $p < .001$, $\eta_p^2 = .17$, and, crucially, better in the source recognition test (with source reinstatement) compared to the standard test, $F(1, 142) = 4.94$, $p = .028$, $\eta_p^2 = .03$. Again mirroring the model-based results, there was an interaction between repetition and test type, $F(1, 142) = 4.71$, $p = .032$, $\eta_p^2 = .03$, and between similarity and test type, $F(1, 142) = 9.72$, $p = .002$, $\eta_p^2 = .06$. Simple main effect analyses following up on the repetition × test type interaction, revealed that repeated encoding substantially benefits source memory in the second, source recognition test, $F(1, 142) = 14.19$, $p < .001$, $\eta_p^2 = .09$, but not in the first, standard test, $F(1, 142) = 1.91$, $p = .170$, again corroborating the model-based results and our assumption that the source recognition test with source reinstatement primarily depends on source storage. Simple main effect analyses following up on the similarity × test type interaction, revealed that the increase in ACSIM from the standard test to the source recognition test (i.e., reinstatement effect) was only significant when the sources were similar and thus difficult to discriminate, $F(1, 142) = 14.26$, $p < .001$, $\eta_p^2 = .09$. In contrast, there was no significant reinstatement effect for dissimilar sources, $F < 1$. Thus, in line with the SDT-based findings, however in contrast to the MPT-based results, there was no support for a reinstatement effect for dissimilar sources in the repetition condition, as indicated by an absent similarity × test type × encoding frequency three-way interaction, $F < 1$, and a non-significant ($p = .624$) pair-wise comparison in the repetition & dissimilar-sources condition. We thus refrain from interpreting the reinstatement effect for repeated dissimilar sources in the MPT-based analyses.

Table B1. Means and standard deviations of corrected hit rates (PR) and averaged conditional source identification measures (ACSIM) for each condition and test.

Test	Condition	PR	ACSIM
Test 1	no-repetition & dissimilar-sources	.48 (.21)	.77 (.13)
(standard)	repetition & dissimilar-sources	.57 (.23)	.79 (.14)
	no-repetition & similar-sources	.40 (.18)	.62 (.15)
	repetition & similar-sources	.50 (.21)	.66 (.17)
Test 2	no-repetition & dissimilar-sources	.42 (.20)	.74 (.13)
(with reinstatement)	repetition & dissimilar-sources	.51 (.23)	.80 (.12)
	no-repetition & similar-sources	.32 (.15)	.65 (.12)
	repetition & similar-sources	.42 (.21)	.74 (.11)

Note: Standard deviation in brackets. PR = Corrected hit rate (item memory). ACSIM = Averaged conditional source identification measures (source memory).

Appendix C. Response frequencies

Table C1. Aggregated response frequencies per test and condition.

		Participant's response					
		Test 1 (standard)			Test 2 (with reinstatement)		
Condition	Correct response	Source A	Source B	New	Source A	Source B	New
no-repetition & dissimilar-sources	Source A	681	213	366	687	210	363
	Source B	203	715	342	238	658	364
	New	159	148	953	207	166	887
repetition & dissimilar-sources	Source A	810	195	290	841	182	272
	Source B	207	816	272	209	831	255
	New	141	131	1020	190	178	927
no-repetition & similar-sources	Source A	546	362	387	555	334	406
	Source B	328	580	387	298	622	375
	New	194	197	904	231	255	809
repetition & similar-sources	Source A	609	305	346	705	257	298
	Source B	313	619	328	251	721	288
	New	133	159	968	251	183	826

Note: Source A = male speaker "Jakob" in all conditions, Source B = female speaker "Susanne" in the dissimilar-sources conditions, Source B = male speaker "Johan" in the similar-sources conditions.

Appendix D. Bayesian-hierarchical MPT analysis

Table D1. Parameter estimates of the two-high-threshold multinomial model of source monitoring (2HTSM) based on a Bayesian-hierarchical estimation approach.

		Parameter estimates				
Test	Condition	*D*	*d*	*b*	*g*	*a*
Test 1	no-repetition & dissimilar-sources	.47 [.40; .54]	.87 [.74; .98]	.43 [.33; .52]	.50 [.46; .54]	.95 [.85; 1.00]
(standard)	repetition & dissimilar-sources	.59 [.50; .68]	.84 [.71; .95]	.43 [.33; .52]	.51 [.46; .55]	.95 [.86; 1.00]
	no-repetition & similar-sources	.39 [.33; .45]	.34 [.14; .54]	.49 [.40; .57]	.48 [.45; .51]	.73 [.24; .99]
	repetition & similar-sources	.50 [.42; .58]	.55 [.27; .84]	.43 [.34; .51]	.49 [.45; .52]	.69 [.34; .96]
Test 2	no-repetition & dissimilar-sources	.40 [.33; .47]	.87 [.73; .99]	.49 [.42; .56]	.54 [.49; .58]	
(with reinstatement)	repetition & dissimilar-sources	.54 [.46; .62]	.95 [.86; 1.00]	.55 [.48; .63]	.51 [.46; .56]	
	no-repetition & similar-sources	.31 [.25; .36]	.66 [.36; .94]	.56 [.50; .62]	.46 [.43; .49]	
	repetition & similar-sources	.42 [.35; .50]	.94 [.83; 1.00]	.58 [.49; .66]	.53 [.49; .56]	

Note: Parameter estimation of the two-high-threshold multinomial model of source monitoring (Bayen et al., 1996) based on the observed individual response frequencies with the latent-trait approach (Klauer, 2010) as implemented in the *R* package TreeBUGS (Heck et al., 2018). Brackets indicate 95% Bayesian credibility intervals. D = probability of detecting a word as previously presented or not presented; d = probability of correctly recalling the source of a recognised word; b = probability of guessing that a word was previously presented; g = probability of guessing that a detected or undetected word was spoken by Source A (i.e., speaker "Jakob"); a = proportional change in d from the first to the second test (quantifies the size of the reinstatement effect: the higher a the lower the reinstatement effect). Overall, the model fit the data well, all $T_1 \geq .061$ for the mean structure and all $T_2 \geq .103$ for the covariance structure.

Appendix E. Reinstatement effect for sources presented first versus second in Test 2

Table E1. Parameter estimates and model fit of the two-high-threshold multinomial model of source monitoring (2HTSM) with separate estimation of source memory for sources presented first versus second in Test 2.

		Parameter estimates						
Test	Condition	*D*	d_1	d_2	*b*	*g*	a_d_1	a_d_2
Test 1	no-repetition & dissimilar-sources	.48	.79	.85	.46	.54	.88	1.00
(standard)		[.45; .50]	[.65; .92]	[.73; .96]	[.44; .49]	[.48; .59]	[.68; 1.08]	[.79; 1.21]
	repetition & dissimilar-sources	.57	.80	.85	.49	.51	.81	.91
		[.55; .60]	[.70; .91]	[.75; .94]	[.46; .52]	[.46; .57]	[.67; .94]	[.76; 1.05]
	no-repetition & similar-sources	.40	.29	.54	.50	.53	.63	.67
		[.37; .43]	[.09; .48]	[.38; .79]	[.48; .53]	[.48; .58]	[.10; 1.17]	[.42; .91]
	repetition & similar-sources	.50	.54	.42	.46	.46	.63	.48
		[.47; .53]	[.41; .67]	[.25; .58]	[.43; .49]	[.41; .52]	[.45; .82]	[.28; .68]
Test 2	no-repetition & dissimilar-sources	.42	.90	.81	.51	.53		
(with reinstatement)		[.38; .45]	[.76; 1.05]	[.67; .95]	[.48; .53]	[.48; .58]		
	repetition & dissimilar-sources	.51	1.00	.93	.58	.51		
		[.48; .54]	[.89; 1.10]	[.83; 1.04]	[.55; .61]	[.46; .56]		
	no-repetition & similar-sources	.32	.46	.81	.55	.56		
		[.29; .35]	[.22; .69]	[.63; .98]	[.53; .58]	[.51; .60]		
	repetition & similar-sources	.42	.85	.87	.60	.50		
		[.39; .45]	[.71; 1.00]	[.73; 1.01]	[.57; .62]	[.46; .55]		

Note: Brackets indicate 95% confidence intervals. D = probability of detecting a word as previously presented or not presented; $d_{1/2}$ = probability of correctly recalling the source presented first versus second in Test 2 (and left versus right in Test 1); b = probability of guessing that a word was previously presented; g = probability of guessing the source presented first in Test 2 (and left in Test 1); $a\_d_{1/2}$ = reinstatement effect for items in which correct source was reinstated first versus second in Test 2 (the higher a the lower the reinstatement effect). Overall, the model fit the data well, $G^2(8) = 5.36$, $p = .718$ (across all four conditions and both tests).